Adversarially Robust Distillation by Reducing the Student-Teacher Variance Gap

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

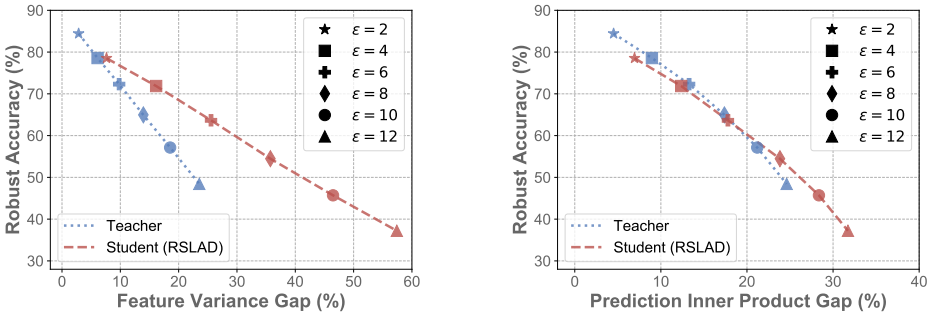
[♣]Nanyang Technological University, Singapore
[♦]Centre for Frontier AI Research, IHPC, A*STAR, Singapore
{junhao003,asysong}@ntu.edu.sg
[♠]Australian National University, Canberra, Australia
[♥]Data61♥CSIRO, Canberra, Australia
piotr.koniusz@data61.csiro.au
[♢]Sun Yat-sen University, Guangzhou, China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

Abstract. Adversarial robustness generally relies on large-scale architectures and datasets, hindering resource-efficient deployment. For scalable solutions, adversarially robust knowledge distillation has emerged as a principle strategy, facilitating the transfer of robustness from a large-scale teacher model to a lightweight student model. However, existing works focus solely on sample-to-sample alignment of features or predictions between the teacher and student models, overlooking the vital role of their statistical alignment. Thus, we propose a novel adversarially robust knowledge distillation method that integrates the alignment of feature distributions between the teacher and student backbones under adversarial and clean sample sets. To motivate our idea, for an adversarially trained model (*e.g.*, student or teacher), we show that the robust accuracy (evaluated on testing adversarial samples under an increasing perturbation radius) correlates negatively with the gap between the feature variance evaluated on testing adversarial samples and testing clean samples. Such a negative correlation exhibits a strong linear trend, suggesting that aligning the feature covariance of the student model toward the feature covariance of the teacher model should improve the adversarial robustness of the student model by reducing the variance gap. A similar trend is observed by reducing the variance gap between the gram matrices of the student and teacher models. Extensive evaluations highlight the state-of-the-art adversarial robustness and natural performance of our method across diverse datasets and distillation scenarios.

1 Introduction

Deep Neural Networks (DNNs) [20, 22, 45] have been demonstrated to be susceptible to adversarial samples [49]. These malicious inputs, subtly altered with visually imperceptible perturbations, can disrupt deep learning-based systems

⋆ Corresponding authors.



(a) Gap between Feature-wise Variances. **(b)** Gap between Pred. Gram-wise Variances.

Fig. 1: Analysis on robust teacher and lightweight student (RSLAD [59]). Fig. 1a: The robust accuracy (on test adversarial samples under an increasing perturbation radius $255 \cdot \epsilon$) correlates negatively with the gap between the feature variance evaluated on testing adversarial samples and testing clean samples. Fig. 1b: A similar trend occurs for the gap between variances of prediction score-based gram matrices (inner product). The negative correlation with linear characteristics suggests that reducing the variance gaps of the student toward the variance gaps of the teacher should improve robustness.

[14, 27]. Thus, establishing robustness against adversarial samples becomes vital for the practical applicability of DNNs in trustworthy real-world applications [11, 30]. To this end, adversarial training has emerged as the most effective defense approach against adversarial samples [12, 21, 37]. However, achieving substantial adversarial robustness essentially necessitates the use of large-scale models or abundant training data [47, 52]. To circumvent this and achieve robustness in resource-efficient scenarios, adversarially robust knowledge distillation [10, 19, 25, 58, 59] transfers adversarial robustness from a large-scale *teacher* model to a lightweight *student* model under some performance trade-off. Existing robust distillation approaches typically align feature representations or predictions between the teacher and student models. However, such a sample-to-sample feature alignment scheme differs from the feature-statistics alignment which often serves as a better mechanism for modeling properties of sets of samples. This oversight leads to a significant gap between the variance of features of adversarial samples and the variance of features of clean samples, consequently limiting the robust generalization ability of the student, as detailed in the motivation below.

Motivation. The feature variance has been linked with the generalization gap. For example, Huang *et al.* [26] observed that for functions $f \in \mathcal{F}$, where $\mathcal{F}$ is a finite class of functions, for each $\delta \in (0, 1)$ with probability at least $1-\delta$, the gap between expected and empirical risks r and $\tilde{r}$ is bounded as:

$$|r(f, \mathcal{Q}) - \tilde{r}(f(\mathbf{X}), \mathbf{Y})| \leq \zeta + \sqrt{8(N-1)\sigma^2\big(f(\mathbf{X})\big)\iota/N},$$

where $\iota = \ln(3|\mathcal{F}|/\delta)$, $\zeta = 3.5\iota/N + 3/$, $|\mathcal{F}|$ is the cardinality of a class of functions $\mathcal{F}$, N is the number of samples, $\mathcal{Q}$ is the data distribution (*e.g.*, $(\mathbf{x}, y) \sim \mathcal{Q}$), whereas $f(\mathbf{X})$ and $\mathbf{Y}$ are features and labels for empirical samples (*e.g.*, column vectors of $\mathbf{X}$), and $\sigma^2(\cdot)$ estimates their variance.

Considering the above bound, we suggest that matching the variance of features of adversarial samples with the variance of features of clean samples can naturally tighten the gap between the expected risk and the empirical risks of adversaries, *i.e.*, $|r(f, \mathcal{Q}) - \tilde{r}(f(\mathbf{X}_{adv}), \mathbf{Y})|$ gets closer to $|r(f, \mathcal{Q}) - \tilde{r}(f(\mathbf{X}), \mathbf{Y})|$ when $\sigma^2\big(f(\mathbf{X}_{adv})\big)$ gets closer to $\sigma^2\big(f(\mathbf{X})\big)$. The bound in the equation is fulfilled simultaneously for both $\mathbf{X}$ and $\mathbf{X}_{adv}$ with probability at least $(1-\delta)^2$.

To support our claim, we study the robust accuracy of adversarial test samples w.r.t. the gap between the variance of features from adversarial test samples and that from clean samples (see Figure 1a). Note that the larger the variance gap is, the worse the robust accuracy is, irrespective of whether the experiment is performed on the robust teacher or the student model (RSLAD [59]). Moreover, as RSLAD performs sample-to-sample distillation, its variance gap remains larger than that of the teacher across diverse perturbation radii, resulting in comparatively lower robustness. Figure 1b presents similar findings related to the variance gap of prediction score-based gram matrices (inner products).

Drawing on the insights from the above analyses, we hypothesize that bridging the gap in the robust accuracy of student and teacher can be achieved by aligning (i) the student's covariance to the teacher's covariance, (ii) the student's statistics for adversarial samples with the student's statistics on clean samples if the student is unable to distill the robust knowledge from the teacher, (iii) the student's prediction-based gram matrix with the teacher's gram matrix, as gram matrices capture the complementary information to covariance matrices.

Moreover, we introduce cost-effective parameter-level adversarial perturbations on the feature projection head alone to improve the student-teacher alignment of the multivariate Normal distributions (covariances). Looking at the spectrum of covariances, one may imagine that this mechanism helps find dominant directions in the eigenspace that are prone to misalignment, and the parameters of the feature projection head are minimized to overcome this effect.

Our key contributions can be summarized as follows:

i. We provide an intuitive motivation by analyzing the robust performance *vs.* the gap between variances of features of adversarial samples and clean samples. Such an exploration reveals a strong negative correlation with a linear trend, suggesting that aligning the covariance for adversarial features with the covariance for clean features helps improve the adversarial robustness.
ii. Contrary to prior robust distillation methods that typically conduct sample-to-sample feature alignment between the teacher and student models, or even align adversarial features of the student with the corresponding clean features of the teacher, we investigate the covariance alignment between the student and the teacher, as well as aligning the sub-covariance built from adversarial samples toward those from clean samples. Similarly, we investigate aligning score-based gram matrices for better robustness transfer.
iii. We provide comprehensive experiments across various datasets and scenarios, highlighting the superiority of our method, dubbed adverSarially robusT distillAtion by Reducing Student-teacHer varIance gaP (STARSHIP), over state-of-the-art adversarially robust knowledge distillation approaches.

2 Related Works

Adversarial Training. Goodfellow *et al*. [21] proposed to enhance robustness by adaptively incorporating adversaries [49] into the training data, which represents a "train-from-scratch" scheme. Existing works primarily focus on optimizing the input-loss landscape to mitigate the disruption impact of input-level noise on final predictions [13,16,37,43,55]. Wu *et al*. [51] introduced a double-perturbation mechanism that combines both input-level and parameter-level perturbations, flattening the parameter-loss landscape during adversarial training. Our method leverages parameter-level perturbations but in the context of exacerbating the student-teacher disagreement at the covariance level. To this end, the parameters of the feature projection head are optimized to overcome the introduced perturbations to attain robust alignment of student-teacher covariances, narrowing the variance gap between the student and teacher models. Despite the effectiveness of adversarial training, such methods heavily depend on large network architectures [52] and a massive amount of training data [15, 47]. In contrast, we focus on obtaining the robustness w.r.t. lightweight models by robust distillation.

Vanilla Knowledge Distillation. Hinton *et al*. [23] proposed knowledge distillation with the goal of transferring the learned knowledge from a large-scale model (*teacher*) to a lightweight model (*student*) by leveraging ground-truth labels and the teacher's predictions. Subsequent knowledge distillation studies focus on the feature-level distillation [4, 5, 42, 50], and the prediction-level distillation [2, 6, 28, 39]. However, vanilla knowledge distillation has been identified as insufficient for robustness transfer from an adversarially trained teacher [19].

Adversarially Robust Knowledge Distillation. To bridge the robustness gap between large-scale and lightweight models, recent studies have investigated adversarially robust knowledge distillation [19, 25, 58, 59]. Zi *et al*. [59] primarily resort to the soft labels of clean samples predicted by the teacher model to guide the student's predictions. Huang *et al*. [25] further integrated the gradient flow of the teacher model into the robust knowledge distillation to search for the worst-case matching point (adversarial sample). In contrast to these sample-to-sample alignment methods, our robust distillation method emphasizes a statistics-level robustness transfer that considers the properties of sample sets rather than individual samples. Thus, we align the covariance and gram matrices of the student toward those of the teacher for a more effective robust knowledge transfer.

3 Proposed Method

Below, we introduce the background of our work and propose our robust knowledge distillation method, STARSHIP, motivated by observations in Figure 1.

Background. Adversarially robust knowledge distillation generally requires an adversarially pre-trained teacher obtained via adversarial training from scratch. Let the network backbone and the classifier head be denoted as $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}^d$ and $g_{\boldsymbol{\theta}'} : \mathbb{R}^d \to \mathbb{R}^c$, given network parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, respectively. Let d be the backbone feature dimension, and C be the number of categories. Adversarial

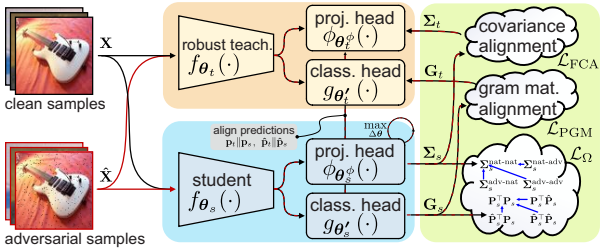


Fig. 2: Our STARSHIP pipeline. For brevity, assume we have a set of clean samples $\mathbf{X}$ and adversarial samples $\hat{\mathbf{X}}$. The student and the teacher backbones are equipped with the projection head and the classification head. Apart from aligning predictions which is a mere sample-to-sample strategy, in the green box we form covariance matrices $\boldsymbol{\Sigma}_t$ and $\boldsymbol{\Sigma}_s$, and align $\boldsymbol{\Sigma}_s$ toward $\boldsymbol{\Sigma}_t$ by $\mathcal{L}_{\text{FCA}}$. We also form prediction score based gram matrices $\mathbf{G}_t$ and $\mathbf{G}_s$, and align $\mathbf{G}_s$ toward $\mathbf{G}_t$ by $\mathcal{L}_{\text{PGM}}$. Moreover, $\mathcal{L}_{\Omega}$ aligns adversarial-adversarial, natural-adversarial and adversarial-natural sub-matrices of $\boldsymbol{\Sigma}_s$ and $\mathbf{G}_s$ with the natural-natural sub-matrices. $\boldsymbol{\Delta\theta}$ denotes parameter perturbations.

training [37] on a dataset from distribution $\mathcal{D}$ solves the following min-max game:

$$\min_{\boldsymbol{\theta},\boldsymbol{\theta}'} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\max_{\|\boldsymbol{\delta}\|_\infty<\epsilon} \mathcal{L}_{\text{CE}}\Big(g_{\boldsymbol{\theta}'}\big(f_{\boldsymbol{\theta}}(\mathbf{x}+\boldsymbol{\delta})\big), y\Big)\right], \tag{1}$$

where the perturbation $\boldsymbol{\delta}$ forms the adversarial sample $\hat{\mathbf{x}}=\mathbf{x}+\boldsymbol{\delta}$ restricted within radius ϵ of the original sample $\mathbf{x}$. The inner optimization maximizes the Cross-Entropy (CE) loss $\mathcal{L}_{\text{CE}}$ to find the worst-case adversarial sample posing misclassification, while the outer minimization optimizes the empirical risk on these adversaries. In the context of adversarially robust knowledge distillation, we generate adversarial samples by maximizing the prediction discrepancy between the teacher model $g_{\boldsymbol{\theta}'_t}\big(f_{\boldsymbol{\theta}_t}(\cdot)\big)$ and the student model $g_{\boldsymbol{\theta}'_s}\big(f_{\boldsymbol{\theta}_s}(\cdot)\big)$ as follows:

$$\hat{\mathbf{x}}^{t+1}=\Pi_{\mathbb{B}(\mathbf{x},\epsilon)}\Big[\hat{\mathbf{x}}^t+\alpha\ \text{sign}\Big(\nabla_{\hat{\mathbf{x}}^t}\mathcal{L}_{\text{KL}}\Big(g_{\boldsymbol{\theta}'_t}\big(f_{\boldsymbol{\theta}_t}(\mathbf{x})\big)\Big\|g_{\boldsymbol{\theta}'_s}\big(f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^t)\big)\Big)\Big)\Big], \tag{2}$$

where $\mathcal{L}_{\text{KL}}(\cdot\|\cdot)$ is the Kullback–Leibler (KL) divergence, α is the step size, $\hat{\mathbf{x}}^t$ is an adversarial sample at the t^{th} iteration in an m-step generation, and $\hat{\mathbf{x}}=\hat{\mathbf{x}}^m$ is the final adversarial sample. As Zi *et al.* [59], we use soft labels of clean samples predicted by the teacher model to guide the adversarial sample generation.

In addition to adversary generation via Eq. (2), we also consider its adaptive variant following [25], which integrates the gradient flow of the teacher model into the adversary generation, albeit at a higher computational cost:

$$\hat{\mathbf{x}}^{t+1}_{\text{ada}}=\Pi_{\mathbb{B}(\mathbf{x},\epsilon)}\Big[\hat{\mathbf{x}}^t_{\text{ada}}+\alpha\ \text{sign}\Big(\nabla_{\hat{\mathbf{x}}^t_{\text{ada}}}\mathcal{L}_{\text{KL}}\Big(g_{\boldsymbol{\theta}'_t}\big(f_{\boldsymbol{\theta}_t}(\hat{\mathbf{x}}^t_{\text{ada}})\big)\Big\|g_{\boldsymbol{\theta}'_s}\big(f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^t_{\text{ada}})\big)\Big)\Big)\Big], \tag{3}$$

where $\hat{\mathbf{x}}^t_{\text{ada}}$ denotes the t^{th} iteration during adaptive adversary generation. The primary distinction of Eq. (3) from Eq. (2) is the use of adaptive predictions from the teacher model for iterative adversarial samples, which can be regarded as adversaries with the worst-case prediction alignment. While Eq. (3) brings performance gains, it implicitly poses an increased computational cost.

3.1 Adversarially Robust Knowledge Distillation

In contrast to vanilla knowledge distillation, which aligns only clean samples, robust knowledge distillation additionally incorporates adversarial samples into knowledge transfer. Existing works primarily rely on fixed references (*e.g.*, ground-truth labels or the teacher's predictions on a certain type of samples) to guide the distillation process [25, 58, 59]. In contrast, we treat adversarial samples not just as noise but as specialized augmentations that retain the visual semantic content of their original counterpart. For brevity, denote the teacher's predictions on clean samples as $\mathbf{p}_t = g_{\boldsymbol{\theta}'_t}\big(f_{\boldsymbol{\theta}_t}(\mathbf{x})\big)$ and on adversarial samples as $\hat{\mathbf{p}}_t = g_{\boldsymbol{\theta}'_t}\big(f_{\boldsymbol{\theta}_t}(\hat{\mathbf{x}})\big)$. For the student model, let $\mathbf{p}_s = g_{\boldsymbol{\theta}'_s}\big(f_{\boldsymbol{\theta}_s}(\mathbf{x})\big)$ and $\hat{\mathbf{p}}_s = g_{\boldsymbol{\theta}'_s}\big(f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}})\big)$ be the student's predictions on clean and adversarial samples. Then, the baseline loss used in our work, *Adversarially Robust Knowledge Distillation* (ARKD), is given as:

$$\mathcal{L}_{\text{ARKD}} = \underbrace{(1-\beta)\,\mathcal{L}_{\text{KL}}\big(\mathbf{p}_t \| \mathbf{p}_s\big)}_{\text{align clean predictions}} \quad + \quad \underbrace{\beta\,\mathcal{L}_{\text{KL}}\big(\hat{\mathbf{p}}_t \| \hat{\mathbf{p}}_s\big)}_{\text{align adversarial predictions}}, \tag{4}$$

where $\beta \in [0, 1]$ balances the focus of distillation between the clean samples and their adversarial counterparts. By adopting the ARKD loss, the student model can effectively probe responses of the adversarially pre-trained teacher model to learn its implicit robust behavior against adversarial samples.

3.2 Clean and Adversarial Feature Covariance Alignment

Following the motivation from Figure 1 and our hypothesis that reducing the feature variance gap between the student and teacher models can facilitate distilling the robust behavior from the teacher model, we attach feature projection heads $\phi_{\boldsymbol{\theta}^{\phi}_s}(\cdot)$ and $\phi_{\boldsymbol{\theta}^{\phi}_t}(\cdot)$ with parameters $\boldsymbol{\theta}^{\phi}_s$ and $\boldsymbol{\theta}^{\phi}_t$ to the student and teacher models, respectively. Let $\boldsymbol{\phi}_t(\mathbf{x}; \boldsymbol{\theta}^{\phi}_t) \equiv \phi_{\boldsymbol{\theta}^{\phi}_t}\big(f_{\boldsymbol{\theta}_t}(\mathbf{x})\big)$ and $\boldsymbol{\phi}_t(\hat{\mathbf{x}}; \boldsymbol{\theta}^{\phi}_t) \equiv \phi_{\boldsymbol{\theta}^{\phi}_t}\big(f_{\boldsymbol{\theta}_t}(\hat{\mathbf{x}})\big)$ be the features obtained by combining the teacher projection head with the teacher backbone, given the clean sample $\mathbf{x}$ and its adversarial counterpart $\hat{\mathbf{x}}$.

Similarly, let $\boldsymbol{\phi}_s(\mathbf{x}; \boldsymbol{\theta}^{\phi}_s) \equiv \phi_{\boldsymbol{\theta}^{\phi}_s}\big(f_{\boldsymbol{\theta}_s}(\mathbf{x})\big)$ and $\boldsymbol{\phi}_s(\hat{\mathbf{x}}; \boldsymbol{\theta}^{\phi}_s) \equiv \phi_{\boldsymbol{\theta}^{\phi}_s}\big(f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}})\big)$ be the features obtained by combining the student projection head with the student backbone, given the clean sample $\mathbf{x}$ and its adversarial counterpart $\hat{\mathbf{x}}$.

Considering $\mathbf{X}$ and $\hat{\mathbf{X}}$ as the sets (*e.g.*, mini-batch) comprising clean and adversarial samples, respectively, organized as column vectors. Then, we simply denote the corresponding clean and adversarial features obtained from the teacher projection head as $\boldsymbol{\Phi}_t$ and $\hat{\boldsymbol{\Phi}}_t$, and from the student head as $\boldsymbol{\Phi}_s$ and $\hat{\boldsymbol{\Phi}}_s$. To highlight that these features depend on the parameters of projection heads, we denote $\boldsymbol{\Phi}_t | \boldsymbol{\theta}^{\phi}_t$ and $\hat{\boldsymbol{\Phi}}_t | \boldsymbol{\theta}^{\phi}_t$ for the teacher, and $\boldsymbol{\Phi}_s | \boldsymbol{\theta}^{\phi}_s$ and $\hat{\boldsymbol{\Phi}}_s | \boldsymbol{\theta}^{\phi}_s$ for the student.

We concatenate the feature sets $\boldsymbol{\Phi}_t$ and $\hat{\boldsymbol{\Phi}}_t$ along the channel dimension and compute covariance $\boldsymbol{\Sigma}_t | \hat{\boldsymbol{\Phi}}_t$ or $\boldsymbol{\Sigma}_t \in \mathbb{Q}^{2d \times 2d}$ for brevity. Symbol $\mathbb{Q}$ is the set of symmetric positive definite matrices. Similarly, we concatenate $\boldsymbol{\Phi}_s$ and $\hat{\boldsymbol{\Phi}}_s$ along the channel dimension and compute $\boldsymbol{\Sigma}_s | \hat{\boldsymbol{\Phi}}_s$, or equivalently, $\boldsymbol{\Sigma}_s \in \mathbb{Q}^{2d \times 2d}$. Note that the feature covariance matrices for the teacher and student models,

$$\boldsymbol{\Sigma}_t = \begin{bmatrix} \boldsymbol{\Sigma}_t^{\text{nat-nat}} & \boldsymbol{\Sigma}_t^{\text{nat-adv}} \\ \boldsymbol{\Sigma}_t^{\text{adv-nat}} & \boldsymbol{\Sigma}_t^{\text{adv-adv}} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_s = \begin{bmatrix} \boldsymbol{\Sigma}_s^{\text{nat-nat}} & \boldsymbol{\Sigma}_s^{\text{nat-adv}} \\ \boldsymbol{\Sigma}_s^{\text{adv-nat}} & \boldsymbol{\Sigma}_s^{\text{adv-adv}} \end{bmatrix}, \tag{5}$$

contain four types of interactions, *i.e.*, *nat-nat*, *nat-adv*, *adv-nat*, and *adv-adv*, which capture clean-clean, clean-adversarial, adversarial-clean, and adversarial-adversarial feature interactions, respectively. Then, the *Feature Covariance Alignment* (FCA), as a function of the projection parameters $\boldsymbol{\theta}_t^\phi$ and $\boldsymbol{\theta}_s^\phi$, is given by:

$$\mathcal{L}_{\text{FCA}}(\boldsymbol{\theta}_t^\phi, \boldsymbol{\theta}_s^\phi) = \frac{1}{4} d^2\big(\boldsymbol{\Sigma}_t|\boldsymbol{\theta}_t^\phi, \boldsymbol{\Sigma}_s|\boldsymbol{\theta}_s^\phi\big), \tag{6}$$

where $d^2(\cdot,\cdot)$ can be any suitable distance metrics between the covariance matrices, *e.g.*, the Frobenius norm, the Power-Euclidean (Pow-E) metric [17, 33, 34], the fast spectral expectation of Max-pooling (MaxExp) [33, 34], or a metric between Multivariate normal distributions, *e.g.*, the KL divergence (similar to [56,57]). The distance is scaled by 0.25, as $\boldsymbol{\Sigma}_t$ and $\boldsymbol{\Sigma}_s$ comprise four sub-matrices.

To improve the resilience of the learned feature-level statistical information and achieve better covariance alignment, we employ adversarial perturbations to optimize parameters of the projection heads, inspired by Wu *et al.* [51], who suggested that the robust generalization gap is bounded by the flatness of the parameter-loss landscape. We introduce parameter-level perturbations to induce a feature- and covariance-level disagreement between the teacher and student:

$$\min_{\boldsymbol{\theta}_t^\phi, \boldsymbol{\theta}_s^\phi} \max_{\Delta\boldsymbol{\theta} \in \mathcal{V}_s} \Big[(1-\beta) \Big\| \boldsymbol{\Phi}_t|\boldsymbol{\theta}_t^\phi - \boldsymbol{\Phi}_s|\big(\boldsymbol{\theta}_s^\phi + \Delta\boldsymbol{\theta}\big) \Big\|_F^2 + \beta \Big\| \hat{\boldsymbol{\Phi}}_t|\boldsymbol{\theta}_t^\phi - \hat{\boldsymbol{\Phi}}_s|\big(\boldsymbol{\theta}_s^\phi + \Delta\boldsymbol{\theta}\big) \Big\|_F^2 + \mathcal{L}_{\text{FCA}}\big(\boldsymbol{\theta}_t^\phi, \boldsymbol{\theta}_s^\phi + \Delta\boldsymbol{\theta}\big) \Big], \tag{7}$$

where $\Delta\boldsymbol{\theta}$ denotes the parameter-level perturbations, and $\mathcal{V}_s$ is the perturbation region set $\big\{\Delta\boldsymbol{\theta} \in \mathcal{V}_s : \|\Delta\boldsymbol{\theta}\|_F \leq \eta \|\boldsymbol{\theta}_s^\phi\|_F\big\}$ for the perturbation intensity η. The parameter-level perturbation is then optimized using iterative gradient ascent.

3.3 Matching Gram Matrices for Prediction Scores

As evidenced by Figure 1b, there exists a direct correlation between enhancing robust performance and reducing the variance gap across prediction score-based gram matrices derived from adversarial and clean samples. Inspired by this nearly linear negative correlation of robustness with the gap, we propose to align the prediction score-based student's gram matrix towards the teacher's counterpart. In analogy to Section 3.2, given clean and adversarial samples $\mathbf{X}$ and $\hat{\mathbf{X}}$ (represented as column vectors), we obtain prediction score matrices $\mathbf{P}_t$ and $\hat{\mathbf{P}}_t$ by stacking prediction score column vectors ($\mathbf{p}_t$ and $\hat{\mathbf{p}}_t$) obtained from the teacher's prediction head, as detailed in Section 3.1. By analogy, we also obtain prediction score matrices $\mathbf{P}_s$ and $\hat{\mathbf{P}}_s$ based on outputs of the student's prediction head for B samples. We form the gram matrices for the teacher and the student:

$$\mathbf{G}_t = \frac{1}{C} \begin{bmatrix} \mathbf{P}_t^\top \mathbf{P}_t & \mathbf{P}_t^\top \hat{\mathbf{P}}_t \\ \hat{\mathbf{P}}_t^\top \mathbf{P}_t & \hat{\mathbf{P}}_t^\top \hat{\mathbf{P}}_t \end{bmatrix} \quad \text{and} \quad \mathbf{G}_s = \frac{1}{C} \begin{bmatrix} \mathbf{P}_s^\top \mathbf{P}_s & \mathbf{P}_s^\top \hat{\mathbf{P}}_s \\ \hat{\mathbf{P}}_s^\top \mathbf{P}_s & \hat{\mathbf{P}}_s^\top \hat{\mathbf{P}}_s \end{bmatrix}. \tag{8}$$

Modeling sample-to-sample relations does not take the spectrum of gram matrices into account. Thus, we propose *Prediction-score Gram Matching* (PGM):

$$\mathcal{L}_{\text{PGM}} = \frac{1}{4} d^2\big(\mathbf{G}_t, \mathbf{G}_s\big), \tag{9}$$

which aligns the student's gram matrix with the teacher's gram matrix. Let $d^2(\cdot,\cdot)$ be any suitable distance measure (as detailed in Section 3.2) between symmetric positive definite matrices $\mathbf{G}_t \in \mathbb{Q}^{2B\times 2B}$. $\mathbf{G}_s \in \mathbb{Q}^{2B\times 2B}$, given B clean samples and their B adversarial counterparts in a batch of data.

3.4 Aligning Adversarial Statistics with Clean Statistics

Drawing insights from the standard adversarially robust distillation that aligns the student's representations of adversarial samples with the teacher's (or even student's) representations of clean samples, the analogous alignment can be performed between covariance sub-matrices or gram sub-matrices. Let a sub-matrix extractor $\psi_{ij}(\cdot)$ split a matrix into two halves along the rows and two halves along the columns. For $\mathbf{\Sigma}_s$ in Eq. (5), operations $\psi_{11}(\cdot)$, $\psi_{12}(\cdot)$, $\psi_{21}(\cdot)$ and $\psi_{22}(\cdot)$ return $\mathbf{\Sigma}_s^{\text{nat-nat}}$, $\mathbf{\Sigma}_s^{\text{nat-adv}}$, $\mathbf{\Sigma}_s^{\text{adv-nat}}$, and $\mathbf{\Sigma}_s^{\text{adv-adv}}$, respectively, as defined in Eq. (5). $\psi_{ij}(\cdot)$ acts on $\mathbf{G}_s$ by analogy. Define a set of index pairs $\mathcal{I} = \big\{(1,2),(2,1),(2,2)\big\}$. The loss aligning sub-matrices of adversarial-adversarial and clean-adversarial statistical interactions toward clean-clean interactions is thus defined as:

$$\mathcal{L}_{\Omega} = \frac{1}{2}\sum_{(i,j)\in\mathcal{I}} \big\|\psi_{11}(\mathbf{\Sigma}_s), \psi_{ij}(\mathbf{\Sigma}_s)\big\|_F^2 + \big\|\psi_{11}(\mathbf{G}_s), \psi_{ij}(\mathbf{G}_s)\big\|_F^2. \tag{10}$$

The above alignment function leverages the Frobenius norm rather than non-Euclidean distances, as generally, sub-matrices extracted by $\psi_{12}(\cdot)$ and $\psi_{21}(\cdot)$ are non-symmetric indefinite (they become symmetric positive definite if they fully converge to $\psi_{11}(\cdot)$). Taking $\mathbf{\Sigma}_s$ as an example, the interpretation of the above loss is as follows: term $\psi_{12}(\cdot)$ (and $\psi_{21}(\cdot)$) captures correlations across feature channels of the clean samples and their adversarial counterparts.

Discussion. The loss functions $\mathcal{L}_{FCA}$ and $\mathcal{L}_{PGM}$ contain highly complementary statistics. Observe that even for perfectly aligned $\mathbf{\Sigma}_s = \mathbf{\Sigma}_t$, consider N zero-centered feature vectors $\boldsymbol{\phi}_s^{(n)}$ and $\boldsymbol{\phi}_t^{(n)}$, which were used to compute their covariance matrices, and note that the corresponding feature vectors are not aligned at all, *i.e.*, $\sum_{n=1,\ldots,N} \boldsymbol{\phi}_s^{(n)} {\boldsymbol{\phi}_s^{(n)}}^\top = \sum_{n=\Pi(1,\ldots,N)} \boldsymbol{\phi}_t^{(n)} {\boldsymbol{\phi}_t^{(n)}}^\top, \forall \Pi$, where $\Pi(1,\ldots,N)$ is a random permutation of the sample indexes $\{1,\ldots,N\}$. Indeed, covariance is invariant to the sample order. In contrast, for two matched gram matrices $\mathbf{P}_s = \mathbf{P}_t$, we know that assuming classification scores lie on the ℓ_1 simplex, then $\sum_{c=1,\ldots,C} {\boldsymbol{P}_s^{(c,:)}}^\top \boldsymbol{P}_s^{(c,:)} = \sum_{c=\Pi(1,\ldots,C)} {\boldsymbol{P}_t^{(c,:)}}^\top \boldsymbol{P}_t^{(c,:)}, \forall \Pi$, where $\Pi(1,\ldots,C)$ is a random permutation of class indexes $\{1,\ldots,C\}$. Thus, $\mathcal{L}_{FCA}$ is invariant to the sample order, whereas $\mathcal{L}_{PGM}$ is invariant to the class order.

3.5 Objective Function

Below, we formalize the objective function of our STARSHIP by combining individual loss components introduced in the preceding sections as follows:

$$\mathcal{L} = \mathcal{L}_{\text{ARKD}} + \lambda_1 \mathcal{L}_{\text{FCA}} + \lambda_2 \mathcal{L}_{\text{PGM}} + \lambda_3 \mathcal{L}_{\Omega}, \tag{11}$$

Table 1: CIFAR-10 and CIFAR-100: Comparisons of our STARSHIP with other adversarially robust knowledge distillation methods when **distilled from a large-scale WRN-28 teacher model**. Adversarial perturbations are restricted within the ℓ_∞-norm radius $\epsilon = 8/255$. We report both clean and robust accuracies (%).

Type	Architecture	Method		CIFAR-10				CIFAR-100			
				Clean	PGD	CW	AA	Clean	PGD	CW	AA
Teacher	WRN-28	SCORE	[41]	88.61	64.95	61.79	61.03	63.64	35.46	32.14	31.13
Student	ResNet-18	ARD	[19]	84.35	53.21	51.58	49.40	58.20	30.87	27.97	26.02
		IAD	[58]	83.46	53.34	51.69	49.09	57.35	31.11	28.12	26.22
		RSLAD	[59]	84.42	54.52	53.46	51.36	57.97	32.57	29.28	27.52
		AKD	[38]	86.04	53.95	52.39	50.11	60.79	31.37	28.85	26.93
		AdaAD	[25]	86.38	56.23	54.16	52.36	61.26	32.24	29.19	27.46
		STARSHIP		86.47	57.45	55.50	53.78	61.54	**34.45**	**30.75**	**29.30**
		Ada-STARSHIP		**87.04**	**58.30**	**56.03**	**54.47**	**62.19**	33.52	30.24	28.28
	MNV2	ARD	[19]	82.10	52.82	50.80	48.44	57.26	30.77	27.96	25.79
		IAD	[58]	81.62	52.77	50.61	48.53	56.88	30.54	27.59	25.69
		RSLAD	[59]	84.33	53.98	52.37	50.38	59.38	31.32	28.38	26.54
		AKD	[38]	84.52	51.80	50.29	48.09	59.32	30.13	27.61	25.46
		AdaAD	[25]	85.45	53.16	51.25	49.49	59.97	31.31	28.69	26.49
		STARSHIP		85.71	55.95	53.79	51.85	60.25	**34.20**	**30.72**	**28.57**
		Ada-STARSHIP		**86.44**	**56.39**	**54.87**	**52.62**	**61.33**	32.89	29.57	27.69

where λ_1, λ_2, and λ_3 are loss weighting factors. Generally, $\lambda_1 = 1$ in all experiments, as $\mathcal{L}_{\text{FCA}}$ is a covariance-based alignment equivalent of the sample-to-sample alignment $\mathcal{L}_{\text{ARKD}}$ (which has the default weight of 1). Moreover, as $\mathcal{L}_\Omega$ operates on the student only by encouraging the "adversarial parts" of matrices $\mathbf{\Sigma}_s$ and $\mathbf{G}_s$ to align with their "non-adversarial parts", we set $\lambda_3 = 1$ in all experiments, and we only vary λ_2 to ensure a desired level of aligning the student's statistics with the teacher's statistics. We optimize the network parameters of the student model by minimizing $\mathcal{L}$. In inference, we use the distilled student.

4 Experiments

In this section, we provide our experimental settings and compare our STARSHIP method with other adversarially robust knowledge distillation approaches.

Datasets. We conduct all the experiments on three standard image classification datasets: CIFAR-10, CIFAR-100 [35], and ImageNet-100 [9]. Further details of these datasets can be found in Appendix A.1.

Implementation details. Following previous works [25,58,59] & RobustBench [7], we adopt ResNet-18/34 [22], MobileNetV2 (MNV2) [46], and Wide-ResNet-28-10 (WRN-28) [53] as the teacher and student models. We also adopt the Vision Transformer (ViT) as the teacher architecture. Unless specified otherwise, we conduct adversarial sample generation based on the ℓ_∞-norm threat model with the perturbation radius $\epsilon = 8/255$. More experimental details are included in Appendix A.2. In all the experiments, we adopt the loss weighting factors $\lambda_1 = \lambda_3 = 1.0$. We determine the hyper-parameters $\lambda_2 = 2.0$ and $\beta = 0.8$ through cross-validation on CIFAR-10 and consistently apply them across all datasets without modifications. We provide hyper-parameter evaluations in Appendix E.

Table 2: CIFAR-10 and CIFAR-100: Comparisons of our STARSHIP with other adversarially robust knowledge distillation methods in the **self-distillation** setting. We report both clean and robust accuracies (%).

Type	Architecture	Method		CIFAR-10				CIFAR-100			
				Clean	PGD	CW	AA	Clean	PGD	CW	AA
Teacher	ResNet-18	TRADES	[55]	82.45	52.21	50.29	48.90	56.37	28.68	24.87	23.78
Student	ResNet-18	ARD	[19]	81.64	52.62	51.35	49.19	57.96	31.34	27.84	26.13
		IAD	[58]	80.66	52.63	52.21	48.90	56.45	31.87	28.00	26.66
		RSLAD	[59]	81.30	53.80	52.32	50.78	55.17	31.21	27.82	26.46
		AKD	[38]	82.30	52.84	51.39	49.71	56.37	30.02	26.48	25.44
		AdaAD	[25]	82.34	52.75	51.26	49.92	56.62	29.62	26.11	24.78
		STARSHIP		81.97	**55.72**	**54.06**	**52.42**	57.60	**32.19**	**28.19**	**27.05**
		Ada-STARSHIP		**82.62**	55.05	53.40	51.89	**58.09**	31.95	28.07	26.92
Teacher	MNV2	TRADES	[55]	81.04	50.87	48.46	47.15	54.11	27.28	23.39	22.36
Student	MNV2	ARD	[19]	81.25	53.02	50.69	48.85	55.64	30.93	27.47	26.05
		IAD	[58]	79.36	53.45	50.93	49.14	54.00	31.01	27.59	26.11
		RSLAD	[59]	80.01	53.35	51.04	49.74	53.52	29.95	26.66	25.47
		AKD	[38]	80.86	52.74	50.60	49.13	54.26	28.99	25.46	24.31
		AdaAD	[25]	80.48	50.80	48.14	46.98	53.97	28.31	24.69	23.51
		STARSHIP		80.97	**54.28**	51.58	**50.46**	56.21	**32.03**	**27.98**	**26.74**
		Ada-STARSHIP		**81.45**	54.17	**51.63**	50.28	**56.70**	31.79	27.85	26.36

4.1 Results

Robust distillation from a larger teacher. We compare our STARSHIP method with the state-of-the-art adversarially robust knowledge distillation approaches in Table 1. We report the classification accuracies on clean samples and their adversarial counterparts generated by three standard adversarial attack approaches: 20-step Projected Gradient Descent (PGD) [37] with a fixed step size $\alpha = 2/255$, CW [3], and Auto-Attack (AA) [8]. For a fair comparison, all the results are obtained by robust knowledge distillation from the same WRN-28 teacher model. Table 1 shows that our STARSHIP and its adaptive variant achieve the best accuracy on both clean and adversarial samples. Notably, the incorporation of the adaptive adversary generation strategy in Ada-STARSHIP leads to further improvements in natural performance. The superior performance on both ResNet-18 [22] and MNV2 [46] showcases the versatility of our method.

Robust self-distillation. In self-distillation setting, the teacher and student architectures are identical. Table 2 shows that our STARSHIP and its adaptive extension, Ada-STARSHIP, outperform other robust knowledge distillation methods across CIFAR-10/100. In addition, our method achieves comparable or even better performance on clean samples than the teacher model. We attribute such a gain to our $\mathcal{L}_\Omega$ (Eq. (10)), which facilitates an intrinsic self-distill within the student, *i.e.*, sub-matrices in this loss entirely operate on the student backbone. Given the teacher backbone in this experiment is not stronger than the student backbone, such an internal self-distillation enables the student model to autonomously refine its decision boundaries. More experimental results w.r.t. diverse settings of $\mathcal{L}_\Omega$ can be found in Appendix C.4.

Robust distillation on ImageNet-100. Below, we extend our investigation to evaluate the generalization ability of our STARSHIP method in the context

Table 3: ImageNet-100: Robust accuracy (%) on the distilled student models using ResNet-18 and MNV2 backbones.

Type	Architecture	Method		Clean	PGD	AA
Teacher	ResNet-34	TRADES	[55]	72.66	40.88	34.70
Student	ResNet-18	ARD	[19]	65.41	38.34	30.94
		IAD	[58]	65.62	39.09	32.63
		RSLAD	[59]	66.60	39.12	32.18
		AKD	[38]	67.00	38.85	31.56
		AdaAD	[25]	68.64	38.99	32.24
		STARSHIP		69.68	**40.08**	**33.18**
		Ada-STARSHIP		**71.10**	39.86	32.74
	MNV2	ARD	[19]	65.90	37.28	30.20
		IAD	[58]	64.96	38.00	31.40
		RSLAD	[59]	65.82	37.86	31.66
		AKD	[38]	66.09	36.47	30.03
		AdaAD	[25]	67.26	37.16	30.55
		STARSHIP		68.20	**39.84**	**32.87**
		Ada-STARSHIP		**69.74**	39.21	32.58

Table 4: Robust accuracy (%) of models distilled from ViTs using ResNet-18 and MNV2 student backbones on CIFAR-10.

Type	Architecture	Method		Clean	PGD	AA
Teacher	ViT-B	AT-PRM	[40]	83.98	53.10	49.66
Student	ResNet-18	ARD	[19]	82.76	52.95	49.03
		IAD	[58]	82.27	53.42	49.48
		RSLAD	[59]	82.33	54.89	49.74
		AKD	[38]	82.86	53.44	49.26
		AdaAD	[25]	82.51	54.30	50.02
		STARSHIP		83.53	**55.89**	**52.07**
		Ada-STARSHIP		**83.60**	55.31	51.49
Teacher	DeiT-S	AT-PRM	[40]	82.68	52.47	49.27
Student	MNV2	ARD	[19]	81.59	53.45	49.20
		IAD	[58]	80.41	54.12	49.62
		RSLAD	[59]	80.86	53.91	50.18
		AKD	[38]	81.62	53.08	49.02
		AdaAD	[25]	82.11	53.85	49.57
		STARSHIP		83.45	**55.04**	51.48
		Ada-STARSHIP		**83.52**	54.92	**51.54**

Table 5: Robust distillation (WRN-28 → ResNet-18) on CIFAR-10/100 **with auxiliary synthetic training data**. We report both clean and (Auto-Attack) robust accuracies (%) associated with the robustness gain Δ_{Rob} with additional data.

Type	Method		CIFAR-10			CIFAR-100		
			Clean	Robust	Δ_{Rob}	Clean	Robust	Δ_{Rob}
Teacher	SCORE	[41]	88.61	61.03	—	63.64	31.13	—
Student	ARD	[19]	83.93	51.04	+1.64	57.36	26.20	+0.18
	IAD	[58]	83.16	50.30	+1.21	56.13	26.54	+0.32
	RSLAD	[59]	83.79	52.42	+1.06	56.17	27.84	+0.32
	AKD	[38]	85.50	51.34	+1.23	58.95	27.26	+0.33
	AdaAD	[25]	85.93	53.08	+0.72	60.54	27.62	+0.16
	STARSHIP		85.67	54.49	+0.71	60.68	**29.43**	+0.13
	Ada-STARSHIP		**86.52**	**55.75**	+0.83	**61.31**	28.63	+0.35

Table 6: Extension of robust distillation (WRN-28 → ResNet-18/MNV2) with **single-step adversary strategy** (N-FGSM [29]) on CIFAR-10. We report accuracy (%) with the average training time per epoch.

Type	Architecture	Method		Clean	Robust	Time (s)
Teacher	WRN-28	SCORE	[41]	88.61	61.03	—
Student	ResNet-18	IAD	[58]	83.87	46.21	68
		RSLAD	[59]	84.74	48.71	42
		AdaAD	[25]	85.50	50.22	112
		STARSHIP		86.68	51.06	50
		Ada-STARSHIP		**87.49**	**51.90**	121
	MNV2	IAD	[58]	81.62	45.53	77
		RSLAD	[59]	84.55	46.69	51
		AdaAD	[25]	84.49	47.37	126
		STARSHIP		86.22	48.65	61
		Ada-STARSHIP		**86.76**	49.17	138

of larger-scale and practical images. As presented in Table 3, we perform robust knowledge distillation on ImageNet-100 [9] using different student backbones. We can observe that our STARSHIP method is superior to existing robustness transfer methods by retaining more comprehensive knowledge inherited from the teacher model in terms of clean accuracy and adversarial robustness.

Robust distillation from a ViT-based teacher. ViT has emerged as a competitive alternative to the convolutional network, showcasing good adversarial robustness [1, 40]. Thus, we investigate if the intrinsic robustness of ViTs can also be transferred to lightweight student backbones. Table 4 shows that our proposed STARSHIP consistently achieves superior robust accuracy that even surpasses the ViT-based teacher model. Such a robustness improvement also demonstrates the viability of our method in the context of inheriting robustness from ViT-based teachers without compromising natural performance.

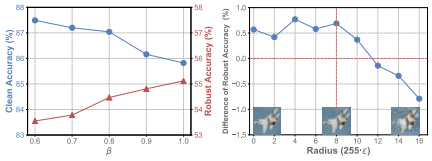
Robust distillation with additional synthetic data. Auxiliary data generated by Denoising Diffusion Probabilistic Models (DDPMs) [24] can further improve adversarial training. Nevertheless, there exists a gap in applying the auxiliary data in the context of robust knowledge distillation. To bridge this gap, we incorporate an additional 1M synthetic training data for CIFAR-10/100

Table 7: Ablation study (WRN-28 → ResNet-18) of three main components in our STARSHIP on CIFAR-10/100.

	FCA	PGM	Ω	CIFAR-10			CIFAR-100		
				Clean	PGD	AA	Clean	PGD	AA
1				84.13	54.49	51.27	58.52	32.43	27.23
2	✓			84.56	56.94	52.43	58.83	33.58	28.60
3	✓	✓		85.35	57.28	53.12	59.79	33.94	28.93
4	✓		✓	86.66	56.80	52.36	61.20	33.46	28.38
	✓	✓	✓	86.47	57.45	53.78	61.54	34.45	29.30

Table 8: Comparison of different distance metrics for our STARSHIP during robust knowledge distillation (WRN-28 → ResNet-18) on CIFAR-10/100.

Metric	CIFAR-10			CIFAR-100		
	Clean	PGD	AA	Clean	PGD	AA
Frobnius norm	85.15	56.03	52.10	59.66	33.23	28.16
Pow-E	86.14	56.85	52.79	60.27	33.51	28.54
MaxExp	**86.47**	**57.45**	**53.78**	**61.54**	**34.45**	**29.30**
KL Divergence	86.32	57.06	53.16	60.49	33.92	28.87



(a) (b)

Fig. 3: Robust distillation on CIFAR-10 (WRN-28 → ResNet-18). Fig. 3a: Trade-off between clean and (Auto-Attack) robust accuracies w.r.t. β that balances distillation between clean and adversarial samples. Fig. 3b: Difference of robust accuracy between our Ada-STARSHIP and STARSHIP under diverse perturb. radii ϵ.

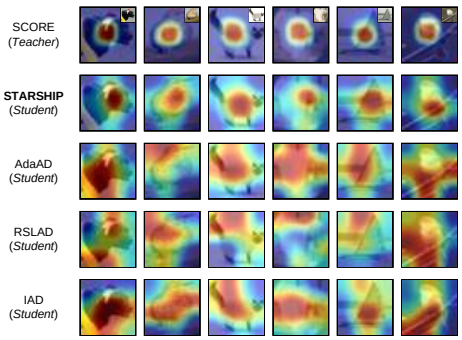


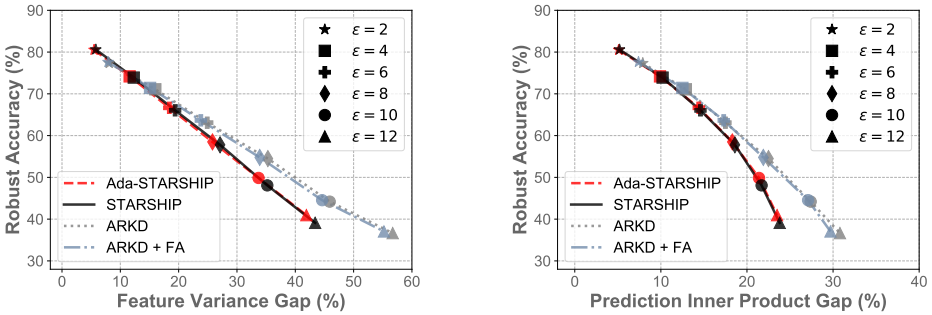
Fig. 4: Visualizations of attention of both the teacher and student models distilled via different robust distillation methods.

when distilling from a large-scale teacher model. Table 5 shows that the auxiliary training data improves the adversarial robustness of distilled students compared to the original training in Table 1. STARSHIP achieves the most significant gain in robustness by incorporating auxiliary DDPM-generated data during robust knowledge distillation.

Single-step robust distillation. Due to the multi-step adversary generation scheme, achieving adversarial robustness requires several times more computation resources than the standard training [18]. To mitigate such a computational burden, we explore the adversarially robust knowledge distillation with the single-step adversary generation strategy [29]. Table 6 shows that our STARSHIP can efficiently inherit non-trivial robustness from the teacher model based on single-step adversarial samples. Further details and more experiments can be found in Appendix B.2 and Appendix C.5, respectively.

5 Further Analyses

Ablation studies. Below, we investigate modules of STARSHIP: (i) Feature Covariance Alignment (FCA) in Sec. 3.2, (ii) Prediction-score Gram Matching (PGM) in Eq. (9), and (iii) Statistical Sub-matrices Alignment (Ω) in Eq. (10).



(a) Gap between Feature-wise Variances. **(b)** Gap between Pred. Gram-wise Variances.

Fig. 5: The robust accuracy (%) of several student models w.r.t. their average gap (%) between feature-wise variances evaluated on the adversarial samples and clean samples (5a). In Fig. 5b, we show equivalent evaluations with the use of prediction-based gram-wise variances. The variance gaps are captured under several adversarial perturbation radii. Note that the experimental settings are the same as Fig. 1.

Table 7 shows results on CIFAR-10. Our baseline (first row) uses the standard prediction alignment in Eq. (4). Both FCA and PGM improve adversarial robustness. The statistical sub-matrices alignment (Ω) boosts the clean accuracy of student by promoting the prediction invariance between clean and adv. samples.

Clean *vs*. robust accuracy trade-off. The trade-off between clean accuracy and adversarial robustness has been widely investigated in adversarial training [12,43,55]. Less is known about such a trade-off within the context of robust knowledge distillation. Thus, we study the effect of β, which balances the prediction alignment on clean and adversarial samples. Fig. 3a shows an improved adversarial robustness associated with a decrease in clean performance as β gets larger. Furthermore, we explore the robustness of our STARSHIP and its adaptive variant (Ada-STARSHIP) against adversarial samples across diverse attack strengths in Fig. 3b. We adopt adversaries with $\epsilon = 8/255$ (signified by the dashed red line) for robust distillation. Compared with the vanilla STARSHIP, Ada-STARSHIP has a better performance on clean samples and their adversarial counterparts of lower attack strength ($\epsilon \leq 10$). In comparison, STARSHIP achieves better robustness against strong adversarial samples ($\epsilon > 10$). Such an intrinsic robustness trade-off makes our method applicable in diverse scenarios.

Visualization. As shown in Figure 4, we provide visualizations of adversarial samples ($\epsilon = 8/255$) related to both the teacher and student models via Grad-CAM [48]. In comparison with other robust distillation methods, the student model distilled by our STARSHIP method shares similar attention regions to that of the teacher model. Such an observation further underscores the efficacy of our method in preserving the adversarially robust prediction alignment between the teacher and student. Additional visualization results are in Appendix D.

Discussion on effectiveness of our method. Building on the insights from Figure 1, we further analyze the robustness w.r.t. the variance gap between adversarial and clean representations under different attack strengths. Figure 5

evaluates our STARSHIP method and its adaptive variant. For comparison, we also introduce a Feature Alignment (FA) variant of our baseline method (ARKD) from Eq. (4). According to Figure 5, relying on the feature or prediction alignment alone cannot reduce such a statistical gap, leading to relatively weaker robustness against adversarial samples. Unlike the sample-to-sample matching paradigm (*e.g.*, FA), our STARSHIP method introduces the alignment of second-order statistics [31, 32, 34, 36, 44, 54] at both the feature and prediction levels, which introduces the multivariate Normal distribution alignment prior to the alignment process (*e.g.*, correlation alignment instead of mere matching of individual features or instances). Analysis of the parameter-level perturbations (Eq. (7)) is in Appendix C.6.

Performance comparison w.r.t. different distance measures. Below, we explore the efficacy of several distance metrics on the alignment of covariance and gram matrices between the teacher and student distilled via our STARSHIP. As such matrices are symmetric positive (semi-)definite, distance $d^2(\cdot,\cdot)$ in Eq. (6) for feature covariance alignment and Eq. (9) for prediction-score gram matching can be achieved by non-Euclidean distances. Table 8 reports clean and robust accuracies of the distilled student model based on different distance metrics: (i) the Frobenius norm, (ii) the Power-Euclidean (Pow-E) metric [17,33,34], (iii) fast spectral expectation of Max-pooling (MaxExp) [33,34], and (iv) the KL divergence. Note that we conduct the alignment of multivariate Normal distributions (statistics) between the teacher and student models when adopting the KL divergence as the distance metric. We observe that the statistical alignment via MaxExp achieves the highest accuracy on both clean and adversarial examples, even surpassing the performance of distribution alignment via the KL divergence that adopts both the mean and covariance information. The main reason is that MaxExp can balance (and partially equalize) the spectrum of the aligned statistics. It dampens the significance of leading eigenvalues and boosts the impact of non-leading eigenvalues. Thus, it encourages the student to distill all orthogonal variance vectors of the teacher covariance, not only those corresponding to the leading eigenvalues, preventing overfitting to leading principal directions.

6 Conclusions

Motivated by the implicit link between feature variance and model generalization, we have investigated the relation between the adversarial robustness and its correlation to the variance gap between feature variances of adversarial and clean samples. We have observed that a similar phenomenon holds for prediction-based gram matrices. We devised several alignment strategies leveraging second-order statistics to align the student's statistics with the teacher's statistics. We have also shown that a degree of self-distillation within the student model is beneficial for robustness. Leveraging second-order statistics to align the student toward the teacher helps capture well intricacies of the teacher's robust boundary through aligning correlations rather than sample-to-sample scores. Appendix F discusses limitations of our method.

Acknowledgments

This research is supported by National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative, the Centre for Frontier Artificial Intelligence Research, Institute of High Performance Computing, A*Star, and the College of Computing and Data Science at Nanyang Technological University. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, and Infocomm Media Development Authority. PK is supported by CSIRO's Science Digital.

References

1. Bai, Y., Mei, J., Yuille, A.L., Xie, C.: Are transformers more robust than cnns? Advances in neural information processing systems **34**, 26831–26843 (2021)
2. Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R., Kolesnikov, A.: Knowledge distillation: A good teacher is patient and consistent. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10925–10934 (2022)
3. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. IEEE (2017)
4. Chen, D., Mei, J.P., Zhang, H., Wang, C., Feng, Y., Chen, C.: Knowledge distillation with the reused teacher classifier. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11933–11942 (2022)
5. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling knowledge via knowledge review. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5008–5017 (2021)
6. Cho, J.H., Hariharan, B.: On the efficacy of knowledge distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4794–4802 (2019)
7. 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)
8. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International conference on machine learning. pp. 2206–2216. PMLR (2020)
9. 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)
10. Dong, J., Koniusz, P., Chen, J., Wang, Z.J., Ong, Y.S.: Robust distillation via untargeted and targeted intermediate adversarial samples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28432–28442 (2024)
11. Dong, J., Koniusz, P., Chen, J., Xie, X., Ong, Y.S.: Adversarially robust few-shot learning via parameter co-distillation of similarity and class concept learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28535–28544 (2024)

12. 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)
13. Dong, J., Wang, Y., Lai, J.H., Xie, X.: Improving adversarially robust few-shot image classification with generalizable representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9025–9034 (2022)
14. Dong, J., Wang, Y., Lai, J., Xie, X.: Restricted black-box adversarial attack against deepfake face swapping. IEEE Transactions on Information Forensics and Security **18**, 2596–2608 (2023). https://doi.org/10.1109/TIFS.2023.3266702
15. Dong, J., Wang, Y., Xie, X., Lai, J., Ong, Y.S.: Generalizable and discriminative representations for adversarially robust few-shot learning. IEEE Transactions on Neural Networks and Learning Systems pp. 1–14 (2024). https://doi.org/10.1109/TNNLS.2024.3379172
16. Dong, J., Yang, L., Wang, Y., Xie, X., Lai, J.: Towards intrinsic adversarial robustness through probabilistic training. IEEE Transactions on Image Processing (2023)
17. Dryden, I.L., Koloydenko, A., Zhou, D.: Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. The Annals of Applied Statistics **3**(3), 1102 – 1123 (2009)
18. Gao, R., Wang, J., Zhou, K., Liu, F., Xie, B., Niu, G., Han, B., Cheng, J.: Fast and reliable evaluation of adversarial robustness with minimum-margin attack. In: International Conference on Machine Learning. pp. 7144–7163. PMLR (2022)
19. 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)
20. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS, pp. 2672–2680 (2014)
21. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015)
22. 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)
23. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
24. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
25. 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)
26. Huang, R., Sun, H., Liu, J., Tian, L., Wang, L., Shan, Y., Wang, Y.: Feature variance regularization: A simple way to improve the generalizability of neural networks. Proceedings of the AAAI Conference on Artificial Intelligence **34**(04), 4190–4197 (Apr 2020). https://doi.org/10.1609/aaai.v34i04.5840
27. Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Adversarial examples are not bugs, they are features. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. pp. 125–136 (2019)

28. Jin, Y., Wang, J., Lin, D.: Multi-level logit distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24276–24285 (2023)
29. 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)
30. Kaur, D., Uslu, S., Rittichier, K.J., Durresi, A.: Trustworthy artificial intelligence: a review. ACM Computing Surveys (CSUR) **55**(2), 1–38 (2022)
31. Koniusz, P., Cherian, A.: Sparse coding for third-order super-symmetric tensor descriptors with application to texture recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
32. Koniusz, P., Wang, L., Cherian, A.: Tensor representations for action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(2), 648–665 (2022). https://doi.org/10.1109/TPAMI.2021.3107160
33. Koniusz, P., Yan, F., Gosselin, P.H., Mikolajczyk, K.: Higher-order Occurrence Pooling on Mid- and Low-level Features: Visual Concept Detection. Technical report (Sep 2013), https://inria.hal.science/hal-00922524
34. Koniusz, P., Zhang, H.: Power normalizations in fine-grained image, few-shot image and graph classification. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(2), 591–609 (2022). https://doi.org/10.1109/TPAMI.2021.3107164
35. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
36. Lin, T.Y., Maji, S., Koniusz, P.: Second-order democratic aggregation. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
37. 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)
38. Maroto, J., Ortiz-Jiménez, G., Frossard, P.: On the benefits of knowledge distillation for adversarial robustness. arXiv preprint arXiv:2203.07159 (2022)
39. Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 5191–5198 (2020)
40. 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)
41. 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)
42. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3967–3976 (2019)
43. 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)
44. Rahman, S., Koniusz, P., Wang, L., Zhou, L., Moghadam, P., Sun, C.: Learning partial correlation based deep visual representation for image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6231–6240 (June 2023)

45. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
46. 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)
47. Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., Madry, A.: Adversarially robust generalization requires more data. Advances in neural information processing systems **31** (2018)
48. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
49. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: ICLR (2014)
50. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: 8th International Conference on Learning Representations, ICLR (2020)
51. Wu, D., Xia, S.T., Wang, Y.: Adversarial weight perturbation helps robust generalization. Advances in Neural Information Processing Systems **33**, 2958–2969 (2020)
52. Xie, C., Yuille, A.L.: Intriguing properties of adversarial training at scale. In: 8th International Conference on Learning Representations, ICLR (2020)
53. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: Proceedings of the British Machine Vision Conference (2016)
54. Zhang, H., Li, H., Koniusz, P.: Multi-level second-order few-shot learning. IEEE Trans. Multim. **25**, 2111–2126 (2023). https://doi.org/10.1109/TMM.2022.3142955
55. 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)
56. Zhang, Y., Zhu, H., Chen, Y., Song, Z., Koniusz, P., King, I.: Mitigating the popularity bias of graph collaborative filtering: A dimensional collapse perspective. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 67533–67550. Curran Associates, Inc. (2023)
57. Zhang, Y., Zhu, H., Song, Z., Chen, Y., Fu, X., Meng, Z., Koniusz, P., King, I.: Geometric view of soft decorrelation in self-supervised learning. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD'24, Association for Computing Machinery, New York, NY, USA (2024)
58. 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)
59. 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)