Robust Distillation via Untargeted and Targeted Intermediate Adversarial Samples

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

Adversarially robust knowledge distillation aims to compress large-scale models into lightweight models while preserving adversarial robustness and natural performance on a given dataset. Existing methods typically align probability distributions of natural and adversarial samples between teacher and student models, but they overlook intermediate adversarial samples along the "adversarial path" formed by the multi-step gradient ascent of a sample towards the decision boundary. Such paths capture rich information about the decision boundary. In this paper, we propose a novel adversarially robust knowledge distillation approach by incorporating such adversarial paths into the alignment process. Recognizing the diverse impacts of intermediate adversarial samples (ranging from benign to noisy), we propose an adaptive weighting strategy to selectively emphasize informative adversarial samples, thus ensuring efficient utilization of lightweight model capacity. Moreover, we propose a dual-branch mechanism exploiting two following insights: (i) complementary dynamics of adversarial paths obtained by targeted and untargeted adversarial learning, and (ii) inherent differences between the gradient ascent path from class c_i towards the nearest class boundary and the gradient descent path from a specific class c_j towards the decision region of c_i ($i \neq j$). Comprehensive experiments demonstrate the effectiveness of our method on lightweight models under various settings.

1. Introduction

Deep Neural Networks (DNNs) have advanced image classification [17, 18, 24, 27, 39, 41] and retrieval [61], generative models [35, 36, 45], deblurring [59], few-shot learning [23, 29, 51, 52], medical diagnosis [9], and biometrics [32]. However, DNNs are vulnerable to adversarial examples: images with imperceptible perturbations [48]. Despite their visual similarity to the natural images, adversarial examples can fool DNNs, leading to incorrect or harmful predictions with high confidence. Thus, the adversarial vulnerabilities undermine public confidence in the reliability of DNNs and raise trustworthiness concerns [22].

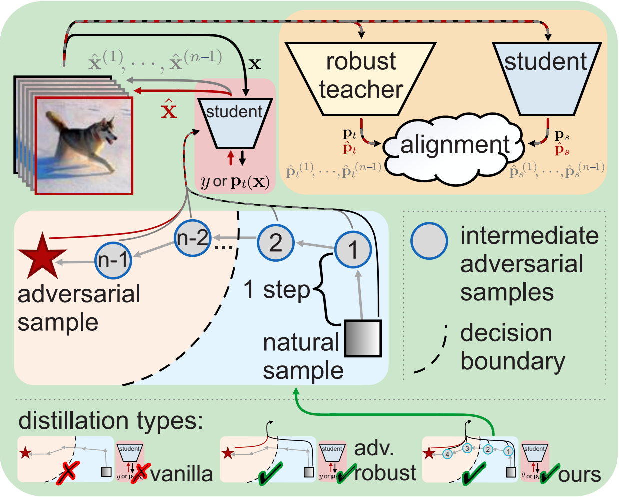

$\hat{\mathbf{x}}^{(1)}, \cdots, \hat{\mathbf{x}}^{(n-1)}$
$\mathbf{x}$
$\hat{\mathbf{x}}$
student
y or $\mathbf{p}_t(\mathbf{x})$
robust teacher
student
$\mathbf{p}_t$
$\hat{\mathbf{p}}_t$
alignment
$\mathbf{p}_s$
$\hat{\mathbf{p}}_s$
$\hat{\mathbf{p}}_t^{(1)}, \cdots, \hat{\mathbf{p}}_t^{(n-1)}$
$\hat{\mathbf{p}}_s^{(1)}, \cdots, \hat{\mathbf{p}}_s^{(n-1)}$
n-1
n-2
2
1
adversarial sample
1 step
natural sample
intermediate adversarial samples
decision boundary
distillation types:
vanilla
adv. robust
ours

Figure 1. Comparison of distillation methods. **(i) Vanilla distillation** employs a naturally trained teacher model without adversarial samples, resulting in susceptibility to adversarial attacks. **(ii) Standard adversarially robust distillation** incorporates adversary generation from each natural sample $\mathbf{x}$ (black frame) to its adversarial counterpart $\hat{\mathbf{x}}$ (red frame). The robust teacher is adversarially pre-trained, and thus the student can distill its responses for natural and adversarial samples. **(iii) Our adversarially robust knowledge distillation with intermediate adversarial samples** extends upon this by introducing intermediate adversarial samples $\hat{\mathbf{x}}^{(1)}, \cdots, \hat{\mathbf{x}}^{(n-1)}$ collected from intermediate steps along the "adversarial path", thereby enabling a more nuanced transfer of robust knowledge based on the comprehensive adversarial landscape. Distillation types are further illustrated in Appendix A.

Adversarial training [16] helps models become adversarially robust by augmenting the training set with adversarial samples. However, adversarial training is computationally costly, with limited usage in resource-constrained scenarios. Moreover, the robustness achieved by adversarial training is mainly observed in large models, leaving lightweight models vulnerable to adversarial attacks [3]. Unlike vanilla knowledge distillation [19], "adversarially robust knowl-

[*]Corresponding author.

edge distillation” [15] overcomes this issue by compressing an adversarially robust large-scale model into a lightweight one without sacrificing much performance.

Adversarially robust knowledge distillation typically focuses on transferring the robustness of a large-scale model to a resource-efficient model by training the student model to mimic outputs of the robust teacher model using merely natural and adversarial samples [15, 63, 64]. However, neglecting intermediate adversarial samples from the “adversarial path” obtained during iterative adversary generation (*e.g*., multi-step gradient ascent) is suboptimal. Specifically, weighted intermediate adversarial samples contribute auxiliary information about the decision boundary, while appropriate weights can amplify informative samples and suppress noisy ones. Fig. 1 shows a comparison of our robust distillation with the adversarial path alongside both the standard adversarially robust and vanilla distillation.

We also notice that previous methods primarily rely on untargeted adversarial samples (robustness w.r.t. the nearest decision boundary of a natural sample). In contrast, our work extends this by exploiting complex decision boundaries between pairs of classes via “targeted adversarial paths”. Specifically, we propose a dual-branch mechanism that comprises: (i) an untargeted adversarial branch, where the predictions of natural samples of class c_i from teacher and untargeted adversarial samples from student should align, (ii) a targeted adversarial branch, which aligns the prediction scores of natural samples of class c_j $(i \neq j)$ from teacher and targeted adversaries from student, and (iii) a divergence criterion that push apart predictions of samples between both branches to differentiate untargeted and targeted adversaries based on underlying semantic distinctions.

Extensive experiments and analyses demonstrate the superior performance of our method compared with other robust knowledge distillation approaches under various scenarios. Our contributions can be summarized as follows:

i. We propose a novel adversarially robust knowledge distillation method that integrates intermediate adversaries along the adversarial path. An adaptive weighting mechanism is proposed to calibrate the influence of each intermediate sample to facilitate the distillation of adversarial paths, refining a robust “understanding” of the decision boundary. Our strategy also leads to minimizing an upper bound of the adversarially robust risk.

ii. To capture relations between decision boundaries, we devise a dual-branch mechanism by harnessing the complementary characteristics of untargeted and targeted adversarial samples. This inter-class relational learning facilitates a more effective robustness transfer.

iii. Extensive experiments showcase the superiority of our method compared with the state-of-the-art approaches across various settings, including diverse backbones, auxiliary data, and cross-dataset distillation.

Related works. Given the security risks posed by adversarial attacks [11, 47], extensive solutions have been proposed to improve DNN robustness [7, 30, 56]. Among them, adversarial training [12, 38, 53, 58] has emerged as a powerful technique to achieve non-trivial robustness by augmenting adversaries into training data. However, its performance highly relies on the network capacity [3], limiting its practical applicability for small models. To bridge this gap, recent studies [15, 54, 62–64] have delved into robust knowledge distillation to transfer adversarial robustness to lightweight models. Zi et al. [64] incorporated soft labels from a robust teacher model as fixed references for distribution alignment. Besides the logit-level alignment, Bai et al. [1] introduced contrastive learning to robustness transfer via latent features. However, previous works primarily focus on the use of natural and adversarial samples for distillation, overlooking the significance of untargeted and targeted intermediate adversaries (paths) to the decision boundary.

Background. The goal of adversarially robust knowledge distillation is to transfer adversarial robustness from a large-scale model (*teacher* model) to a lightweight model (*student* model), where the teacher model is adversarially pretrained to obtain robustness. Let a DNN-based classifier $f_{\boldsymbol{\theta}} : \mathcal{X} \rightarrow [0,1]^C$ with network parameters $\boldsymbol{\theta}$, which outputs probabilities of C classes. Given a dataset with distribution $\mathcal{D}$, standard adversarial training [30] under the ℓ_∞-norm threat model solves the following minimax problem:

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

where $\boldsymbol{\delta}$ denotes the adversarial perturbation bounded within the ℓ_∞-norm of magnitude ϵ, and $\mathcal{L}_{\mathrm{CE}}$ represents the Cross-Entropy (CE) loss. The outer minimization optimizes the adversarial empirical risk over the network parameters $\boldsymbol{\theta}$, while the inner maximization finds the worst-case adversarial examples $\hat{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}$. The Projected Gradient Descent (PGD) [30] is used to optimize the inner maximization:

$$\begin{aligned} \hat{\mathbf{x}}^{(i+1)} &= \vartheta_\alpha(\hat{\mathbf{x}}^{(i)}, y) \\ &= \prod_{\mathbb{B}(\mathbf{x},\epsilon)} \left(\hat{\mathbf{x}}^{(i)} + \alpha \cdot \mathrm{sign} \left(\nabla_{\hat{\mathbf{x}}^{(i)}} \mathcal{L}_{\mathrm{CE}} (f_{\boldsymbol{\theta}} (\hat{\mathbf{x}}^{(i)}), y) \right) \right), \end{aligned} \quad (2)$$

where α is the gradient descent step size. $\Pi_{\mathbb{B}(\mathbf{x},\epsilon)}(\cdot)$ is the projection into the constraints box with ℓ_∞ radius ϵ around $\mathbf{x}$. We randomly initialize $\hat{\mathbf{x}}^{(0)} \sim \mathbf{x} + 0.001 \cdot \mathcal{N}(\mathbf{0}, \mathbf{I})$. After n steps, one obtains $\hat{\mathbf{x}}$ and intermediate samples $\mathcal{I}(\mathbf{x}) = \{\hat{\mathbf{x}}^{(i)}\}_{i=1}^{n-1}$ that are also adversarial due to the non-linear dynamics of decision boundaries. DNNs are also vulnerable to intermediate adversarial samples, which motivates our idea. See Appendix C.1 for further details of teacher pre-training.

2. Proposed Approach

In this section, we propose our novel dual-branch (untargeted and targeted) adversarially robust knowledge distillation method based on intermediate adversarial samples.

2.1. Distillation with Intermediate Adversaries

In contrast to vanilla knowledge distillation, which merely relies on natural samples (non-robust features), robust knowledge distillation additionally incorporates adversaries (robust features) for robustness transfer. Existing works typically utilize predictions of *teacher* on natural examples as the only reference points for aligning with predictions of *student* for both natural and adversarial samples, yet such a strategy fails to mimic the responses of *teacher* to adversarial samples. In our work, we refine this alignment by matching predictions (softmax logits) between the robust teacher model $f_{\boldsymbol{\theta}_t}(\cdot)$ and the student model $f_{\boldsymbol{\theta}_s}(\cdot)$ on both natural samples and their adversarial counterparts. We define the Adversarially Robust Knowledge Distillation (ARKD) as:

$$\mathcal{L}_{\text{ARKD}}=\underbrace{\mathcal{L}_{\text{KL}}\big(f_{\boldsymbol{\theta}_t}(\mathbf{x})\,\|\,f_{\boldsymbol{\theta}_s}(\mathbf{x})\big)}_{\text{alignment of ``natural distributions''}}+\beta\cdot\underbrace{\mathcal{L}_{\text{KL}}\big(f_{\boldsymbol{\theta}_t}(\hat{\mathbf{x}})\,\|\,f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}})\big)}_{\text{alignment of ``adversarial distributions''}}, \tag{3}$$

where $\mathcal{L}_{\text{KL}}$ represents Kullback–Leibler (KL) divergence, and $\beta \geq 0$ controls the trade-off between natural performance and adversarial robustness. Note that Eq. (3) aligns prediction scores without relying on ground-truth labels, which facilitates the adversarially robust knowledge transfer instead of adversarial training from scratch. To obtain an adversarial sample $\hat{\mathbf{x}}$ from its natural counterpart $\mathbf{x}$, one may use the true label y of $\mathbf{x}$ or the prediction $\mathbf{p}_t(\mathbf{x})$ of the teacher during the adversary generation process towards $\hat{\mathbf{x}}$. Our method performs such an "adversarially consistent alignment" as adversaries generated against the student model help probe responses of the adversarially pre-trained teacher, capturing the structure of decision boundaries.

However, Eq. (3) and existing methods only employ natural samples and their adversarial counterparts–they ignore the intermediate adversarial samples (see Fig. 1). In contrast, to explore and capture well the decision boundary of the teacher, we propose to augment intermediate adversarial samples $\mathcal{I}_s\big((\mathbf{x},y)\big)=\big\{(\hat{\mathbf{x}}^{(i)},y)\big\}_{i=1}^{n-1}$ generated from a sample $(\mathbf{x},y)$ into the distillation process. Intermediate Adversarial Knowledge Distillation (IAKD) is defined as:

$$\mathcal{L}_{\text{IAKD}}=\sum_{i=1}^{n-1} w\big(\hat{\mathbf{x}}^{(i)}|\mathbf{x}\big)\cdot\mathcal{L}_{\text{KL}}\big(f_{\boldsymbol{\theta}_t}(\hat{\mathbf{x}}^{(i)})\|f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^{(i)})\big), \tag{4}$$

where $w(\cdot)$ denotes the instance-wise training weight (normalized within the range of $[0,1]$) for each intermediate adversarial sample. Weights are designed as interpolating between (i) an index-based prior, *i.e.*, penalty $i/(n-1)$ gives stronger weights to intermediate samples closer to the final adversarial sample $\hat{\mathbf{x}}$, and (ii) batch-level discrepancy between the student model's predictions on intermediate adversaries and the teacher model's predictions on their corresponding natural samples. Specifically, we define:

$$w\big(\hat{\mathbf{x}}^{(i)}|\mathbf{x}\big)=\frac{(1-\gamma)\,i}{n}+\frac{\gamma\,\big|\big(f_{\boldsymbol{\theta}_t}(\mathbf{x})\big)_y-\big(f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^{(i)})\big)_y\big|}{\max\limits_{j\in\mathcal{B}}\big|\big(f_{\boldsymbol{\theta}_t}(\mathbf{x}_j)\big)_{y_j}-\big(f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}_j^{(i)})\big)_{y_j}\big|}, \tag{5}$$

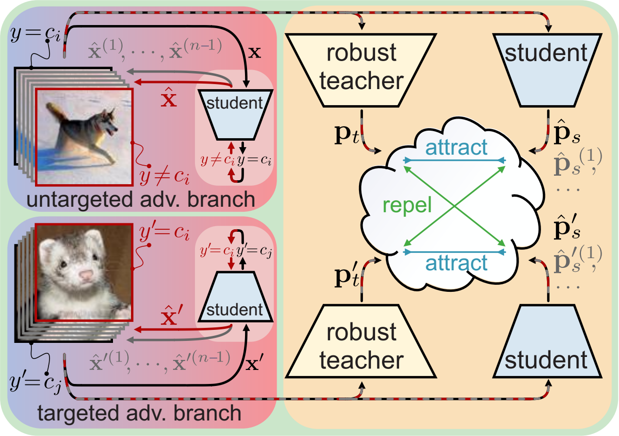


Figure 2. The dual-branch mechanism contains (i) untargeted and (ii) targeted adversary generation branches. The untargeted branch takes sample $\mathbf{x}$ of class $y=c_i$ and produces (intermediate) adversarial samples $\hat{\mathbf{x}},\hat{\mathbf{x}}^{(1)},\cdots,\hat{\mathbf{x}}^{(n-1)}$ by iterative gradient ascent towards $\neg c_i$ (not c_i class). In contrast, the targeted branch takes sample $\mathbf{x}'$ of class $y'=c_j$ and produces (intermediate) adversarial samples $\hat{\mathbf{x}}',\hat{\mathbf{x}}'^{(1)},\cdots,\hat{\mathbf{x}}'^{(n-1)}$ by iterative gradient descent towards class c_i. Subsequently, in each branch, the student's predictions for (intermediate) adversarial samples are attracted towards the teacher's predictions for the natural sample, *i.e.*, $\hat{\mathbf{p}}_s\rightarrow\mathbf{p}_t$ and $\hat{\mathbf{p}}_s^{(1)},\cdots,\hat{\mathbf{p}}_s^{(n-1)}\rightarrow\mathbf{p}_t$ for untargeted adversarial branch and $\hat{\mathbf{p}}'_s\rightarrow\mathbf{p}'_t$ and $\hat{\mathbf{p}}_s'^{(1)},\cdots,\hat{\mathbf{p}}_s'^{(n-1)}\rightarrow\mathbf{p}'_t$ for targeted adversarial branch. Moreover, to improve complementarity of both branches, in each branch, predictions of the student model for (intermediate) adversarial samples are repelled from that of the teacher model for the natural sample of the other branch, *e.g.*, $\hat{\mathbf{p}}'_s\leftrightarrow\mathbf{p}_t$ and $\hat{\mathbf{p}}_s\leftrightarrow\mathbf{p}'_t$.

where $0\leq\gamma\leq1$ interpolates between the prior and the prediction discrepancy for a mini-batch $\mathcal{B}$, and $\big(f(\cdot)\big)_y$ extracts the y-th coefficient from prediction scores of function $f(\cdot)$.

2.2. Dual-branch Adversarially Robust knoWledge dIstillatioN (DARWIN)

Several studies have demonstrated the significance of decision surface modeling to adversarial robustness [8, 28, 34, 60]. The well-established decision boundaries are simultaneously applicable to natural samples and their adversarial counterparts. As existing works typically use untargeted adversarial samples, we propose to use a combination of both untargeted and targeted adversaries with the goal of more effectively capturing the structure of decision boundaries of *teacher*. Fig. 2 illustrates our dual-branch mechanism.

Untargeted adversarial samples are generated via multi-step gradient ascent towards the nearest decision boundary in Eq. (2), where $f_{\boldsymbol{\theta}}(\cdot)$ is replaced with $f_{\boldsymbol{\theta}_s}(\cdot)$. Targeted adversaries are obtained by crossing the decision boundary towards a chosen class. Specifically, we choose $\hat{\mathbf{x}}'^{(0)}\sim\mathbf{x}'+0.001\cdot\mathcal{N}(\mathbf{0},\mathbf{I})$ where $\mathbf{x}'\in\{\mathbf{x}\colon y(\mathbf{x})=c_j\}$ and define:

$$\begin{aligned}\hat{\mathbf{x}}'^{(i+1)}&=\vartheta'_\alpha(\hat{\mathbf{x}}'^{(i)},y')\\&=\prod_{\mathbb{B}(\mathbf{x}',\epsilon)}\Big(\hat{\mathbf{x}}'^{(i)}-\alpha\cdot\text{sign}\left(\nabla_{\hat{\mathbf{x}}'^{(i)}}\mathcal{L}_{\text{CE}}\big(f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}'^{(i)}),y'\big)\right)\Big),\end{aligned} \tag{6}$$

Algorithm 1 Dual-branch Adversarially Robust knoWledge dIstillatioN (DARWIN).

```
Input: Adversarially pre-trained teacher model $f_{\boldsymbol{\theta}_t}$ with parameters $\boldsymbol{\theta}_t$; student model $f_{\boldsymbol{\theta}_s}$ with parameters $\boldsymbol{\theta}_s$; dataset $\mathcal{D}=\{(\mathbf{x}_i,y_i)\}_{i=1}^{D}$ with $D$ data points and $C$ classes; batch size m; learning rate $\tau$; maximum radius of the adversarial perturbation $\epsilon$ and the attack step size $\alpha$; mini-batch size $m$; the number of intermediate adversarial samples $n-1$ (iteration steps $n$); hyper-parameters $\lambda_1$, $\lambda_2$ and $\beta$.
Initialize student network parameters $\boldsymbol{\theta}_s$
while not at end of distillation do
    Form mini-batches for targeted/untargeted adversary generation:
    $\mathcal{B}=\{(\mathbf{x}_j,y_j)\}_{j=1}^{m}$ and $\mathcal{B}'=\{(\mathbf{x}'_j,y'_j): y'_j\neq y_j\}_{j=1}^{m}$
    Set $l_{\text{ARKD}}=0$, $l_{\text{IAKD}}=0$ , and $l_{\text{DBKD}}=0$
    for $j=1,2,\ldots,m$ (in parallel) do
        Draw $\hat{\mathbf{x}}_j^{(0)}\sim\mathbf{x}_j+0.001\cdot\mathcal{N}(\mathbf{0},\mathbf{I})$ , $\hat{\mathbf{x}}_j'^{(0)}\sim\mathbf{x}'_j+0.001\cdot\mathcal{N}(\mathbf{0},\mathbf{I})$
        for $i=1,2,\ldots,n$ do
            Generate untargeted/targeted (intermediate) adv. samples:
            $\hat{\mathbf{x}}_j^{(i)}\leftarrow\vartheta_\alpha(\hat{\mathbf{x}}_j^{(i-1)},y_j)$ and $\hat{\mathbf{x}}_j'^{(i)}\leftarrow\vartheta'_\alpha(\hat{\mathbf{x}}_j'^{(i-1)},y'_j)$
            Use Eq. (8) to accumulate the DBKD loss:
            $l_{\text{DBKD}}\leftarrow l_{\text{DBKD}}$
              $+w(\hat{\mathbf{x}}_j^{(i)}|\mathbf{x}_j)\,\mathcal{L}_{\text{tri}}(f_{\boldsymbol{\theta}_t}(\mathbf{x}_j^{(i)}), f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}_j^{(i)}), f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}_j'^{(i)}))$
              $+w(\hat{\mathbf{x}}_j'^{(i)}|\mathbf{x}'_j)\,\mathcal{L}_{\text{tri}}(f_{\boldsymbol{\theta}_t}(\mathbf{x}_j'^{(i)}), f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}_j'^{(i)}), f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}_j^{(i)}))$
        end for
        Use Eq. (4) to accumulate the IAKD loss:
        $l_{\text{IAKD}}\leftarrow l_{\text{IAKD}}+\sum_{i=1}^{n-1} w(\hat{\mathbf{x}}_j^{(i)}|\mathbf{x}_j)\,\mathcal{L}_{\text{KL}}(f_{\boldsymbol{\theta}_t}(\hat{\mathbf{x}}_j^{(i)})\|f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}_j^{(i)}))$
        Set $\hat{\mathbf{x}}_j=\hat{\mathbf{x}}_j^{(n)}$ (untargeted adversarial sample)
        Use Eq. (3) to accumulate the ARKD loss:
        $l_{\text{ARKD}}\leftarrow l_{\text{ARKD}}$
          $+\mathcal{L}_{\text{KL}}(f_{\boldsymbol{\theta}_t}(\mathbf{x}_j)\|f_{\boldsymbol{\theta}_s}(\mathbf{x}_j))+\beta\cdot\mathcal{L}_{\text{KL}}(f_{\boldsymbol{\theta}_t}(\hat{\mathbf{x}}_j)\|f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}_j))$
    end for
    Update student network parameters:
    $\boldsymbol{\theta}_s\leftarrow\boldsymbol{\theta}_s-\tau\nabla_{\boldsymbol{\theta}_s}[l_{\text{ARKD}}+\lambda_1\,l_{\text{IAKD}}+\lambda_2\,l_{\text{DBKD}}]$
end while
return Student network parameters $\boldsymbol{\theta}_s$.
```

where $y'=c_i$, $c_i\neq c_j$ and $i\neq j$. For simplicity of notations, let adversarial sample $\hat{\mathbf{x}}$ be appended to intermediate adversarial samples $\{\hat{\mathbf{x}}^{(i)}\}_{i=1}^{n-1}$ so that $\hat{\mathbf{x}}^{(n)}=\hat{\mathbf{x}}$. By analogy, we set $\hat{\mathbf{x}}'^{(n)}=\hat{\mathbf{x}}'$. Hence, our dual-branch mechanism performs the following attraction and repulsion steps:

$$\text{attract}\begin{cases}\{f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^{(i)})\}_{i=1}^{n}\rightarrow f_{\boldsymbol{\theta}_t}(\mathbf{x})\\ \{f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}'^{(i)})\}_{i=1}^{n}\rightarrow f_{\boldsymbol{\theta}_t}(\mathbf{x}')\end{cases}, \quad \text{repel}\begin{cases}\{f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^{(i)})\}_{i=1}^{n}\leftrightarrow f_{\boldsymbol{\theta}_t}(\mathbf{x}')\\ \{f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}'^{(i)})\}_{i=1}^{n}\leftrightarrow f_{\boldsymbol{\theta}_t}(\mathbf{x})\end{cases}, \tag{7}$$

where $\rightarrow$ and $\leftrightarrow$ represent the attraction and repulsion operations, respectively. The above steps form our Dual-Branch Knowledge Distillation (DBKD) triplet-based loss:

$$\begin{aligned}\mathcal{L}_{\text{DBKD}}=\sum_{i=1}^{n}\Big[&w(\hat{\mathbf{x}}^{(i)}|\mathbf{x})\,\mathcal{L}_{\text{tri}}(f_{\boldsymbol{\theta}_t}(\mathbf{x}), f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^{(i)}), f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}'^{(i)}))\\ &+w(\hat{\mathbf{x}}'^{(i)}|\mathbf{x}')\,\mathcal{L}_{\text{tri}}(f_{\boldsymbol{\theta}_t}(\mathbf{x}'), f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}'^{(i)}), f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^{(i)}))\Big],\end{aligned} \tag{8}$$

where $\mathcal{L}_{\text{tri}}(\mathbf{p},\mathbf{q},\mathbf{r})$ with a margin constant m is defined as:

$$\mathcal{L}_{\text{tri}}(\mathbf{p},\mathbf{q},\mathbf{r})=\max\left(\|\mathbf{p}-\mathbf{q}\|_2^2-\|\mathbf{p}-\mathbf{r}\|_2^2+m,0\right). \tag{9}$$

Considering the underlying domain shift between untargeted and targeted adversarial samples [14], we further propose to utilize separate Batch Normalization (BN) [21] layers for both branches, which facilitates the disentanglement of the mixed distributions. The main BN layer is primarily used for natural images and untargeted (standard) adversarial samples, while the auxiliary BN is utilized for the targeted adversarial samples. During the inference stage, only the primary BN layer is employed.

Objective function. Our final loss is a combination of the adversarially robust knowledge distillation loss $\mathcal{L}_{\text{ARKD}}$, the intermediate adversarial knowledge distillation loss $\mathcal{L}_{\text{IAKD}}$, and the dual-branch knowledge distillation loss $\mathcal{L}_{\text{DBKD}}$:

$$\mathcal{L}=\mathcal{L}_{\text{ARKD}}+\lambda_1\,\mathcal{L}_{\text{IAKD}}+\lambda_2\,\mathcal{L}_{\text{DBKD}}, \tag{10}$$

where $\lambda_1\geq 0$ and $\lambda_2\geq 0$ are weighting hyper-parameters. The pseudocode of our proposed method is provided in Algorithm 1. During the inference stage, we directly use the distilled student model for robustness evaluations.

2.3. Label-free Adversarially Robust Distillation

As Eq. (2) and (6) primarily rely on ground-truth labels to generate untargeted and targeted adversaries, such reliance can be restrictive when labels are unavailable. To address this, we propose a label-free robust distillation scheme, called DARWIN-LF, by leveraging the predictions of the teacher to simulate ground-truth labels. Hence, untargeted (intermediate) adversarial samples can be obtained by replacing cross-entropy between $f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^{(i)})$ and y in Eq. (2) with KL divergence between $f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^{(i)})$ and $\tilde{\mathbf{y}}=f_{\boldsymbol{\theta}_t}(\mathbf{x})$:

$$\begin{aligned}\hat{\mathbf{x}}^{(i+1)}&=\vartheta_\alpha(\hat{\mathbf{x}}^{(i)},\tilde{\mathbf{y}})\\ &=\prod_{\mathbb{B}(\mathbf{x},\epsilon)}\Big(\hat{\mathbf{x}}^{(i)}+\alpha\cdot\text{sign}\left(\nabla_{\hat{\mathbf{x}}^{(i)}}\mathcal{L}_{\text{KL}}(f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^{(i)})\|\tilde{\mathbf{y}})\right)\Big).\end{aligned} \tag{11}$$

For targeted adversary generation in Eq. (6), we also replace the cross-entropy between $f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^{(i)})$ and y' with KL divergence between $f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^{(i)})$ and $\tilde{\mathbf{y}}'=f_{\boldsymbol{\theta}_t}(\mathbf{x}')$ where $\mathbf{x}'\neq\mathbf{x}$:

$$\begin{aligned}\hat{\mathbf{x}}'^{(i+1)}&=\vartheta'_\alpha(\hat{\mathbf{x}}'^{(i)},\tilde{\mathbf{y}}')\\ &=\prod_{\mathbb{B}(\mathbf{x}',\epsilon)}\Big(\hat{\mathbf{x}}'^{(i)}-\alpha\cdot\text{sign}\left(\nabla_{\hat{\mathbf{x}}'^{(i)}}\mathcal{L}_{\text{KL}}(f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}'^{(i)})\|\tilde{\mathbf{y}}')\right)\Big).\end{aligned} \tag{12}$$

Note that the assertion $y'\neq y$ for Eq. (6) may not always hold if we randomly draw two samples $\mathbf{x}$ and $\mathbf{x}'$. For a dataset of size D partitioned equally into C classes, the chance of them sharing the same label is $p\big(y(\mathbf{x})=y(\mathbf{x}')\big)=1/C^2$, which is extremely low for typical $C\geq 10$. Alternatively, one may compare indexes of the top-k most confident prediction scores $f_{\boldsymbol{\theta}_t}(\mathbf{x})$ and $f_{\boldsymbol{\theta}_t}(\mathbf{x}')$. One may repeat random sampling for $\mathbf{x}'$ if resulting indexes agree, *i.e.*, $\mathbb{1}\big[\big(\text{sort}(f_{\boldsymbol{\theta}_t}(\mathbf{x}))\big)_{1:k}=\big(\text{sort}(f_{\boldsymbol{\theta}_t}(\mathbf{x}'))\big)_{1:k}\big]$, or if $\|f_{\boldsymbol{\theta}_t}(\mathbf{x})-f_{\boldsymbol{\theta}_t}(\mathbf{x}')\|_1\leq h$ where $h=1/C$, *etc*. Here, $\mathbb{1}(\mathbf{v}=\mathbf{v}')$ returns 1 if $\mathbf{v}=\mathbf{v}'$ and $(\mathbf{v})_{1:k}$ returns the first k coefficients of $\mathbf{v}$. By default, we use the "comparison of indexes" strategy for DARWIN-LF. See also Appendix C.2.

2.4. DARWIN Minimizes an Upper Bound of the Adversarially Robust Risk

Let $f^*_{\boldsymbol{\theta}}(\mathbf{x}) = \text{argmax}_c[f_{\boldsymbol{\theta}}(\mathbf{x})]_c$ be the index of maximum predicted probability. Zhang et al. [58] decompose the so-called robust risk $\mathcal{R}_{rob}(f_{\boldsymbol{\theta}_s};\mathcal{V})$ of the distilled student model $f_{\boldsymbol{\theta}_s}$ under set $\mathcal{V}$ as in the following definition.

Definition 1. *The so-called robust, natural, and boundary risks are defined by Zhang et al. [58] as:*

$$\begin{aligned}
&\mathcal{R}_{rob}(f_{\boldsymbol{\theta}_s};\mathcal{V}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{V}}[\mathbb{1}(\exists\hat{\mathbf{x}} \in \mathbb{B}(\mathbf{x},\epsilon) : f^*_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}) \neq y)],\\
&\mathcal{R}_{nat}(f_{\boldsymbol{\theta}_s};\mathcal{V}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{V}}[\mathbb{1}(f^*_{\boldsymbol{\theta}_s}(\mathbf{x}) \neq y)], \qquad (13)\\
&\mathcal{R}_{bdy}(f_{\boldsymbol{\theta}_s};\mathcal{V}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{V}}[\mathbb{1}(\exists\hat{\mathbf{x}} \in \mathbb{B}(\mathbf{x},\epsilon) : f^*_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}) \neq f^*_{\boldsymbol{\theta}_s}(\mathbf{x}) = y)],
\end{aligned}$$

where $\epsilon \geq 0$ is the radius of a ball around $\mathbf{x}$ under the ℓ_∞ norm[1]. Also, $\mathcal{R}_{rob}(f_{\boldsymbol{\theta}_s};\mathcal{V}) = \mathcal{R}_{nat}(f_{\boldsymbol{\theta}_s};\mathcal{V}) + \mathcal{R}_{bdy}(f_{\boldsymbol{\theta}_s};\mathcal{V})$.

The natural risk corresponds to the error on natural examples, while the boundary risk represents how close natural samples are to the decision boundary under ϵ. Typical adversarially robust classification models minimize the robust risk, which ensures good performance on natural and adversarial samples due to low natural and boundary risks.

As DARWIN utilizes intermediate adversarial samples, our proposed weighting mechanism assigns higher weights to such samples proportionally to $\delta_y(\mathbf{x}, \hat{\mathbf{x}}^{(i)}) = \left|\left(f_{\boldsymbol{\theta}_t}(\mathbf{x})\right)_y - \left(f_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}^{(i)})\right)_y\right|$, signifying how fast the change of soft-score for label y is as $\mathbf{x} \to \hat{\mathbf{x}}^{(i)}$ based on the following theorem.

Theorem 1. *The weighting mechanism in Eq. (5) captures the localized κ-Lipschitz smoothness of the student network when $f_{\boldsymbol{\theta}_t}(\cdot) \approx f_{\boldsymbol{\theta}_s}(\cdot)$ (reasonably holds when the student is converging) and $\gamma = 1$ (holds for any $\gamma \in (0, 1]$). We have:*

$$\frac{\left|\delta_y\left(\mathbf{x}, \hat{\mathbf{x}}^{(i)}\right)\right|}{\epsilon} \leq \frac{\left|\delta_y\left(\mathbf{x}, \hat{\mathbf{x}}^{(i)}\right)\right|}{\|\mathbf{x} - \hat{\mathbf{x}}^{(i)}\|_\infty} \leq \kappa \leq \frac{\left|\delta_y\left(\mathbf{x}, \hat{\mathbf{x}}^{(i)}\right)\right|}{\alpha}, \forall_{\substack{i=1,\cdots,\\ n-1}}. \qquad (14)$$

Proof. See Appendix D.3. Here, $|\cdot|$ is the absolute val. □

Hence, higher weights for intermediate adversaries indicate higher κ and higher weighted boundary risk.

Definition 2. *The weighted boundary risk is defined as:*

$$\begin{aligned}
\widehat{\mathcal{R}}_{bdy}(f_{\boldsymbol{\theta}_s};\mathcal{V}) &:= \sum_{(\mathbf{x},y)\sim\mathcal{V}} \frac{w(\hat{\mathbf{x}}|\mathbf{x})}{\omega} \mathbb{1}(\exists\hat{\mathbf{x}} \in \mathbb{B}(\mathbf{x},\epsilon) : f^*_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}) \neq f^*_{\boldsymbol{\theta}_s}(\mathbf{x}) = y),\\
&= \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim\mathcal{V}_w}\left[\mathbb{1}(\exists\hat{\mathbf{x}} \in \mathbb{B}(\mathbf{x},\epsilon) : f^*_{\boldsymbol{\theta}_s}(\hat{\mathbf{x}}) \neq f^*_{\boldsymbol{\theta}_s}(\mathbf{x}) = y)\right], \qquad (15)
\end{aligned}$$

where $\omega = \sum_{(\mathbf{x},y)\sim\mathcal{V}} w(\hat{\mathbf{x}}|\mathbf{x})$ ensures normalized weights and $\mathcal{V}_w$ is the distribution resulting from $\mathcal{V}$ under weights.

Definition 3. *Let $\mathcal{I}_s$ contain intermediate adversarial samples for dataset $\mathcal{D}$. Let $\mathcal{I}_s = \cup_{(\mathbf{x},y)\in\mathcal{D}} \mathcal{I}_s\big((\mathbf{x}, y)\big)$ be split into sets $\mathcal{I}^{\times}_s = \{(\mathbf{x}, y) \in \mathcal{I}_s : f^*_{\boldsymbol{\theta}_s}(\mathbf{x}) \neq y\}$ and $\mathcal{I}^{\checkmark}_s = \{(\mathbf{x}, y) \in \mathcal{I}_s : f^*_{\boldsymbol{\theta}_s}(\mathbf{x}) = y\}$ that contain incorrectly and correctly classified intermediate adversarial samples, respectively.*

> Theorem 1 indicates that applying the IAKD loss from Eq. (4) (which uses weighting) decreases locally the Lipschitz constant κ, resulting in a smoother student function, implying better stability against adversarial attacks. This is signified by Definition 2, which suggests that with high probability, $\widehat{\mathcal{R}}_{bdy}(f_{\boldsymbol{\theta}_s};\mathcal{V}) \geq \mathcal{R}_{bdy}(f_{\boldsymbol{\theta}_s};\mathcal{V})$ holds, which explains that we minimize the (more challenging) upper bound of the robust risk that simultaneously improves the smoothness of the student net.

Theorem 2. *The difference between the robust risk of $\mathcal{D} \cup \mathcal{I}_s$ (dataset and its intermediate adversarial samples) and the robust risk of dataset $\mathcal{D}$ alone is given as:*

$$\mathcal{R}_{rob}(\mathcal{D} \cup \mathcal{I}_s) - \mathcal{R}_{rob}(\mathcal{D}) = \qquad (16)$$
$$\frac{\left|\mathcal{I}^{\times}_s\right| (\mathcal{R}_{rob}(\mathcal{I}^{\times}_s) - \mathcal{R}_{rob}(\mathcal{D}))}{|\mathcal{D}| + |\mathcal{I}_s|} + \frac{\left|\mathcal{I}^{\checkmark}_s\right| (\mathcal{R}_{rob}(\mathcal{I}^{\checkmark}_s) - \mathcal{R}_{rob}(\mathcal{D}))}{|\mathcal{D}| + |\mathcal{I}_s|}$$

Proof. See Appendix D.1. Here, $|\cdot|$ is the cardinality. □

Theorem 3. *DARWIN minimizes the robust risk $\mathcal{R}_{rob}(\mathcal{D} \cup \mathcal{I}_s)$, which is the upper bound of the robust risk of $\mathcal{D}$,* i.e., *$\mathcal{R}_{rob}(\mathcal{D} \cup \mathcal{I}_s) \geq \mathcal{R}_{rob}(\mathcal{D})$ if $\tau_{bdy} \geq \mathcal{R}_{nat}(\mathcal{D})$ where the boundary risk gain obtained from introducing intermediate adversarial samples $\mathcal{I}_s$ is $\tau_{bdy} = \mathcal{R}_{bdy}(\mathcal{I}^{\checkmark}_s) - \mathcal{R}_{bdy}(\mathcal{D}) \geq 0$.*

Proof. See Appendix D.2. □

> Theorem 3 tells that if the boundary risk gain $\tau_{bdy} \geq \mathcal{R}_{nat}(\mathcal{D})$, the use of intermediate adversarial samples $\mathcal{I}_s$ with dataset $\mathcal{D}$ leads to minimization of an upper bound of the robust risk. Moreover, the weighted boundary risk is "focused" on the most perturbing cases. Thus, we expect that $\widehat{\mathcal{R}}_{bdy}(f_{\boldsymbol{\theta}_s};\mathcal{I}^{\checkmark}_s) \geq \mathcal{R}_{bdy}(f_{\boldsymbol{\theta}_s};\mathcal{I}^{\checkmark}_s)$ with high probability, yet we cannot guarantee that (the weights do not capture the decision boundary). We conclude that the use of the weighted boundary risk increases τ_{bdy}, helping ensure the upper bound we optimize holds.

If $\tau_{bdy} < \mathcal{R}_{nat}(f_{\boldsymbol{\theta}_s};\mathcal{D})$, DARWIN achieves improvement in adversarial robustness but we expect a small drop in the natural performance as the boundary risk gain $\tau_{bdy} = \mathcal{R}_{bdy}(\mathcal{I}^{\checkmark}_s) - \mathcal{R}_{bdy}(\mathcal{D}) \geq 0$ is insufficient to compensate for $\mathcal{R}_{nat}(f_{\boldsymbol{\theta}_s};\mathcal{D})$ so that the assertion $\tau_{bdy} \geq \mathcal{R}_{nat}(f_{\boldsymbol{\theta}_s};\mathcal{D})$ from Theorem 3 cannot hold. Section 3.2 shows that when the assertion is violated, a small drop in the natural performance occurs (note that this happens in many such pipelines for adversarially robust self-distillation).

3. Experiments

Below, we provide our experimental setups and compare DARWIN with other adversarially robust distillation works.

Datasets. We conduct experiments on four datasets: CIFAR-10, CIFAR-100 [25], ImageNet-100, and TinyImageNet [42]. See Appendix B.1 for details.

[1] Please note that $(\mathbf{x}, y)$ is a sample-label pair enumerated from the set $\mathcal{V}$ (which can contain clean samples, adversarial samples, *etc.*), so the role and meaning of $\mathbf{x}$ and y depend on what set $\mathcal{V}$ contains.

Table 1. CIFAR-10 and CIFAR-100: Comparisons of our DARWIN with other robust knowledge distillation methods when **distilled from the large-scale WRN-34 teacher model**. The ℓ_∞-norm adversarial perturbations are restricted within $\epsilon = 8/255$. We report both natural accuracy (%) and robust accuracy (%). "**y**" indicates if class labels are required. The best distillation result in each column is in **bold**.

Type	Architecture	Method	**y**	CIFAR-10				CIFAR-100			
				Natural	PGD-20	CW	AA	Natural	PGD-20	CW	AA
Teacher	WRN-34	TRADES [58]	✓	84.92	55.34	54.21	52.55	60.04	31.56	28.64	27.38
Student	ResNet-18	ARD [15]	✓	82.95	52.26	51.69	49.46	57.46	30.14	27.11	25.30
		IAD [63]	✓	82.41	53.06	51.79	49.78	56.38	30.61	27.35	25.51
		RSLAD [64]	✗	83.12	53.91	52.84	51.19	57.23	31.08	28.29	26.62
		CRDND [54]	✗	83.92	52.70	50.95	49.05	58.03	30.16	27.02	25.68
		GACD [1]	✗	82.76	53.42	52.26	50.07	56.82	31.19	27.81	26.12
		DARWIN	✓	**84.48**	**55.07**	53.85	52.24	**59.12**	**32.30**	**28.95**	**27.26**
		DARWIN-LF	✗	84.35	55.02	**53.99**	**52.33**	59.04	32.18	28.62	27.13
	MNV2	ARD [15]	✓	82.44	51.91	50.64	48.40	55.28	30.23	27.05	25.28
		IAD [63]	✓	81.61	52.30	50.19	48.34	54.26	30.46	27.13	25.50
		RSLAD [64]	✗	82.89	52.72	52.04	50.04	57.31	30.48	27.86	25.89
		CRDND [54]	✗	82.77	52.57	50.11	49.28	56.24	29.65	26.68	25.61
		GACD [1]	✗	82.90	52.49	51.40	49.55	56.10	30.49	27.18	25.33
		DARWIN	✓	84.06	**53.94**	**53.11**	**51.28**	**58.45**	**31.53**	**28.36**	26.55
		DARWIN-LF	✗	**84.08**	53.76	52.80	51.09	58.41	31.44	28.33	**26.58**

Table 2. CIFAR-10 and CIFAR-100: Comparisons of our DARWIN with other robust knowledge distillation methods under **the self-distillation scenario**. The ℓ_∞-norm adversarial perturbations are restricted within $\epsilon = 8/255$. We report both natural accuracy (%) and robust accuracy (%). "**y**" indicates whether class labels are required. The best distillation result in each column is in **bold**.

Type	Architecture	Method	**y**	CIFAR-10				CIFAR-100			
				Natural	PGD-20	CW	AA	Natural	PGD-20	CW	AA
Teacher	ResNet-18	TRADES [58]	✓	82.45	52.21	50.29	48.90	56.37	28.68	24.87	23.78
Student	ResNet-18	ARD [15]	✓	81.64	52.62	51.35	49.19	**57.96**	31.34	27.84	26.13
		IAD [63]	✓	80.66	52.63	52.21	48.90	56.45	31.87	28.00	26.66
		RSLAD [64]	✗	81.30	53.80	52.32	50.78	55.17	31.21	27.82	26.46
		CRDND [54]	✗	81.52	52.88	50.85	48.40	56.69	30.13	26.90	26.08
		GACD [1]	✗	81.03	53.26	51.75	49.48	56.95	31.80	28.07	26.30
		DARWIN	✓	82.23	**55.15**	**53.22**	**51.31**	57.74	**32.17**	**28.40**	**26.89**
		DARWIN-LF	✗	**82.25**	55.12	53.15	51.23	57.82	32.11	28.26	26.73
Teacher	MNV2	TRADES [58]	✓	81.04	50.87	48.46	47.15	54.11	27.28	23.39	22.36
Student	MNV2	ARD [15]	✓	81.25	53.02	50.69	48.85	55.64	30.93	27.47	26.05
		IAD [63]	✓	79.36	53.45	50.93	49.14	54.00	31.01	27.59	26.11
		RSLAD [64]	✗	80.01	53.35	51.04	49.74	53.52	29.95	26.66	25.47
		CRDND [54]	✗	80.27	52.92	50.14	49.10	53.70	29.89	26.70	25.67
		GACD [1]	✗	81.17	53.26	50.85	49.18	54.48	31.23	27.38	26.00
		DARWIN	✓	**82.18**	**54.74**	**52.65**	**50.97**	**57.04**	31.89	**28.19**	**26.33**
		DARWIN-LF	✗	82.12	54.52	52.35	50.58	56.90	**31.96**	28.16	26.20

Implementation details. Following [15, 63, 64] & RobustBench [6], we use ResNet-18/34 [18], MobileNetV2 (MNV2) [43], Wide-ResNet-28-10/34-10 (WRN-28/34) [57] as teacher and student models. We also try Vision Transformers (ViTs) [13, 49] as a teacher. We adopt regularization factors $\beta = 4.0$ and $\gamma = 0.5$ with margin $m = 0.1$. In all experiments, loss weighting factors $\lambda_1 = 1.0$ and $\lambda_2 = 0.5$. We use the ℓ_∞-norm threat model with perturbation radius $\epsilon = 8/255$. See Appendix B.2 for more details and Appendix E for hyper-parameter evaluations.

3.1. Main Results

DARWIN with the WRN-34 teacher net. Table 1 reports the classification accuracies on natural examples and their adversarial counterparts based on three standard adversarial attack methods: PGD [30] of 20 iterations with step size $\alpha = 2$, CW [4], and AutoAttack (AA) [5]. AA is a powerful parameter-free robustness evaluation with three white-box (targeted/untargeted) attacks and a black-box attack. For fair comparisons, all experiments use the WRN-34 teacher model. Table 1 shows that DARWIN overall achieves the best natural and adversarial performances. The superior distillation results with ResNet-18 [18] and MNV2 [43] student backbones show the versatility of DARWIN.

DARWIN in the self-distillation setting. Table 2 evaluates DARWIN in the self-distillation setting for which the teacher and student backbones are identical. As one can see, DARWIN and its label-free extension, DARWIN-LF, out-

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

Type	Architecture	Method	Natural	PGD-20	AA
Teacher	ResNet-34	TRADES [58]	72.66	40.88	34.70
Student	ResNet-18	ARD [15]	65.41	38.34	30.94
		RSLAD [64]	66.60	39.12	32.18
		IAD [63]	65.62	39.09	32.63
		CRDND [54]	65.39	39.33	31.36
		GACD [1]	64.85	38.48	31.83
		DARWIN	68.76	**40.37**	**33.58**
		DARWIN-LF	**68.91**	40.16	33.29
	MNV2	ARD [15]	65.90	37.28	30.20
		RSLAD [64]	65.82	37.86	31.66
		IAD [63]	64.96	38.00	31.40
		CRDND [54]	65.09	37.54	30.56
		GACD [1]	64.20	37.27	31.11
		DARWIN	**67.93**	**39.38**	**32.85**
		DARWIN-LF	67.86	39.24	32.51

Table 4. Comparison of robust self-distillation methods (ResNet-18→ResNet-18) on CIFAR-10/CIFAR-100 with auxiliary generative data. We report the natural & (Auto-Attack) robust accuracies.

Dataset	Type	DDPM	Method	Natural	Robust
CIFAR-10	**Teacher**	✗	TRADES [58]	82.45	48.90
	Student	✓	ARD [15]	82.89	53.41
		✓	RSLAD [64]	82.05	52.60
		✓	IAD [63]	82.95	53.47
		✓	**DARWIN**	84.13	55.92
		✓	**DARWIN-LF**	**84.68**	**56.41**
CIFAR-100	**Teacher**	✗	TRADES [58]	56.37	23.78
	Student	✓	ARD [15]	56.07	26.92
		✓	RSLAD [64]	53.40	26.00
		✓	IAD [63]	55.82	26.77
		✓	**DARWIN**	58.18	28.24
		✓	**DARWIN-LF**	**58.74**	**28.45**

perform other methods and the teacher architecture on the robust accuracy for CIFAR-10 and CIFAR-100. In terms of natural performance, DARWIN achieves comparable or even better results than the teacher model. We attribute great robust accuracy to the dual-branch mechanism that helps explore the complex decision boundary.

Robust distillation on ImageNet-100. We here evaluate our approach in the context of larger-scale classification with higher-resolution images and a larger number of classes. Table 3 shows that both DARWIN and DARWIN-LF retain a significant portion of adversarial robustness inherited from the teacher model while simultaneously outperforming other methods on natural performance.

Robust distillation with auxiliary generative data. Auxiliary data generated by the Denoising Diffusion Probabilistic Model (DDPM) [20, 46] has been shown to improve adversarial training [10, 38, 40, 55]. Thus, Table 4 provides the robustness results on CIFAR-10/100 with an additional 1M generated images [40] in the self-distillation setting. For CIFAR-10, our DARWIN significantly outperforms the state-of-the-art method on robust accuracy by a large margin (2%). In addition, DARWIN improves results on natural examples for both CIFAR-10 and CIFAR-100.

DARWIN with the ViT teacher. Vision Transformers (ViT) [13, 49] enjoy good adversarial robustness [2, 26, 31, 33]. Thus, we evaluate DARWIN to see if the intrinsic robustness of ViTs can be distilled into lighter student models. Table 5 shows that DARWIN consistently achieves robust performance that even surpasses the teacher model. Such an improvement in robustness is achieved without compromising the natural performance. Hence, DARWIN is also effective in inheriting robustness from ViT-based teachers.

Table 5. Robust accuracy (%) of models distilled from ViT variants on CIFAR-10 using ResNet-18 and MNV2 student nets.

Type	Architecture	Method	Natural	PGD-20	AA
Teacher	ViT-B	AT-PRM [33]	83.98	53.10	49.66
Student	ResNet-18	ARD [15]	82.76	52.95	49.03
		RSLAD [64]	82.33	54.89	49.74
		IAD [63]	82.27	53.42	49.48
		CRDND [54]	82.19	53.16	48.98
		GACD [1]	81.64	54.24	49.95
		DARWIN	**83.75**	54.80	51.42
		DARWIN-LF	83.73	**54.95**	**51.49**
Teacher	DeiT-S	AT-PRM [33]	82.68	52.47	49.27
Student	MNV2	ARD [15]	81.59	53.45	49.20
		RSLAD [64]	80.86	53.91	50.18
		IAD [63]	80.41	54.12	49.62
		CRDND [54]	80.27	52.21	48.46
		GACD [1]	79.97	54.00	48.91
		DARWIN	83.02	54.46	**51.19**
		DARWIN-LF	**83.15**	**54.62**	51.13

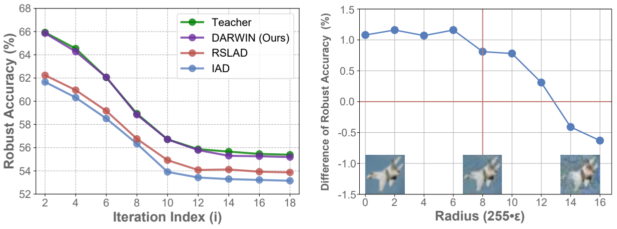
Table 6. Black-box model extraction. WRN-34 teacher pre-trained on CIFAR-10/CIFAR-100 is extracted into a ResNet-18 student with the use of CIFAR-10/CIFAR-100/TinyImageNet. We report the natural accuracy and (Auto-Attack) robust accuracy.

Pre-training Dataset	**Distillation** Dataset	Method	Natural	Robust
CIFAR-10	CIFAR-100	RSLAD [64]	70.03	37.68
		CRDND [54]	69.30	37.29
		GACD [1]	69.22	38.05
		DARWIN-LF	**72.48**	**41.79**
	TinyImageNet	RSLAD [64]	64.44	30.29
		CRDND [54]	64.95	31.83
		GACD [1]	65.78	33.12
		DARWIN-LF	**66.73**	**36.20**
CIFAR-100	CIFAR-10	RSLAD [64]	44.41	18.51
		CRDND [54]	45.38	18.65
		GACD [1]	44.34	18.37
		DARWIN-LF	**46.29**	**21.90**

Black-box model extraction via DARWIN. The black-box model extraction recovers an online black-box model with no access to its model parameters or training data [37, 50]. Thus, we extract the teacher model (pre-trained on an inaccessible source dataset) by DARWIN with a target dataset that differs from the source dataset. We aim to recover the natural performance and adversarial robustness of the black-box teacher. Table 6 shows that DARWIN-LF outperforms other adversarially robust models in both natural and adversarial evaluation metrics upon testing the student on the source test set. Thus, DARWIN-LF is effective in extracting knowledge of the black-box model. Appendix C.3 contains more details about this problem.

Table 7. Ablation study (WRN-34→ResNet-18) of three loss components of DARWIN for accuracy (%) on CIFAR-10/CIFAR-100.

	ARKD	IAKD	DBKD	Natural	PGD-20	AA
1	✓			83.09/57.34	53.95/30.59	50.66/25.32
2	✓	✓		82.74/56.68	54.52/32.11	51.70/26.95
3	✓		✓	84.68/59.23	54.36/31.53	51.21/26.18
	✓	✓	✓	84.48/59.12	55.07/32.30	52.24/27.26



(a) (b)

Figure 3. Experiments on CIFAR-10 (WRN-34→ResNet-18). (a) Average robust accuracy (PGD-20) w.r.t. the index number $i = 6, \ldots, 14$ of the intermediate adversarial test samples ($\alpha = 1/255, \epsilon = 8/255$). (b) Difference of AutoAttack robust accuracy under different attack strengths (radii ϵ) between DARWIN and RSLAD [64]. Zoom on tiny pictures: $\epsilon = 0/255$ is a clean sample, $\epsilon = 8/255$, and $\epsilon = 16/255$ are adversarial samples. We used $\epsilon = 8/255$ (indicated by the dashed vertical line) for training.

3.2. Ablation Studies

Loss components. Below, we investigate individual loss components of DARWIN: (i) our baseline of Adversarially Robust Knowledge Distillation (ARKD) in Eq. (3), (ii) Intermediate Adversarial Knowledge Distillation (IAKD) in Eq. (4), and (iii) Dual-Branch Knowledge Distillation (DBKD) in Eq. (8). Table 7 reports accuracy (CIFAR-10) for both the natural and adversarially robust performance.

Our baseline ARKD-based alignment of the student with the teacher obtains a competitive performance, and using intermediate adversarial samples with the IAKD loss further improves the distillation. The dual-branch module (DBKD) also boosts the natural performance/adversarial robustness.

Performance w.r.t. the index number $i = 1, \cdots, n-1$ of the intermediate adversarial test samples. Figure 3a provides the average robust accuracy w.r.t. the index number on DARWIN, RSLAD [64], and IAD [63]. DARWIN shows the largest gain compared to to RSLAD and IAD on early intermediate adversarial test samples (*e.g.*, $i = 6$).

Performance w.r.t. radius ϵ. Figure 3b shows the robustness gap between DARWIN and RSLAD under several attack radii ϵ (training $\epsilon = 8/255$). DARWIN achieves better natural performance and adversarial robustness under weaker adversaries ($\epsilon \leq 12$). In contrast, RSLAD captures well the adversarial robustness against strong adversarial perturbations ($\epsilon \geq 12$). Thus, DARWIN captures well small visually undetectable attacks (zoom tiny pictures in Fig. 3b) whereas RSLAD only handles visually conspicuous attacks.

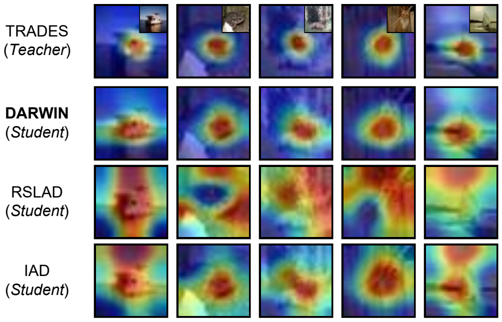


Figure 4. Attention maps of the teacher (WRN-34 obtained by TRADES [58]) and several students based on ResNet-18. Notice the similarity of maps of the teacher and DARWIN (student).

Table 8. The boundary risk gain τ_{bdy} *vs.* the natural risk $\mathcal{R}_{nat}(\mathcal{D}; \boldsymbol{\theta}_s)$ for robust self-distillation (ResNet-18, CIFAR-10).

Epoch	MNV2			ResNet-18		
	τ_{bdy}	$\mathcal{R}_{nat}(\mathcal{D}; \boldsymbol{\theta}_s)$	Δ	τ_{bdy}	$\mathcal{R}_{nat}(\mathcal{D}; \boldsymbol{\theta}_s)$	Δ
20-th	0.320	0.306	0.014	0.275	0.261	0.014
40-th	0.358	0.284	0.014	0.237	0.242	-0.005
60-th	0.315	0.272	0.043	0.226	0.240	-0.014
80-th	0.309	0.287	0.022	0.235	0.234	0.001
100-th	0.281	0.253	0.028	0.228	0.230	-0.002

Performance w.r.t. τ_{bdy} *vs.* $\mathcal{R}_{nat}(\mathcal{D})$. Theorem 3 indicates that if the boundary risk gain $\tau_{bdy} < \mathcal{R}_{nat}(\mathcal{D})$, the use of intermediate adversarial samples $\mathcal{I}_s$ with dataset $\mathcal{D}$ may violate the minimization of an upper bound of the robust risk. Table 8 captures if the boundary risk τ_{bdy} compensates for the natural risk $\mathcal{R}_{nat}(f_{\boldsymbol{\theta}_s}; \mathcal{D})$. For MNV2, the assertion $\tau_{bdy} \geq \mathcal{R}_{nat}(\mathcal{D})$ holds, and thus Table 2 shows a gain in the natural performance of the student over the teacher. For ResNet-18, the assertion fails, leading to a slight drop in the natural performance of the student model (see Table 2).

3.3. Visualization

Figure 4 shows that the student model distilled by DARWIN shares similar attention (Grad-CAM [44]) regions with the teacher model, unlike other methods. Thus, DARWIN captures the complex decision boundaries of the teacher better than RSLAD/IAD. Appendix F shows more visualizations.

4. Conclusions

We propose a novel adversarially robust knowledge distillation approach, DARWIN, that efficiently incorporates dual-branch intermediate adversarial samples into robustness transfer with the goal of capturing the complex decision boundaries of the teacher model. We demonstrate that our DARWIN benefits from an instance-wise weighting scheme, and it minimizes an upper bound of the robust risk. We make a connection between violating such a theoretical bound and a slight degradation in the natural performance exhibited by many adversarially robust distillation methods.

Acknowledgments. This work is partially supported by the National Research Foundation Singapore under its AI Singapore Program [Award No.:AISG2-GC-2023-007], CFAR, IHPC, A*STAR, and SCSE at Nanyang Technological University. PK is supported by CSIRO's Science Digital.

References

[1] Tao Bai, Jun Zhao, and Bihan Wen. Guided adversarial contrastive distillation for robust students. *IEEE Transactions on Information Forensics and Security*, 2023. 2, 6, 7

[2] Philipp Benz, Soomin Ham, Chaoning Zhang, Adil Karjauv, and In So Kweon. Adversarial robustness comparison of vision transformer and mlp-mixer to cnns. In *32nd British Machine Vision Conference 2021, BMVC*, 2021. 7

[3] Sébastien Bubeck, Yuanzhi Li, and Dheeraj M Nagaraj. A law of robustness for two-layers neural networks. In *Conference on Learning Theory*, pages 804–820, 2021. 1, 2

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017. 6

[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 6

[6] 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. 6

[7] Zhijie Deng, Xiao Yang, Shizhen Xu, Hang Su, and Jun Zhu. Libre: A practical bayesian approach to adversarial detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 972–982, 2021. 2

[8] Junhao Dong, Yuan Wang, Jian-Huang Lai, and Xiaohua Xie. Improving adversarially robust few-shot image classification with generalizable representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9025–9034, 2022. 3

[9] Junhao Dong, Junxi Chen, Xiaohua Xie, Jianhuang Lai, and Hao Chen. Adversarial attack and defense for medical image analysis: Methods and applications. *arXiv preprint arXiv:2303.14133*, 2023. 1

[10] 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. 7

[11] Junhao Dong, Yuan Wang, Jianhuang Lai, and Xiaohua Xie. Restricted black-box adversarial attack against deepfake face swapping. *IEEE Transactions on Information Forensics and Security*, 18:2596–2608, 2023. 2

[12] Junhao Dong, Lingxiao Yang, Yuan Wang, Xiaohua Xie, and Jianhuang Lai. Toward intrinsic adversarial robustness through probabilistic training. *IEEE Transactions on Image Processing*, 32:3862–3872, 2023. 2

[13] 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. 6, 7

[14] Liam Fowl, Micah Goldblum, Ping-yeh Chiang, Jonas Geiping, Wojciech Czaja, and Tom Goldstein. Adversarial examples make strong poisons. *Advances in Neural Information Processing Systems*, 34:30339–30351, 2021. 4

[15] 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. 2, 6, 7

[16] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 1

[17] Maryam Haghighat, Peyman Moghadam, Shaheer Mohamed, and Piotr Koniusz. Pre-training with random orthogonal projection image modeling. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[18] 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. 1, 6

[19] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 7

[21] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 4

[22] Wei Jiang, Zhiyuan He, Jinyu Zhan, Weijia Pan, and Deepak Adhikari. Research progress and challenges on application-driven adversarial examples: A survey. *ACM Transactions on Cyber-Physical Systems (TCPS)*, 5(4):1–25, 2021. 1

[23] Dahyun Kang, Piotr Koniusz, Minsu Cho, and Naila Murray. Distilling self-supervised vision transformers for weakly-supervised few-shot classification & segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19627–19638. IEEE, 2023. 1

[24] Piotr Koniusz, Fei Yan, Philippe-Henri Gosselin, and Krystian Mikolajczyk. Higher-order Occurrence Pooling on Mid- and Low-level Features: Visual Concept Detection. Technical report, 2013. 1

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[26] Yanxi Li and Chang Xu. Trade-off between robustness and accuracy of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7568, 2023. 7

[27] Tsung-Yu Lin, Subhransu Maji, and Piotr Koniusz. Second-order democratic aggregation. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III*, pages 639–656. Springer, 2018. 1

[28] Feng Liu, Bo Han, Tongliang Liu, Chen Gong, Gang Niu, Mingyuan Zhou, Masashi Sugiyama, et al. Probabilistic margins for instance reweighting in adversarial training. *Advances in Neural Information Processing Systems*, 34: 23258–23269, 2021. 3

[29] Changsheng Lu and Piotr Koniusz. Detect any keypoints: An efficient light-weight few-shot keypoint detector. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4):3882–3890, 2024. 1

[30] 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. 2, 6

[31] Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7838–7847, 2021. 7

[32] Blaž Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter J. Scheirer, Arun Ross, Peter Peer, and Vitomir Štruc. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security*, 16:4147–4183, 2021. 1

[33] 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. 7

[34] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9078–9086, 2019. 3

[35] Yao Ni and Piotr Koniusz. NICE: NoIse-modulated Consistency rEgularization for Data-Efficient GANs. In *Advances in Neural Information Processing Systems*, pages 13773–13801. Curran Associates, Inc., 2023. 1

[36] Yao Ni and Piotr Koniusz. CHAIN: Enhancing Generalization in Data-Efficient GANs via lipsCHitz continuity constrAIned Normalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*. IEEE, 2024. 1

[37] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4954–4963, 2019. 7

[38] 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. 2, 7

[39] Saimunur Rahman, Piotr Koniusz, Lei Wang, Luping Zhou, Peyman Moghadam, and Changming Sun. Learning partial correlation based deep visual representation for image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023. 1

[40] 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. 7

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[42] 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. 5

[43] 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. 6

[44] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, pages 618–626, 2017. 8

[45] Fatemeh Shiri, Xin Yu, Fatih Porikli, Richard Hartley, and Piotr Koniusz. Recovering faces from portraits with auxiliary facial attributes. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 406–415, 2019. 1

[46] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265, 2015. 7

[47] Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 241–257, 2019. 2

[48] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations, ICLR*, 2014. 1

[49] 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. 6, 7

[50] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pages 601–618, 2016. 7

[51] Lei Wang and Piotr Koniusz. Uncertainty-dtw for time series and sequences. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXI*, pages 176–195. Springer, 2022. 1

[52] Lei Wang and Piotr Koniusz. Temporal-viewpoint transportation plan for skeletal few-shot action recognition. In *Asian Conference on Computer Vision*, page 307–326, Berlin, Heidelberg, 2023. Springer-Verlag. 1

[53] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020. 2

[54] Yuzheng Wang, Zhaoyu Chen, Dingkang Yang, Yang Liu, Siao Liu, Wenqiang Zhang, and Lizhe Qi. Adversarial contrastive distillation with adaptive denoising. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2, 6, 7

[55] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning, ICML*, pages 36246–36263, 2023. 7

[56] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018. 2

[57] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference*, 2016. 6

[58] 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. 2, 5, 6, 7, 8

[59] Hongguang Zhang, Limeng Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Event-guided multi-patch network with self-supervision for non-uniform motion deblurring. *Springer International Journal of Computer Vision (IJCV)*, 131:453–470, 2023. 1

[60] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *International Conference on Learning Representations*, 2021. 3

[61] Zhongyan Zhang, Lei Wang, Luping Zhou, and Piotr Koniusz. Learning spatial-context-aware global visual feature representation for instance image retrieval. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11216–11225. IEEE, 2023. 1

[62] Shiji Zhao, Jie Yu, Zhenlong Sun, Bo Zhang, and Xingxing Wei. Enhanced accuracy and robustness via multi-teacher adversarial distillation. In *European Conference on Computer Vision*, pages 585–602. Springer, 2022. 2

[63] 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. 2, 6, 7, 8

[64] 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. 2, 6, 7, 8