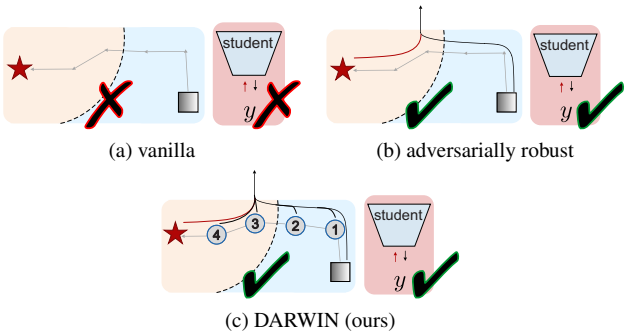
Robust Distillation via Untargeted and Targeted Intermediate Adversarial Samples
– *Supplementary Material* –

Junhao Dong[1,2], Piotr Koniusz[4,3*], Junxi Chen[5], Z. Jane Wang[6], and Yew-Soon Ong[1,2*]
[1]Nanyang Technological University, [2]CFAR, IHPC, A*STAR, [3]Australian National University,
[4]Data61♥CSIRO, [5]Sun Yat-sen University, [6]University of British Columbia
{junhao003, asysong}@ntu.edu.sg, piotr.koniusz@data61.csiro.au,
chenjx353@mail2.sysu.edu.cn, zjanew@ece.ubc.ca

Abstract

In this supplementary material, we commence by giving detailed comparisons of distillation types and experimental configurations (Appendices A and B). Furthermore, we provide more details about our Dual-branch Adversarially Robust knoWledge dIstillatioN (DARWIN) in Appendix C, including adversarially robust teacher models, our extension with label-free adversary generation, and how we conduct the black-box model extraction into a student model by the use of a dataset disjoint with the teacher pre-training dataset. Theoretical analysis is provided in Appendix D. We also present hyper-parameter analyses (Appendix E) and visualization results (Appendix F).

A. Types of Distillation



(a) vanilla (b) adversarially robust

(c) DARWIN (ours)

Figure 5. Zoom on figures (from Fig. 1) illustrating the difference between different distillation types. Fig. 5a shows that the vanilla distillation does not use adversarial generation. Fig. 5b shows that the standard adversarially robust distillation uses only the natural sample and its adversarial counterpart. Fig. 5c shows that our DARWIN also uses intermediate adversarial samples collected from intermediate steps along the "adversarial path".

In this section, we provide more explanations about vanilla knowledge distillation, standard adversarially robust knowledge distillation, and our proposed novel framework, which uses intermediate iterative adversarial samples from multi-step generation. Figure 5 shows the key difference between all three types.

Vanilla knowledge distillation. The teacher model in the context of vanilla distillation is naturally pre-trained using standard methods without specific emphasis on adversarial robustness. Moreover, vanilla knowledge distillation incorporates only clean samples into the knowledge distillation. Thus, such a distillation paradigm transfers knowledge for clean samples instead of the adversarially robust knowledge from the teacher model (Fig. 5a).

Adversarially robust knowledge distillation. Achieving high adversarial robustness usually requires high network capacity, restricting its applicability to large models. To endow lightweight models with adversarial robustness, standard adversarially robust knowledge distillation transfers the robust behavior from an adversarially pre-trained teacher model based on adversarial samples by distilling the clean samples and their adversarial counterparts (Fig. 5b).

DARWIN: our adversarially robust knowledge distillation with intermediate adversarial samples. DARWIN builds upon the concept of adversarially robust knowledge distillation by leveraging the intermediate steps of the adversary generation process. This process is commonly based on an iterative gradient ascent that enlarges the task-related loss to disrupt the target network, as shown in Eq. (2). There exist a dozen intermediate products during such an adversary generation process, which represent a path crossing the decision boundary (Fig. 5c). Existing works primarily resort to two endpoints of adversary generation for distillation: clean samples and their corresponding (final) adversarial samples, overlooking intermediate adversarial samples that also "contain" information about the decision boundary. In this paper, we efficiently incorporate these intermediate adversarial samples into the robust knowledge distillation process with no additional

*Corresponding author.

cost. Moreover, instead of using only untargeted adversarial generation (path towards the nearest decision boundary in Eq. (2)) for robustness transfer, we further incorporate targeted adversary generation (path towards the nearest decision boundary in Eq. (6)) as auxiliary supervision to more effectively capture the structure of decision boundaries of the robust teacher model. Moreover, we also ascertain the importance of intermediate adversarial samples to re-weight their impact on distillation.

B. Experimental Setup

Below, we provide our experimental configurations, including detailed descriptions of the datasets utilized for the adversarially robust knowledge distillation and the implementation details of DARWIN.

B.1. Datasets

Following the evaluation criteria on RobustBench [65], we conduct all the robust knowledge transfer experiments on four standard datasets: CIFAR-10, CIFAR-100 [73], ImageNet-100, and TinyImageNet [78]. The CIFAR-10 dataset contains 60,000 color images of 32×32 resolution across 10 classes. CIFAR-100 shares the same setting as CIFAR-10 except that it contains 100 classes with 600 images per class. Both the ImageNet-100 and TinyImageNet datasets are subsets of the standard ImageNet dataset for evaluating the robustness transfer on real-world data. ImageNet-100 contains 130,000 color images of 100 classes from the original dataset. TinyImageNet is a downsized version of ImageNet that contains 100,000 images over 200 categories, each at a low resolution of 64×64 pixels. For robust knowledge distillation with extra data (Table 4), we additionally include 1M synthetic examples generated by the Denoising Diffusion Probabilistic Model (DDPM) [72] for CIFAR-10/100 following the setting of [66, 76, 77].

B.2. Implementation Details

Following the default experimental settings of robust knowledge distillation methods [70, 87, 88] and also RobustBench [65], we utilize network architectures including ResNet-18/34 [71], MobileNetV2 (MNV2) [79], Wide-ResNet-28-10/34-10 (WRN-28/34) [85] as both teacher and student models. In addition to conventional Convolutional Neural Networks (CNNs), we incorporate Vision Transformers (ViTs) [68, 83] for teacher models to conduct adversarially robust knowledge distillation. We primarily focus on two knowledge transfer paradigms in our experiments: distillation from large-scale models and self-distillation. The former involves robust knowledge transfer from a large-scale adversarially pre-trained teacher model to a relatively lightweight student model. On the contrary, self-distillation means that both the teacher and student models share the same network architecture.

For network parameter optimization during robust knowledge distillation, we adopt the Stochastic Gradient Descent (SGD) optimizer with a momentum factor of 0.9, a weight decay factor of 5×10^{-4}, and a cyclic learning rate schedule [82] with a maximum learning rate of 0.1. Weighting hyper-parameters β and γ are set to 4.0 and 0.5 in Eq. (3) and Eq. (5). We choose loss weighting factors $\lambda_1 = 1.0$ and $\lambda_2 = 0.5$ for all the experiments. The margin factor m in Eq. (9) is set to 0.1. For adversarial robustness, we primarily focus on the ℓ_∞-norm threat model with the maximum perturbation radius $\epsilon = 8/255$ unless stated otherwise. During robust distillation, we adopt the PGD method [74] with $n = 10$ iteration steps (step size $\alpha = 2/255$) to generate adversarial examples. When conducting adversary generation based on class labels, we optimize the Cross-Entropy (CE) loss associated with predictions of the student model. For label-free adversary generation, we leverage the predictions of the teacher model in lieu of ground-truth labels, specifically optimizing the Kullback–Leibler (KL) divergence of predictions between teacher and student models. Further details of our label-free adversary generation strategy are provided in subsequent sections. For a fair and comprehensive assessment, we conduct all robustness evaluations based on the adaptive attack principle. All the experiments are conducted based on a single NVIDIA Tesla A100.

C. Details of DARWIN

Below, we provide more details of DARWIN, including the training paradigms of robust pre-trained teacher models, the extension of our method with the label-free adversary generation, and the black-box model extraction.

C.1. Adversarially Robust Teacher Models

For CNN architectures, we typically obtain the adversarially robust teacher models by one of the most effective adversarial training methods, TRADES [86], which employs the KL divergence for prediction alignment between clean samples and their adversarial counterparts. The objective function of TRADES can be defined as below:

$$\min_{\boldsymbol{\theta}_t} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \Big[\mathcal{L}_{\text{CE}}(f_{\boldsymbol{\theta}_t}(\mathbf{x}), y) + \omega \max_{\|\boldsymbol{\delta}\|_\infty < \epsilon} \mathcal{L}_{\text{KL}}(f_{\boldsymbol{\theta}_t}(\mathbf{x}) \| f_{\boldsymbol{\theta}_t}(\mathbf{x}+\boldsymbol{\delta})) \Big], \tag{17}$$

where $\omega \geq 0$ is a trade-off between clean performance and adversarial robustness. Following the setting of TRADES [86], we set $\omega = 6.0$ for the adversarial training of all the CNN-based teacher models in this paper.

For ViT architectures, we adopt the Projected Gradient Descent (PGD) adversarial training approach [74] with attention random dropping and perturbation random masking

[75] to obtain robust teacher models. These techniques involve dropping gradients from selected attention blocks in ViTs or randomly masking perturbations on certain patches of the input image during the adversarial training process. Specifically, we employ ViT-Base [68] and DeiT-Small [83] architectures to construct our robust teacher models for robust knowledge distillation.

C.2. Label-free DARWIN

We have elaborated upon untargeted and targeted adversary generation strategies based on class labels (CE loss) in Eq. (2) and Eq. (6), respectively. Nevertheless, their reliance on ground-truth labels can hinder their applicability in certain scenarios. In addition to DARWIN based on class labels, we propose a label-free extension of DARWIN in Eq. (11) and Eq. (12), which requires no ground-truth labels during adversarial generation. Consistent with the adversary generation with labels, we randomly initialize untargeted and targeted adversarial examples $\hat{\mathbf{x}}^{(0)}, \hat{\mathbf{x}}'^{(0)} \sim \mathbf{x} + 0.001 \cdot \mathcal{N}(\mathbf{0}, \mathbf{I})$ and conduct an n-step optimization involving a series of intermediate samples $\mathcal{I} = \{\hat{\mathbf{x}}^{(i)}\}_{i=1}^{n-1}$ that are also adversarial towards the student model.

To achieve label-free adversarial generation, we can sample images from "approximately" different classes based on class predictions of the well-optimized teacher model. Such a label-free sampling strategy is applicable to our dual-branch adversary mechanism. Specifically, instead of directly choosing samples based on their semantic labels, we investigate a series of heuristic sampling strategies based on the generated pseudo-labels of input samples, including clustering, negative sampling, and confidence thresholding. We provide details of these sampling strategies below.

Clustering. Deep clustering generally combines DNNs with traditional clustering techniques to organize unlabeled data into meaningful groups [80]. Here, we apply clustering techniques, *e.g.*, k-means and DBSCAN [69], on feature embeddings of clean samples extracted by the teacher model. We can thus assign a pseudo-label for each sample based on the cluster to which it belongs. The rest of the algorithm proceeds as normal, except that the generated pseudo-labels are used in place of class labels.

Negative sampling. Negative sampling aims at identifying instances that are distinct from a given reference. For a given reference sample $\mathbf{x}$, we can examine the top-k class prediction scores of a randomly sampled instance $\mathbf{x}'$ with top-k class prediction scores of sample $\mathbf{x}$. If the top-k probabilities of the sampled instance $\mathbf{x}'$ align with top-k probabilities of $\mathbf{x}$, this instance is discarded, and random sampling is repeated until one finds an instance with unaligned top-k probability scores.

Confidence thresholding. For C classes, the chance that i.i.d. sampled $\mathbf{x}$ and $\mathbf{x}'$ share the same category is $1/C^2$. However, rather than blindly accepting random sampling, one may use a confidence threshold h, *e.g.*, set as $h = 1/C$. This threshold acts as a measurement between the reference sample $\mathbf{x}$ and a randomly sampled instance $\mathbf{x}'$. Subsequently, we can qualify the distinction between them based on some distance, *e.g.*, $\|f_{\boldsymbol{\theta}_t}(\mathbf{x}) - f_{\boldsymbol{\theta}_t}(\mathbf{x}')\|_1$. If the evaluated distance is less than our predefined threshold h, this randomly sampled instance $\mathbf{x}'$ is excluded, and the sampling procedure is repeated until we find a suitable $\mathbf{x}'$.

Table 9. Comparison of label-free strategies for robust knowledge distillation from WRN-34 to ResNet-18 on CIFAR-10.

Strategy	Natural	PGD	AA
Label-based sampling	**84.48**	**55.07**	52.24
Random sampling	83.74	54.59	51.86
Clustering	84.09	54.65	51.98
Negative sampling (top-1)	84.35	55.02	**52.33**
Negative sampling (top-2)	84.30	54.79	52.08
Confidence thresholding	84.12	54.87	52.19

Table 9 provides the evaluation of the aforementioned label-free strategies in comparison to our label-based approach. As one can see, the label-free strategies achieve comparable or even better performance than the label-based DARWIN under Auto-Attack evaluation.

C.3. Black-box model extraction via DARWIN

Given that the label domain of the distilling dataset does not align with the knowledge encapsulated in the teacher model (pre-trained on a different dataset), we resort to label-free adversarial generation and negative sampling to facilitate our DARWIN in the black-box extraction setting. Compared with existing label-free robust distillation methods, we additionally incorporate intermediate adversarial samples into the robust knowledge transfer. The integration of intermediate adversarial samples helps explore the decision boundaries of the teacher, thereby enhancing the robustness of the student against unforeseen adversarial samples.

D. Theoretical Analysis

D.1. Proof of Theorem 2

Proof. It follows from the expansion below:

$$\mathcal{R}_{rob}(\mathcal{D}\cup\mathcal{I}_s) - \mathcal{R}_{rob}(\mathcal{D}) = \mathcal{R}_{rob}(\mathcal{D}\cup\mathcal{I}_f^{\text{✗}}\cup\mathcal{I}_f^{\text{✓}}) - \mathcal{R}_{rob}(\mathcal{D}) \tag{18}$$

$$= \frac{|\mathcal{D}|\,\mathcal{R}_{rob}(\mathcal{D}) + |\mathcal{I}_s^{\text{✗}}|\,\mathcal{R}_{rob}(\mathcal{I}_s^{\text{✗}}) + |\mathcal{I}_s^{\text{✓}}|\,\mathcal{R}_{rob}(\mathcal{I}_s^{\text{✓}})}{|\mathcal{D}| + |\mathcal{I}_s^{\text{✗}}| + |\mathcal{I}_s^{\text{✓}}|} - \frac{|\mathcal{D}| + |\mathcal{I}_s^{\text{✗}}| + |\mathcal{I}_s^{\text{✓}}|}{|\mathcal{D}| + |\mathcal{I}_s^{\text{✗}}| + |\mathcal{I}_s^{\text{✓}}|}\mathcal{R}_{rob}(\mathcal{D}) \tag{19}$$

$$= \frac{|\mathcal{I}_s^{\text{✗}}|\,(\mathcal{R}_{rob}(\mathcal{I}_s^{\text{✗}}) - \mathcal{R}_{rob}(\mathcal{D}))}{|\mathcal{D}| + |\mathcal{I}_s|} + \frac{|\mathcal{I}_s^{\text{✓}}|\,(\mathcal{R}_{rob}(\mathcal{I}_s^{\text{✓}}) - \mathcal{R}_{rob}(\mathcal{D}))}{|\mathcal{D}| + |\mathcal{I}_s|}. \tag{20}$$

□

D.2. Proof of Theorem 3

Proof. For $\mathcal{R}_{rob}(\mathcal{D}\cup\mathcal{I}_s)\geq\mathcal{R}_{rob}(\mathcal{D})$ to hold, we require that (i) $\mathcal{R}_{rob}(\mathcal{I}_s^{\text{✗}})-\mathcal{R}_{rob}(\mathcal{D})\geq 0$ & (ii) $\mathcal{R}_{rob}(\mathcal{I}_s^{\text{✓}})-\mathcal{R}_{rob}(\mathcal{D})\geq 0$.

Point (i) follows directly from Definition 1 of the natural and robust risks. Specifically, we have $\mathcal{R}_{rob}(f_{\boldsymbol{\theta}_s};\mathcal{I}_s^{\text{✗}})=\mathcal{R}_{nat}(f_{\boldsymbol{\theta}_s};\mathcal{I}_s^{\text{✗}})=1\geq\mathcal{R}_{rob}(f_{\boldsymbol{\theta}_s};\mathcal{D})\geq\mathcal{R}_{nat}(f_{\boldsymbol{\theta}_s};\mathcal{D})$. For example, $\mathcal{R}_{nat}(f_{\boldsymbol{\theta}_s};\mathcal{I}_s^{\text{✗}})=1$ as all samples of $\mathcal{I}_s^{\text{✗}}$ are misclassified by Definition3, and $\mathcal{R}_{bdy}(f_{\boldsymbol{\theta}_s};\mathcal{I}_s^{\text{✗}})=0$ as the boundary risk requires correctly classified samples, contradicting Definition 3 of $\mathcal{I}_s^{\text{✗}}$. □

Point (ii) evaluates $\mathcal{R}_{rob}(\mathcal{I}_s^{\text{✓}})-\mathcal{R}_{rob}(\mathcal{D})$. It is easy to observe that $\mathcal{R}_{bdy}(\mathcal{I}_s^{\text{✓}})\geq 0$ and $\mathcal{R}_{nat}(\mathcal{I}_s^{\text{✓}})=0$, where the latter point follows from Definition 1 of the natural risk and Definition 3 of $\mathcal{I}_s^{\text{✓}}$ (correctly classified samples have the natural risk equal zero). The boundary risk gain $\tau_{bdy}=\mathcal{R}_{bdy}(\mathcal{I}_s^{\text{✓}})-\mathcal{R}_{bdy}(\mathcal{D})\geq 0$ according to Proposition 1.

Then it remains to conclude that point (ii) holds if $\tau_{bdy}\geq\mathcal{R}_{nat}(\mathcal{D})$.

D.3. Proof of Theorem 1

Proof. The following inequality follows directly from the definition of the Lipschitz smoothness and $g:=f_{\boldsymbol{\theta}_s}$ and $f_{\boldsymbol{\theta}_s}\to f_{\boldsymbol{\theta}_t}$, *i.e.*, the student network converges to the teacher network:

$$\frac{\left|\delta_y(\mathbf{x},\hat{\mathbf{x}}^{(i)})\right|}{\|\mathbf{x}-\hat{\mathbf{x}}^{(i)}\|_\infty}=\frac{\left|\left(f_{\boldsymbol{\theta}_t}(\mathbf{x})\right)_y-\left(f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^{(i)})\right)_y\right|}{\|\mathbf{x}-\hat{\mathbf{x}}^{(i)}\|_\infty}\leq\kappa, \tag{21}$$

$$\frac{\left|g(\mathbf{x})-g(\hat{\mathbf{x}}^{(i)})\right|}{\|\mathbf{x}-\hat{\mathbf{x}}^{(i)}\|_\infty}\leq\kappa', \tag{22}$$

where $\kappa\approx\kappa'$. Moreover, the locality follows from the fact that for each $\mathbf{x}$ we have a finite subset $\{\hat{\mathbf{x}}^{(i)}\}_{i=1}^{n-1}$ where $\alpha\leq\|\mathbf{x}-\hat{\mathbf{x}}^{(i)}\|_\infty\leq\epsilon$. The chain of inequalities in Eq. (14) follows from the above bounds. □

D.4. Relation of the Boundary Risk to the Distance between Sample and the Decision Boundary

Madry *et al.* [74] observed that introducing adversarial samples increases the robust risk, which can be used in devising adversarially robust networks. Let $d(\mathbf{x},\mathcal{B}_{\boldsymbol{\theta}})$ be the distance between an input sample $\mathbf{x}$ and the closest point from it to a decision boundary $\mathcal{B}_{\boldsymbol{\theta}}$ of the classifier $f_{\boldsymbol{\theta}}(\cdot)$. Xu *et al.* [84] observed that adversarial examples tend to lie closer to the decision boundary compared with their clean samples [84]. Although it is challenging to directly quantify the exact distance of a sample from the decision boundary in high-dimensional spaces, the higher value of the loss (*e.g.*, Cross-entropy or KL divergence) implies closer proximity to the decision boundary if the decision boundary corresponds to regions of high uncertainty (rapid change in predicted probabilities). Thus, for $\hat{\mathbf{x}}\in\mathcal{I}_s^{\text{✓}}$ and its corresponding clean sample $\mathbf{x}\in\mathcal{D}$, we can assume $d(\hat{\mathbf{x}},\mathcal{B}_{\boldsymbol{\theta}_s})\leq d(\mathbf{x},\mathcal{B}_{\boldsymbol{\theta}_s})$ if $\hat{\mathbf{x}}$ is obtained by translating $\mathbf{x}$ towards the decision boundary by a sufficiently small ϵ. Then the following inequality holds $\mathbb{E}_{\hat{\mathbf{x}}\in\mathcal{I}_s^{\text{✓}}}[d(\hat{\mathbf{x}},\mathcal{B}_{\boldsymbol{\theta}_s})]\leq\mathbb{E}_{\mathbf{x}\in\mathcal{D}}[d(\mathbf{x},\mathcal{B}_{\boldsymbol{\theta}_s})]$.

If $d(\mathbf{x},\mathcal{B}_{\boldsymbol{\theta}_s})$ is large, $\mathbf{x}$ is far from the decision boundary, and the chance that its ϵ-perturbation will cause a misclassification is low. Thus, the chance $p\big(\mathcal{R}_{bdy}(f_{\boldsymbol{\theta}_s};\{(\mathbf{x},y)\})=1\big)$ is low. Conversely, if $d(\mathbf{x},\mathcal{B}_{\boldsymbol{\theta}_s})$ is small, perturbing $\mathbf{x}$ by ϵ is more likely to make it cross the decision boundary, causing misclassification. The chance $p\big(\mathcal{R}_{bdy}(f_{\boldsymbol{\theta}_s};\{(\mathbf{x},y)\})=1\big)$ becomes high. Therefore, we can form an inverse relationship between the boundary risk and the distance to the decision boundary and define its soft variant ($\gamma\geq 1$):

$$\mathcal{R}_{bdy}^{\gamma}\big(f_{\boldsymbol{\theta}_s};\{(\mathbf{x},y)\}\big):=\left(\frac{\epsilon}{\max\big(d(\mathbf{x},\mathcal{B}_{\boldsymbol{\theta}_s}(\mathbf{x})),\epsilon\big)}\right)^{\gamma}, \tag{23}$$

$$\mathcal{R}_{bdy}^{\gamma}\big(f_{\boldsymbol{\theta}_s};\mathcal{V}\big):=\mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{V}}\left[\left(\frac{\epsilon}{\max\big(d(\mathbf{x},\mathcal{B}_{\boldsymbol{\theta}_s}(\mathbf{x})),\epsilon\big)}\right)^{\gamma}\right], \tag{24}$$

where $\mathcal{B}_{\boldsymbol{\theta}_s}(\mathbf{x})$ is the nearest boundary for sample $\mathbf{x}$. It holds that $\mathcal{R}_{bdy}\big(f_{\boldsymbol{\theta}_s};\mathcal{V}\big)=\lim_{\gamma\to\infty}\mathcal{R}_{bdy}^{\gamma}\big(f_{\boldsymbol{\theta}_s};\mathcal{V}\big)$ and that $\mathcal{R}_{bdy}^{\gamma}\big(f_{\boldsymbol{\theta}_s};\{(\mathbf{x},y)\}\big)\in[0,1]$.

Proposition 1. *As* $\mathbb{E}_{\hat{\mathbf{x}}\in\mathcal{I}_s^{\text{✓}}}[d(\hat{\mathbf{x}},\mathcal{B}_{\boldsymbol{\theta}_s})]\leq\mathbb{E}_{\mathbf{x}\in\mathcal{D}}[d(\mathbf{x},\mathcal{B}_{\boldsymbol{\theta}_s})]$, *we obtain* $\mathcal{R}_{bdy}^{\gamma}(f_{\boldsymbol{\theta}_s};\mathcal{I}_s^{\text{✓}})\geq\mathcal{R}_{bdy}(f_{\boldsymbol{\theta}_s};\mathcal{I}_s^{\text{✓}})\geq\mathcal{R}_{bdy}(f_{\boldsymbol{\theta}_s};\mathcal{D})$.

Proposition 1 means that correctly classified intermediate adversarial samples have a higher boundary risk than their clean samples.

E. Hyper-parameter Analysis

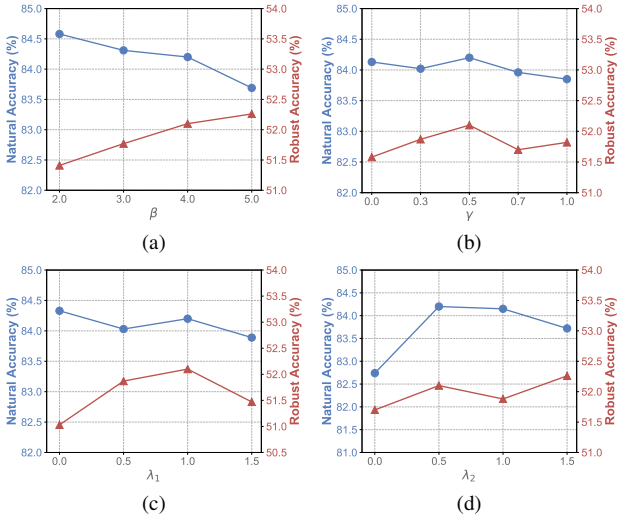


Figure 6. Hyper-parameter sensitivity of DARWIN on natural and (Auto-Attack) robust accuracies on CIFAR-10 when distilling from large-scale models. We report the hyper-parameters adjustment of β in Fig. 6a and γ in Fig. 6b. The tuning for loss weighting factors λ_1 and λ_2 are in Fig. 6c and 6d, respectively.

To improve the understanding of DARWIN, we analyze the impact of hyper-parameters on the performance. Figure 6 shows both natural and adversarially robust accuracies w.r.t. hyper-parameters of DARWIN on CIFAR-10.

Figure 6 shows that the hyper-parameter β is responsible for a trade-off between the natural performance and adversarial robustness of the student model during adversarially robust knowledge distillation. The robust accuracy increases when we enlarge β, whereas the natural performance drops. Conversely, the natural accuracy improves when we lower β, along with the degradation of robustness. Moreover, $\gamma = 0.5$ seems optimal when determining weights in Eq. (5). In addition, appropriately choosing the loss weighting factors λ_1 and λ_2 for adversarially robust knowledge distillation leads to better performance on natural samples and their adversarial counterparts.

F. Visualization

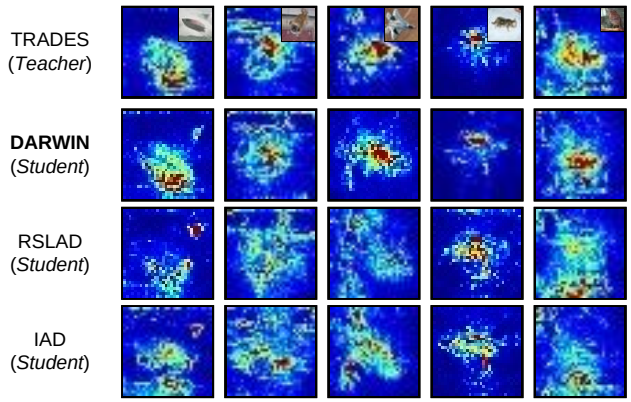


Figure 7. Saliency visualization (SmoothGrad) of adversarial samples. Student models are based on ResNet-18, which are distilled from the teacher model using WRN-34 (trained by TRADES).

In addition to attention visualizations presented in the main manuscript, we have also incorporated saliency visualizations on the CIFAR-10 dataset, as shown in Fig. 7. The saliency maps are obtained via the SmoothGrad [81] technique, which smoothens raw gradients of the class score function over the input space. All the adversarial samples are generated via the PGD method with the maximum perturbation radius $\epsilon = 8/255$. To ensure a fair comparison, we utilize adaptive adversarial generation to generate a corresponding adversarial counterpart for each natural sample across teacher and student models.

We can observe that the saliency maps of the robust student model, derived using our DARWIN method, exhibit greater overlap with the teacher model's saliency maps than with those obtained by other adversarially robust knowledge distillation techniques. This alignment suggests that our DARWIN method effectively captures and transfers robust knowledge from the teacher model. Furthermore, these saliency regions predominantly converge on the discriminative parts of the target objects, indicating that the distilled robustness also aligns well with the human vision. Such findings further corroborate the resilience of our distilled models against unforeseen adversarial examples.

G. Additional Analyses

Table 10. Comparison of the robust accuracy (%) between our DARWIN (distillation approach) (WRN-34→ResNet-18) and adversarial training approaches using ResNet-18 on CIFAR-10.

Method	CIFAR-10			CIFAR-100		
	Natural	PGD	AA	Natural	PGD	AA
PGD-AT [27]	83.80	51.40	47.68	57.39	28.36	23.18
TRADES [55]	82.45	52.21	48.90	56.37	28.68	23.78
MART [49]	82.20	53.94	48.04	54.78	28.79	24.58
DARWIN	**84.48**	**55.07**	**52.24**	**59.12**	**32.30**	**27.26**

Comparisons between the robust distillation and adversarial training approaches. In the main text, we primarily compare our method with previous adversarially robust knowledge distillation approaches. Thus, below, we study a simple baseline for adversarial robustness, *i.e.*, adversarial training, for a comprehensive understanding of our proposed method. Table 10 compares our DARWIN (WRN-34→ResNet-18) with adversarial approaches on ResNet-18. We report the classification accuracy on both clean and adversarial samples. The robust accuracy of DARWIN is ∼5% better.

Table 11. Comparison of the robust accuracy (%) (WRN-34→ResNet-18) of our DARWIN given diverse weighting strategies on CIFAR-10/100.

Weighting Strategies	CIFAR-10			CIFAR-100		
	Natural	PGD	AA	Natural	PGD	AA
Uniform Weighting (no weights)	83.32	53.84	50.92	57.81	30.95	25.70
First Term switched on ($\gamma=0$)	84.13	54.18	51.58	58.53	31.44	26.20
Second Term switched on ($\gamma=1$)	83.85	54.45	51.82	58.16	31.82	26.47
Both Terms switched on ($\gamma=0.5$)	**84.48**	**55.07**	**52.24**	**59.12**	**32.30**	**27.26**

Weighting for intermediate samples. Below, we provide further explanations of the weighting mechanism presented in Eq. (5). The first term, $(1-\gamma)i/n$, is a prior emphasizing intermediate adversaries of later steps i. The second term concentrates on the prediction discrepancy for a local mini-batch. Note the γ in Figure 6b controls the trade-off between these two terms: $\gamma = 0$ does not mean that the weighting is off. As shown in Table 11, the uniform weighting for all the intermediate samples leads to a ∼1.5% drop in both natural performance and adversarial robustness.

Impact of "early" *vs*. "late" intermediate samples. The repulsion term operates between targeted and untargeted branches in DBKD, whose original classes differ. Thus, predictions of samples from two different classes are pushed away, leading to better separable decision regions. Note that

Table 12. Robustness (%) (WRN-34→ResNet-18) of our DARWIN with diverse intermediate adversaries on CIFAR-10/100.

Used Intermediate Adversaries of Different Iterations	CIFAR-10			CIFAR-100		
	Natural	PGD	AA	Natural	PGD	AA
1-5 (early steps)	**84.56**	53.98	50.85	58.98	31.02	25.57
6-10 (later steps)	83.11	54.94	51.73	57.03	31.86	26.73
1-10 (**our default setup**)	84.48	**55.07**	**52.24**	**59.12**	**32.30**	**27.26**

the student's features for intermediate adversaries of early steps (*e.g.*, $i = 1, 2, 3$) are relatively close to features of the clean sample. Thus, repelling them from a clean sample of different classes also boosts natural accuracy (see Table 12).

Table 13. Black-box robustness results (%) (WRN-34→ResNet-18) of our DARWIN on CIFAR-10.

Method	ResNet-18			MNV2		
	FGSM	PGD	MIM	FGSM	PGD	MIM
IAD	67.47	66.19	65.79	67.33	65.80	65.15
RSLAD	67.59	66.06	65.71	67.12	65.45	65.03
DARWIN (Ours)	**69.22**	**67.14**	**66.80**	**68.97**	**66.98**	**66.56**

Black-box robustness evaluation. In addition to the evaluation against white-box attacks, we also provide black-box robustness evaluations below to simulate a more realistic defense scenario. By using the teacher model (WRN-34) as the substitute model, we show our black-box robustness against transferable adversaries on CIFAR-10 (see Table 13). Note that MIM denotes the momentum iterative method [67]. Compared to other methods, our DARWIN enjoys better black-box robustness results under iterative and non-iterative attacks.

References

[65] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[66] Junhao Dong, Seyed-Mohsen Moosavi-Dezfooli, Jianhuang Lai, and Xiaohua Xie. 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)*, pages 24678–24687, 2023.

[67] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[68] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR*, 2021.

[69] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996.

[70] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3996–4003, 2020.

[71] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[72] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

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

[74] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[75] Yichuan Mo, Dongxian Wu, Yifei Wang, Yiwen Guo, and Yisen Wang. When adversarial training meets vision transformers: Recipes from training to architecture. *Advances in Neural Information Processing Systems*, 2022.

[76] Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

[77] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948, 2021.

[78] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[79] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[80] Amit Saxena, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, Aruna Tiwari, Meng Joo Er, Weiping Ding, and Chin-Teng Lin. A review of clustering techniques and developments. *Neurocomputing*, 267:664–681, 2017.

[81] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[82] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386. SPIE, 2019.

[83] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training

data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[84] Yuancheng Xu, Yanchao Sun, Micah Goldblum, Tom Goldstein, and Furong Huang. Exploring and exploiting decision boundary dynamics for adversarial robustness. In *International Conference on Learning Representations*, 2023.

[85] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference*, 2016.

[86] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[87] Jianing Zhu, Jiangchao Yao, Bo Han, Jingfeng Zhang, Tongliang Liu, Gang Niu, Jingren Zhou, Jianliang Xu, and Hongxia Yang. Reliable adversarial distillation with unreliable teachers. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

[88] Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16443–16452, 2021.