Adversarially Robust Distillation by Reducing the Student-Teacher Variance Gap
– *Supplementary Material* –

Junhao Dong[♣,♦], Piotr Koniusz[⋆,♥,♠], Junxi Chen[◇], and Yew-Soon Ong[⋆,♣,♦]

[♣]Nanyang Technological University, Singapore
[♦]Centre for Frontier AI Research, IHPC, A*STAR, Singapore
{junhao003,asysong}@ntu.edu.sg
[♠]Australian National University, Canberra, Australia
[♥]Data61♥CSIRO, Canberra, Australia
piotr.koniusz@data61.csiro.au
[◇]Sun Yat-sen University, Guangzhou, China
chenjx353@mail2.sysu.edu.cn

Abstract. In this supplementary material, we provide a detailed experimental configuration (Appendix A). Furthermore, we present more details about our *adverSarially robusT distillAtion by Reducing Student-teacHer varIance gaP* (STARSHIP) in Appendix B, including the adversarially pre-trained teacher models and the efficient extension with a single-step adversary generation. Appendix C incorporates further evaluations of robustness. Moreover, we provide visualization results (Appendix D) and hyper-parameter analysis (Appendix E). Potential limitations are also discussed in Appendix F.

A Experimental Configuration

In this section, we elaborate on the experimental settings used in this paper, including detailed descriptions of related datasets for adversarially robust knowledge distillation and the implementation details of our STARSHIP.

A.1 Dataset Description

Following the existing works on adversarially robust knowledge distillation [14, 33, 34], we conduct all the experiments on three standard datasets: CIFAR-10, CIFAR-100 [16], and ImageNet-100 [5]. The CIFAR-10 dataset comprises a collection of 60,000 color images, each with a resolution of 32×32 pixels, categorized into 10 distinct classes. CIFAR-100 mirrors CIFAR-10 but divides images into 100 classes, each represented by 600 samples. ImageNet-100 is a subset of the standard ImageNet dataset, which is utilized to evaluate the transferability of adversarial robustness in the context of real-world data. Specifically, ImageNet-100 contains 130K color images spanning a subset of 100 classes chosen from the

⋆ Corresponding authors.

original 1,000 classes. For robustness transfer with auxiliary data (referenced in Table 5), we incorporate 1 million synthetic images generated by the Denoising Diffusion Probabilistic Model (DDPM) [13], explicitly tailored for the CIFAR-10 and CIFAR-100 datasets in line with the established protocols [6, 21, 22].

A.2 Implementation Details

Following the experimental settings in previous works [14, 33, 34] and RobustBench [4], we adopt ResNet-18/34 [12], MobileNetV2 (MNV2) [23], and Wide-ResNet-28-10 (WRN-28) [31] for both the teacher and student models. Beyond these deep Convolutional Neural Networks (CNNs), we also incorporate Vision Transformers (ViTs) [11] as the teacher architecture for robustness transfer to lightweight models. Our experiments mainly concentrate on two principal knowledge distillation settings: (1) distillation from a large-scale teacher model to a lightweight but more efficient student model, and (2) self-distillation, where the teacher and student models are of identical network architecture.

For optimizing network parameters during adversarially robust knowledge distillation, we employ the Stochastic Gradient Descent (SGD) as the optimization algorithm, characterized by a momentum factor of 0.9, a weight decay factor of 5×10^{-4}, and a cyclic learning rate tuning strategy [25], peaking at a learning rate of 0.1. The weighting hyper-parameters β in Eq. (4&7) and γ for power normalization are fixed at values of 0.8 and 0.5, respectively. Furthermore, we determine the loss weighting coefficients $\lambda_1 = 1.0$, $\lambda_2 = 2.0$, and $\lambda_3 = 1.0$ across all experimental setups. Detailed analysis of different settings for hyper-parameters can be found in Appendix E. For robustness evaluation, we mainly focus on the ℓ_∞-norm threat model with a maximum perturbation intensity of $\epsilon = 8/255$, except where explicitly noted otherwise. During adversarially robust knowledge distillation, we utilize the Projected Gradient Descent (PGD) method [17] with $n = 10$ iteration steps (step size=$\alpha = 2/255$) to generate adversarial samples. Recall that our original STARSHIP mainly relies on the standard adversary generation strategy during the robustness transfer (see Eq. (2)), where the student's prediction on adversarial samples is optimized to deviate from the fixed prediction of the teacher model with respect to their clean counterpart. Despite its training efficiency, such a training scheme overlooks the gradient flow of the teacher model, leading to a suboptimal robustness transfer. Following [14], we introduce an adaptive version of our STARSHIP, dubbed Ada-STARSHIP, which replaces the fixed prediction alignment reference (clean samples) with their adversarial counterparts for the teacher model. In other words, this adaptive adversary generation scheme maximizes the prediction gap between the teacher and student models with regard to the same adversary (see Eq. (3)). To ensure a fair and comprehensive robustness evaluation, we conduct all the evaluations in accordance with the adaptive attack principle. All the experiments are conducted based on a single NVIDIA Tesla A100 GPU.

B Details of STARSHIP

In this section, we provide further details of our STARSHIP method, highlighting the adversarially pre-trained teacher models and introducing an efficient extension of our method in the context of single-step adversary generation.

B.1 Adversarially Pre-trained Teacher Models

In this study, we derive all the teacher models from scratch via adversarial training. We mainly resort to two standard adversarial training methods, TRADES [32] and SCORE [20], to construct robust teacher models for CNN architectures. Specifically, TRADES [32] minimizes a surrogate upper bound of the robust risk by aligning the predictions between clean samples and their adversarial counterparts using the KL divergence, as follows:

$$\min_{\boldsymbol{\theta}_t, \boldsymbol{\theta}'_t} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\Big[\mathcal{L}_{\mathrm{CE}}(g_{\boldsymbol{\theta}'_t}(f_{\boldsymbol{\theta}_t}(\mathbf{x})), y) + \omega \cdot \max_{\|\boldsymbol{\delta}\|_\infty<\epsilon} \mathcal{L}_{\mathrm{KL}}(g_{\boldsymbol{\theta}'_t}(f_{\boldsymbol{\theta}_t}(\mathbf{x})) \| g_{\boldsymbol{\theta}'_t}(f_{\boldsymbol{\theta}_t}(\mathbf{x}+\boldsymbol{\delta})))\Big], \quad (13)$$

where $\omega \geq 0$ controls the trade-off between natural performance and adversarial robustness, with higher ω values prioritizing adversarial robustness. We adopt $\omega = 6.0$, following the original configuration of TRADES [32]. Furthermore, to mitigate the inductive bias towards local invariance during adversarial training, SCORE [20] theoretically justifies the effectiveness of the squared error variant of Eq. (13), replacing the KL divergence. Note that the adversarial training process of SCORE [20] is based on a hybrid dataset composed of the original dataset associated with auxiliary generated data for improved robustness. During the adversarially robust knowledge distillation, we merely incorporate the original training dataset unless stated otherwise. For teacher models based on ViT architectures, we utilize PGD-based adversarial training [17], augmented with the strategies of attention random dropping and perturbation random masking as introduced in [18]. Specifically, we adopt ViT-Base [8] and DeiT-Small [26] as architectures of teacher models for adversarially robust knowledge distillation.

B.2 Single-step Robust Distillation

In this section, we elaborate on an efficient extension with the single-step adversary generation strategy to reduce the computational cost in adversarially robust knowledge distillation. Generally, the predominant computational overhead in robustness transfer historically stems from the multi-step adversary generation, necessitating multiple gradient backpropagations. Although previous studies have shown the feasibility and the potential of single-step adversary generation during adversarial training [1, 15, 30], rare efforts have been made in the context of efficient robustness transfer. To relieve such a computationally intensive demand, we made the first attempt to explore adversarially robust knowledge distillation with the single-step adversary generation strategy. Specifically, we replace the multi-step adversary generation in STARSHIP (see Eq. (2))

with its single-step counterpart $\widetilde{\mathbf{x}} = \mathbf{x} + \widetilde{\boldsymbol{\delta}}$ as below:

$$\widetilde{\boldsymbol{\delta}} = \psi\left[\boldsymbol{\delta}_0 + \alpha' \operatorname{sign}\left(\nabla_{\boldsymbol{\delta}_0} \mathcal{L}_{\mathrm{KL}}\left(g_{\boldsymbol{\theta}'_t}(f_{\boldsymbol{\theta}_t}(\mathbf{x})) \| g_{\boldsymbol{\theta}'_s}(f_{\boldsymbol{\theta}_s}(\mathbf{x}+\boldsymbol{\delta}_0))\right)\right)\right], \tag{14}$$

where $\boldsymbol{\delta}_0$ is the randomly initialized perturbation drawn from a certain distribution $\boldsymbol{\Omega}$, while $\psi[\cdot]$ denotes the projection operator that constrains the adversarial perturbation within the ℓ_∞-norm hypersphere. The single-step variant of our Ada-STARSHIP can be easily obtained by replacing the fixed prediction alignment reference $g_{\boldsymbol{\theta}'_t}(f_{\boldsymbol{\theta}_t}(\mathbf{x}))$ with the single-step adversary $g_{\boldsymbol{\theta}'_t}(f_{\boldsymbol{\theta}_t}(\mathbf{x}+\boldsymbol{\delta}_0))$. Furthermore, the extension of other robustness transfer methods can also be achieved in such an efficient scheme. In particular, N-FGSM [15] has emerged as one of the most effective single-step strategies for robustness enhancement by adding strong noise augmentations to regularize the loss landscape, which can be conducted by simply disabling the projection operator. Highlighted by its scalability and efficiency, we investigate the extension of robust distillation methods with N-FGSM [15] in Table 6. Despite a minor performance drop compared to its multi-step counterpart (see Table 1), single-step robust distillation can achieve overall better training efficiency by reducing the gradient computing iterations. In addition, we show that the single-step extensions of our STARSHIP method and its variant can also maintain comparable efficacy on both clean and adversarial samples in comparison with their multi-step counterparts.

C Additional Robustness Evaluations and Analyses

In this section, we provide additional experimental results to further justify the efficacy and generalization ability of our method, including comparisons with adversarial training, the extension to black-box model extraction, and black-box robustness evaluations. Moreover, we incorporate further analyses of our STARSHIP method w.r.t. different distance types for statistics alignment. All the settings are consistent with the main text.

C.1 Comparisons with Adversarial Training Baselines

To further justify the efficacy of robustness transfer from large-scale teacher models, we compare our STARSHIP method with adversarial training baselines that build robust models from scratch. As shown in Table 9, our method can significantly outperform adversarial training methods in terms of both clean accuracy and adversarial robustness. This also indicates that the guidance of a well-optimized and large-scale teacher model can lead to better performance compared with the sole supervision by one-hot labels.

Table 9: Comparison between our STARSHIP method (WRN-28 → ResNet-18) with adversarial training approaches. We report both clean and robust accuracies (%) on CIFAR-10/100.

Method		CIFAR-10			CIFAR-100		
		Clean	PGD	AA	Clean	PGD	AA
PGD-AT	[17]	83.80	51.40	47.68	57.39	28.36	23.18
TRADES	[32]	82.45	52.21	48.88	54.36	27.49	24.19
MART	[28]	82.20	53.94	48.04	54.78	28.79	24.58
HAT	[21]	84.86	52.04	48.85	58.73	27.92	23.34
STARSHIP		**86.47**	**57.45**	**53.78**	**61.54**	**32.24**	**27.46**

C.2 Black-box Model Extraction

Black-box model extraction [19, 27] has been recognized as one of the most fundamental security threats to existing deep learning-based systems. Specifically, such an attack scheme can effectively recover an online black-box model without access to its network parameters or even training data. Different from the data-relevant knowledge transfer setting we evaluated in the main text, where the data for pre-training the teacher model and the distillation are from the same domain or even identical, we primarily focus on a data-irrelevant knowledge distillation setting to transfer the data-independent robustness from the teacher model to the student model.

Table 10: Black-box model extraction to distill the pre-trained teacher model (WRN-28) to the student model (ResNet-18) on CIFAR-10/CIFAR-100. We report both clean and robust accuracies (%).

Pre-training Dataset	**Distillation Dataset**	Method	Clean	PGD	AA
CIFAR-10	CIFAR-100	RSLAD [34]	69.86	42.65	37.24
		CRDND [29]	69.15	42.93	37.50
		GACD [2]	69.93	43.16	37.89
		AdaAD [14]	70.30	43.47	38.24
		STARSHIP	**72.19**	**45.76**	**41.13**
CIFAR-100	CIFAR-10	RSLAD [34]	43.57	22.71	18.29
		CRDND [29]	44.20	22.13	17.82
		GACD [2]	43.87	22.90	18.40
		AdaAD [14]	44.98	23.66	18.92
		STARSHIP	**45.73**	**25.31**	**20.96**

To investigate the cross-domain generalization ability of our STARSHIP method, we here explore the extension of adversarially robust knowledge distillation in the context of black-box model extraction. Specifically, the pre-training dataset for the teacher model is completely irrelevant to the dataset used for knowledge distillation to simulate the black-box scenarios of model extraction, where these two datasets have disjoint image categories. Here, we conduct model extraction in a pairwise scheme between CIFAR-10 and CIFAR-100 datasets, as shown in Table 10. Note that the evaluation is conducted on the pre-training dataset. As observed, our STARSHIP consistently achieves better clean accuracy as well as adversarial robustness in such a data-irrelevant scenario. This further indicates that our method can promote the transfer of generalizable robust knowledge rather than the robust knowledge that is strongly correlated with the data distribution.

C.3 Black-box Adversarial Robustness Evaluation

In addition to the white-box robustness evaluation in the main text, we here explore the black-box adversarial robustness of the student model distilled via our STARSHIP method (see Table 11). Following the evaluation setting from [3], we conduct black-box transferable adversarial attacks using iterative and non-iterative attack approaches. The adversarially

Table 11: Black-box robustness against iterative & non-iterative adversarial attacks when distilling from a large-scale teacher model on CIFAR-10.

Method	ResNet-18			MNV2		
	FGSM	PGD	MIM	FGSM	PGD	MIM
ARD [9]	67.87	66.77	66.56	66.72	65.32	65.04
IAD [33]	67.59	66.30	66.25	65.99	64.81	64.72
RSLAD [34]	68.46	67.27	66.99	68.56	67.05	66.94
AdaAD [14]	69.04	67.68	67.46	68.25	66.90	66.58
STARSHIP	**70.59**	**69.28**	**69.01**	**70.06**	**68.66**	**68.39**

pre-trained teacher model is adopted as the substitute model for adversary generation. We can observe that our method can simultaneously achieve better black-box robustness against both iterative (PGD [17] and MIM [7]) and non-iterative (FGSM [10]) adversarial attacks compared to other adversarially robust knowledge distillation approaches.

C.4 Performance w.r.t. Diverse Settings of $\mathcal{L}_\Omega$

We have shown the superior performance brought by the statistics alignment $\mathcal{L}_\Omega$ (Eq. (10)) within the student model under the self-distillation setting, where the teacher and student models share the same network architecture. To provide a better understanding of our method, we here explore the effect of each sub-matrix (statistical interaction) in $\mathcal{L}_\Omega$ (Eq. (10)) by blocking its gradient when optimizing such an alignment loss. Specifically, we report both clean and robust accuracies when detaching the gradient flow of both the covariance and gram sub-matrices extracted by $\psi_{ij}(\cdot)$ in $\mathcal{L}_\Omega$ during our robust knowledge distillation (see Table 12). As observed, blocking the gradient of statistical interactions related to adversaries (ψ_{12}, ψ_{21}, and ψ_{22}) leads to a drop in adversarial robustness, while eliminating the effect of the statistical interaction between clean samples (ψ_{11}) suppresses the natural performance.

Table 12: Diverse sub-matrices for gradient blocking in $\mathcal{L}_\Omega$ of our STARSHIP during **self-distillation** on CIFAR-10. We report both clean and robust accuracies (%).

Gradient Blocking	ResNet-18			MNV2		
	Clean	PGD	AA	Clean	PGD	AA
ψ_{11}	81.14	55.65	52.29	80.20	54.23	50.18
ψ_{12} & ψ_{21}	81.46	54.86	51.60	80.55	54.01	49.94
ψ_{22}	81.78	54.33	51.38	80.83	53.78	49.61
No Blocking	**81.97**	**55.72**	**52.42**	**80.97**	**54.28**	**50.46**

C.5 Robust Distillation with Diverse Single-step Strategies

In addition to the extension with the single-step adversary generation strategy (N-FGSM [15]) we discussed in the main manuscript, we also investigate such an efficient extension via other single-step strategies for adversarially robust knowledge distillation. Specifically, we adopted RS-FGSM [30] and GradAlign [1] for adversary generation and applied them to different robust distillation approaches when distilling from a large-scale teacher model (See Table 13). As observed, our STARSHIP method and its adaptive variant can effectively achieve superior performance on clean samples and their adversarial counterparts when

Table 13: Extension of robust distillation (WRN-28 → ResNet-18) with diverse **single-step adversary strategies** on the CIFAR-10 dataset. We report both clean and (Auto-Attack) robust accuracies (%) with the average training time per epoch.

Type	Strategy	Method		Clean	Robust	Time (s)
Teacher	—	SCORE	[20]	88.61	61.03	—
Student	RS-FGSM	IAD	[33]	84.07	46.11	68
		RSLAD	[34]	85.16	48.30	42
		AdaAD	[14]	86.35	49.74	112
		STARSHIP		86.93	50.58	50
		Ada-STARSHIP		87.60	51.14	121
	GradAlign	IAD	[33]	83.92	46.62	106
		RSLAD	[34]	84.77	48.69	85
		AdaAD	[14]	85.87	50.35	230
		STARSHIP		86.50	51.22	97
		Ada-STARSHIP		87.34	51.83	247

efficiently combing with different single-step adversary generation strategies, further highlighting the generalization ability of our proposed method.

C.6 Diverse Objective Functions for Parameter-level Perturbations

We here investigate the contribution of three main components within our parameter-level perturbations (Eq. (7)): (1) Clean Feature Alignment (CFA), (2) Adversarial Feature Alignment (AFA), and (3) Feature Covariance Alignment. Furthermore, we evaluate the efficacy of our adversarial optimization scheme at the parameter level against the natural optimization strategy. As shown in Table 14, we report the accuracy of clean samples and their adversarial counterparts w.r.t. different combinations of the component modules and optimization strategies. As observed, both CFA and AFA significantly aid the feature projection head in aligning the feature spaces of the teacher and student models, thereby enhancing the efficacy of knowledge transfer. The feature-level covariance alignment further improves both natural performance and adversarial robustness. Note that both the natural and adversarial optimization strategies can achieve excellent performance on clean and adversarial samples, while the adversarial optimization enhances the flatness of the parameter-loss landscape, thus reducing the robust generalization gap.

Table 14: Diverse settings for the parameter-level perturbation and its corresponding optimization strategy (Eq. (7)) within our STARSHIP (WRN-28 → ResNet-18) on CIFAR-10. We report both clean and robust accuracies (%).

Setting	Adv. Optim.			Nat. Optim.		
	Clean	PGD	AA	Clean	PGD	AA
w/o CFA	85.64	57.11	53.26	85.73	56.54	52.70
w/o AFA	86.28	55.21	51.59	86.40	54.84	51.34
w/o FCA	85.75	56.53	52.66	85.52	56.13	52.28
w/ All Modules	**86.47**	**57.45**	**53.78**	86.18	56.89	53.22

C.7 Aligning Variances Rather than Reducing Variances

In this paper, we mainly focus on aligning the variances in terms of features and prediction scores (Gram matrices) between the teacher and student models. It has been noted that the robust teacher model typically exhibits a lower variance gap compared to the student model, as illustrated in Figure 1. This observation raises a critical question: Could merely reducing variances, rather than aligning them with those of the teacher model, also improve the robustness transfer? To investigate this query, we introduce regularization terms aimed at reducing the values of the feature-level covariance matrix or the prediction-level Gram matrix into our baseline method (*i.e.*, prediction alignment in Eq. (4)), as shown in Table 15. As observed, reducing the feature variances or the prediction Gram variances does indeed improve the model robustness, which is accompanied by a significant trade-off with a decrease in clean accuracy. Conversely, aligning

Table 15: Diverse distillation strategies for our STARSHIP method (WRN-28 → ResNet-18). We report both clean and robust accuracies (%).

Distillation Strategy	CIFAR-10			CIFAR-100		
	Clean	PGD	AA	Clean	PGD	AA
Baseline (Eq. (4))	84.13	54.49	51.27	58.52	32.43	27.23
Reducing Feature Variance	83.43	55.25	51.76	57.94	33.19	28.10
Reducing Prediction Gram	83.98	55.03	51.60	58.41	32.96	27.65
Reducing Both	84.19	55.47	51.96	58.77	33.25	28.32
Ours (Aligning Variances)	**86.47**	**57.45**	**53.78**	**61.54**	**34.45**	**29.30**

these variances of the student model to mirror those of the teacher model can obtain a better trade-off between natural performance and adversarial robustness in comparison with merely reducing variances.

Feature variance gap & error bar. We report accuracy and Feature Variance Gap (**FVG**, %) in the table (right) (WRN-28 → ResNet-18, ℓ_∞-norm $\epsilon = 8/255$). Our method yields a low variance gap and high robustness. All baselines and ours share a single seed (same split). The error bars (last 2 rows in cyan) are computed on 10 random splits.

Table 16: Feature variance gap and error bars on CIFAR-10/100 (WRN-28 → ResNet-18).

Type	Method		CIFAR-10 Clean	CIFAR-10 AA	CIFAR-10 **FVG**	CIFAR-100 Clean	CIFAR-100 AA	CIFAR-100 **FVG**
Teacher	SCORE	[38]	88.61	61.03	13.95	63.64	31.13	11.53
Student	ARD	[18]	84.35	49.40	37.42	58.20	26.02	19.79
	IAD	[59]	83.46	49.09	37.54	57.35	26.22	19.54
	RSLAD	[60]	84.42	51.36	35.73	57.97	27.52	18.37
	AKD	[34]	86.04	50.11	36.65	60.79	26.93	18.97
	AdaAD	[25]	86.38	52.36	29.34	61.26	27.46	18.62
	STARSHIP		86.47	53.78	27.15	61.54	**29.30**	**16.02**
	Ada-STARSHIP		**87.04**	**54.47**	**25.89**	**62.19**	28.28	16.90
	STARSHIP		86.45±0.09	53.76±0.12	-	61.56±0.07	29.35±0.13	-
	Ada-STARSHIP		87.01±0.12	54.42±0.16	-	62.17±0.05	28.25±0.11	-

Robustness of ViT-based students. We have reported the adversarially robust knowledge distillation from a ViT-based teacher model in the main text. We here also evaluate the adversarial robustness of ViTs as the student model in the table right (**ViT-B → ViT-S**, ℓ_∞-norm $\epsilon = 8/255$). As expected, ViT-based students do not perform well on smaller datasets. However, our STARSHIP method still achieves the best performance in terms of clean examples and their adversarial counterparts.

Table 17: Robustness of ViT students on CIFAR-10/100 (WRN-28 → ResNet-18).

Type	Method		CIFAR-10 Clean	CIFAR-10 PGD	CIFAR-10 AA
Teacher	AT-PRM	[36]	83.98	53.10	49.66
Student	ARD	[18]	81.43	51.05	47.12
	IAD	[59]	81.15	51.59	47.46
	RSLAD	[60]	81.13	51.87	47.66
	AKD	[34]	81.78	51.36	47.27
	AdaAD	[25]	82.16	52.14	47.84
	STARSHIP		82.74	52.30	48.08
	Ada-STARSHIP		**83.19**	**52.63**	**48.31**

D Visualization

In addition to the attention visualizations shown in the main text, we also incorporate saliency visualizations of adversarial examples against different robustness transfer methods derived from the CIFAR-10 dataset, as illustrated in Fig. 6. These saliency maps are generated via the SmoothGrad technique [24], which smoothens the raw gradient of the class score function across the input domain. Recall that the adversary generation is based on the PGD method with the maximum perturbation intensity of $\epsilon = 8/255$. To maintain a fair and comprehensive analysis, we

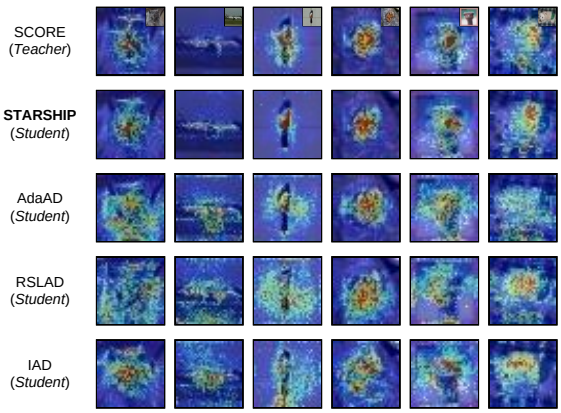
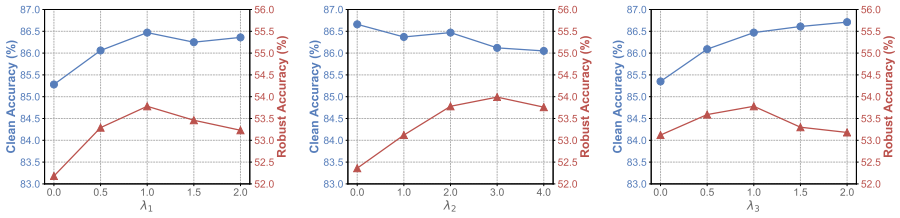


Fig. 6: Saliency visualizations of both the teacher and student models distilled via different methods.



(a) **(b)** **(c)**

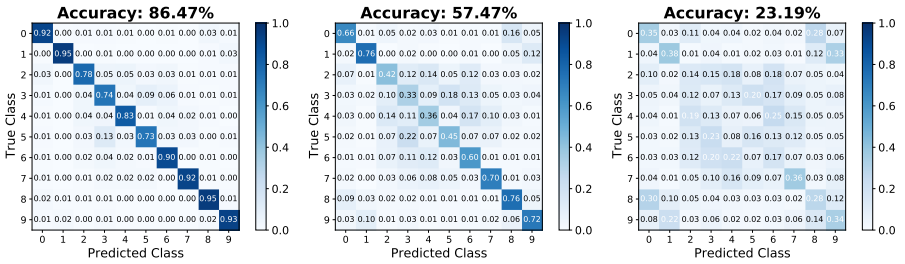
Fig. 7: Sensitivity analyses of our STARSHIP method (WRN-28 → ResNet-18) with diverse hyper-parameter settings. We report both the clean accuracy and (Auto-Attack) robust accuracy on CIFAR-10 when tuning diverse loss weighting factors during robust distillation from a large-scale teacher model to a lightweight student model.

conduct adaptive attacks for all the models to create a corresponding adversarial counterpart for each clean example.

Notably, the saliency maps generated from the robust student model, utilizing our STARSHIP method, demonstrate a higher degree of consistency with saliency maps of the teacher model compared to other robust distillation approaches. Such a saliency alignment also indicates the efficacy of our STARSHIP method in capturing and transferring the adversarially robust knowledge from the teacher model. In the meantime, the saliency regions are primarily concentrated on the discriminative features of the target object, suggesting the implicit alignment between the transferred knowledge and human vision. These visual observations lend further support to the robustness of our distilled student models against unforeseen adversarial examples.

E Hyper-Parameter Analyses

We have indicated that the hyper-parameter β controls the trade-off between natural performance and adversarial robustness in the main manuscript. To further deepen the comprehension of our work, we delve into the analyses with respect to the efficacy of key hyper-parameters in our STARSHIP method. As shown in Fig. 7, we report both clean and (Auto-Attack) robust accuracy of our method under diverse hyper-parameter settings. Note that all hyper-parameters in this study were tuned based on a tiny subset of the CIFAR-10 training set to ensure fairness. This hyper-parameter configuration is subsequently applied across other datasets to maintain consistency. We can easily observe a performance boost on both clean and adversarial examples when enlarging the weighting for feature covariance alignment (λ_1) and statistics alignment (λ_3). In the meantime, appropriately choosing the loss weighting factor λ_2 for adversarially robust knowledge distillation can also lead to a reasonable trade-off between natural performance and adversarial robustness.



(a) Radius $\epsilon = 0$ **(b)** Radius $\epsilon = 8/255$ **(c)** Radius $\epsilon = 16/255$

Fig. 8: Confusion matrices of our STARSHIP method (WRN-28 → ResNet-18) from CIFAR-10 on clean samples and their PGD-based adversarial samples using ResNet-18.

F Limitations

To ensure a comprehensive evaluation, we discuss the potential limitations of our work and the corresponding solutions. For error analyses, we provide a confusion matrix analysis of our method under different attack strengths on CIFAR-10 (see Figure 8). The confusion matrix suffers from a more severe disruption when the attack strength (perturbation radius) increases. In other words, adversarial examples with a higher attack strength can deceive the model, leading to an increase in off-diagonal elements of the confusion matrix. However, this trend is not unique to our model but rather a common challenge for all the adversarially robust models. The adoption of larger models and the augmentation with more data have been identified as effective solutions to counteract such a robustness degradation. In this paper, we primarily focus on facilitating robust knowledge transfer from large-scale models to lightweight models, thereby enhancing their practical deployability and adaptability in real-world settings.

Another limitation lies in the additional training cost of our robust knowledge distillation method, stemming from building the feature projection head. To mitigate such a potential drawback, we extend our method with single-step adversary generation to further improve its efficiency, as presented in Table 6 of the main text. Our method can achieve better training efficiency while maintaining performance efficacy. Kindly note that our method does not incur additional test-time computational costs compared to other approaches, as our test-time classification model shares the same architecture with classifiers of other methods, which ensures that our method remains viable for practical deployment.

References

1. Andriushchenko, M., Flammarion, N.: Understanding and improving fast adversarial training. Advances in Neural Information Processing Systems **33**, 16048–16059 (2020)
2. Bai, T., Zhao, J., Wen, B.: Guided adversarial contrastive distillation for robust students. IEEE Transactions on Information Forensics and Security (2023)

3. Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., Kurakin, A.: On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705 (2019)
4. Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., Hein, M.: Robustbench: a standardized adversarial robustness benchmark. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (2021)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Dong, J., Moosavi-Dezfooli, S.M., Lai, J., Xie, X.: The enemy of my enemy is my friend: Exploring inverse adversaries for improving adversarial training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24678–24687 (June 2023)
7. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9185–9193 (2018)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations, ICLR (2021)
9. Goldblum, M., Fowl, L., Feizi, S., Goldstein, T.: Adversarially robust distillation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 3996–4003 (2020)
10. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015)
11. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al.: A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence **45**(1), 87–110 (2022)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Huang, B., Chen, M., Wang, Y., Lu, J., Cheng, M., Wang, W.: Boosting accuracy and robustness of student models via adaptive adversarial distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24668–24677 (2023)
15. de Jorge Aranda, P., Bibi, A., Volpi, R., Sanyal, A., Torr, P., Rogez, G., Dokania, P.: Make some noise: Reliable and efficient single-step adversarial training. Advances in Neural Information Processing Systems **35**, 12881–12893 (2022)
16. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
17. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: 6th International Conference on Learning Representations, ICLR (2018)
18. Mo, Y., Wu, D., Wang, Y., Guo, Y., Wang, Y.: When adversarial training meets vision transformers: Recipes from training to architecture. Advances in Neural Information Processing Systems **35**, 18599–18611 (2022)

19. Orekondy, T., Schiele, B., Fritz, M.: Knockoff nets: Stealing functionality of black-box models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4954–4963 (2019)
20. Pang, T., Lin, M., Yang, X., Zhu, J., Yan, S.: Robustness and accuracy could be reconcilable by (proper) definition. In: International Conference on Machine Learning. pp. 17258–17277. PMLR (2022)
21. Rade, R., Moosavi-Dezfooli, S.: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In: The Tenth International Conference on Learning Representations, ICLR (2022)
22. Rebuffi, S.A., Gowal, S., Calian, D.A., Stimberg, F., Wiles, O., Mann, T.A.: Data augmentation can improve robustness. Advances in Neural Information Processing Systems **34**, 29935–29948 (2021)
23. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
24. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
25. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. In: Artificial intelligence and machine learning for multi-domain operations applications. vol. 11006, pp. 369–386. SPIE (2019)
26. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
27. Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T.: Stealing machine learning models via prediction {APIs}. In: 25th USENIX security symposium (USENIX Security 16). pp. 601–618 (2016)
28. Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., Gu, Q.: Improving adversarial robustness requires revisiting misclassified examples. In: 8th International Conference on Learning Representations, ICLR (2020)
29. Wang, Y., Chen, Z., Yang, D., Liu, Y., Liu, S., Zhang, W., Qi, L.: Adversarial contrastive distillation with adaptive denoising. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
30. Wong, E., Rice, L., Kolter, J.Z.: Fast is better than free: Revisiting adversarial training. In: 8th International Conference on Learning Representations, ICLR (2020)
31. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: Proceedings of the British Machine Vision Conference (2016)
32. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning. pp. 7472–7482. PMLR (2019)
33. Zhu, J., Yao, J., Han, B., Zhang, J., Liu, T., Niu, G., Zhou, J., Xu, J., Yang, H.: Reliable adversarial distillation with unreliable teachers. In: The Tenth International Conference on Learning Representations, ICLR (2022)
34. Zi, B., Zhao, S., Ma, X., Jiang, Y.G.: Revisiting adversarial robustness distillation: Robust soft labels make student better. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16443–16452 (2021)