Adversarial local distribution regularization for knowledge distillation

Thanh Nguyen-Duc[1], Trung Le[1], He Zhao[2], Jianfei Cai[1] and Dinh Phung[1]
[1]Department of Data Science and AI, Monash University
[2]CSIRO's Data61
[1]{thanh.nguyen4,trunglm,jianfei.cai,dinh.phung}@monash.edu, [2]he.zhao@ieee.org

Abstract

Knowledge distillation is a process of distilling information from a large model with significant knowledge capacity (teacher) to enhance a smaller model (student). Therefore, exploring the properties of the teacher is the key to improving student performance (e.g., teacher decision boundaries). One decision boundary exploring technique is to leverage adversarial attack methods, which add crafted perturbations within a ball constraint to clean inputs to create attack examples of the teacher called adversarial examples. These adversarial examples are informative examples because they are near decision boundaries. In this paper, we formulate a teacher adversarial local distribution, a set of all adversarial examples within the ball constraint given an input. This distribution is used to sufficiently explore the decision boundaries of the teacher by covering the full spectrum of possible teacher model perturbations. The student model is then regularized by matching the loss between teacher and student using these adversarial example inputs. We conducted a number of experiments on CIFAR-100 and Imagenet datasets to illustrate this teacher adversarial local distribution regularization (TALD) can be applied to improve performance of many existing knowledge distillation methods (e.g., KD, FitNet, CRD, VID, FT, etc.).

1. Introduction

Transferring knowledge from an excessive deep learning model (teacher) to a lighter model (student) is known as knowledge distillation (KD). The light-weight student is advantageous when deployment costs need to be lowered due to the devices' constrained computing and memory capabilities. Hinton et al. [21] originally introduced the objective of KD loss, which minimizes the KL divergence between the teacher and student outputs. This KD loss extracts knowledge from the teacher's class probabilities with the temperature softmax to guide the training of the student. Therefore, the student network is developed to be a better classifier than the student developed without KD loss.

Many studies improve the KD loss [21] for matching the teacher and student outputs such as label smoothing [25], virtual teacher [48], and decouple KD loss [51]. Moreover, deep learning models are well-learned multiple levels of feature representation [3]. Feature-based KD is adopted in the works [40, 37, 9], in which teacher provides intermediate representations and hints for training the student. These previous approaches attempt to manipulate various network components to enhance the knowledge distillation process. The work [12] shows that input samples close to the classifier's decision boundaries affect performance more than samples further from it, which can help to regularize the student [20]. Therefore, we can effectively transfer the teacher properties to the student by utilizing informative samples near the teacher decision boundaries.

One strategy for exploring decision boundaries is utilizing adversarial attack approaches. The adversarial attack [18, 50, 13, 31, 20] transports clean inputs to the model's decision boundaries by iteratively adding specially designed perturbations inside of a ball constraint to produce adversarial examples. Although finding a decision boundary is not the primary objective of an adversarial attack, they are closely related to each other [7]. Moreover, the vanilla adversarial attacks [18, 50, 13, 31, 20, 5, 6] can only create one adversarial example, which may not be enough to examine the full spectrum of possible teacher model perturbations. The works [42, 34] also show how attacks with random initialization can lie together and lose diversity, which reduces the quality of adversarial examples.

In this paper, we introduce a *teacher adversarial local distribution* (TALD) regularization for knowledge distillation, which can be used to improve many existing KD approaches. Our contributions are summarized as follows:

- We explore the teacher decision boundaries by introducing the teacher adversarial local distribution, a set of all adversarial examples within the constraint given an input that maximize a teacher loss function.

- We find TALD by using a multiple particle-based search named Stein Variational Gradient Decent

(SVGD) [29]. The SVGD sufficiently approximates the TALD without any assumptions and creates more diverse adversarial examples. The student decision boundaries are then regularized by matching the loss between teacher and student using these adversarial example inputs.

- We show that our method can be adapted well to various existing methods. We conduct various experiments on CIFAR-100 and ImageNet to demonstrate our TALD regularization can improve performance of many existing methods such as KD [21], FitNet [40], AT [49], SP [44], CC [38], VID [1], RKD [36], PKT [37], AB [20], FT [24], NST [22], CRD [43], SSKD [47], and HSAKD[11].

2. Related Works

Knowledge distillation. Hinton et al. [21] originally introduced knowledge distillation (KD) which extracts knowledge from class probabilities of a large deep learning model (teacher) to a lighter model (student). Many approaches have adopted the class-probability knowledge transfer perspective of KD [21] to improve model compression such as class-distance loss [25], label smoothing [33], adaptive regularization [15], virtual teacher [48], and decoupling KD loss [51]. In addition, deep learning models are well-learned in multiple levels of feature representation [3]. Romero et al. exploited the intermediate teacher representation that the student model is trained by matching the teacher responses from multiple hidden layers named FitNets [40]. Various approaches have been proposed inspired by [40] to significantly improve the student model. The teacher feature maps are selected using attention mechanism to omit redundant knowledge transfer from the teacher to student [26, 22]. The knowledge from the teacher can distill to the student and be explained [35]. Zhou et al. [52] explored parameter sharing of intermediate layers of the teacher model. Cross-layer knowledge distillation [9] adaptively assigns proper teacher layers for each student layer via attention allocation that matches the semantics between teacher and student. The work [19] proposed an efficient use of the pre-trained teacher's intermediate representations. Contrastive learning loss was proposed by Tian et al. [43] to capture correlations and higher order output dependencies. Hierarchical self-supervised learning technique was proposed in the work [11] to improve the student. However, these above approaches have not considered the teacher decision boundary perspective. In this paper, we introduce the regularization using the teacher decision boundary information to add an additional help to enhance the teacher property transfer to improve the student. Our regularization loss only requires the teacher to be differentiable and has no additional learnable module. Therefore, we can easily add the regularization loss to many existing KD methods.

Adversarial attack. State-of-the-art deep neural networks are reportedly vulnerable to attacks [18, 41]. Fast Gradient Sign Method (FGSM) [18], Projected Gradient Descent (PGD) [30], and Auto-Attack [13] are a few examples of adversarial attacks that add specially crafted perturbations to clean inputs to produce adversarial examples. The most popular technique for finding perturbations is using gradients to maximize a model's loss on given a clean input while limiting the perturbation size smaller than a specified amount referred to a radius constraint epsilon. In other words, the adversarial attacks find a path to transport clean inputs to cross model decision boundaries, which means to fool the model prediction. Due to the threats, many methods have been proposed for defense techniques using adversarial examples such as [31, 39, 46, 50, 28, 4]. Recently, the works [16, 34] proposed adversarial distribution training to improve the model robustness. In knowledge distillation, exploring the properties of the teacher (e.g., decision boundaries) is the key to improving student performance. Therefore, we leverage the attack to explore the teacher decision boundaries using generated adversarial examples. These adversarial examples are then used to regularize the student.

Knowledge distillation using adversarial attacks. Many previous approaches use knowledge distillation for transferring robustness from a well-defended teacher to a student. Robustness transfer from a robust teacher to a student using KD loss [21] technique was proposed by the work [17]. Chan et al. [8] trained a student model's input gradients to match those of the robust teacher to gain robustness. In addition, the work [10] proposed a noisy feature distillation, a new transfer learning method that improve robustness. Other works [23, 2] used contrastive learning loss to transfer robustness. These above distillation approaches only attempt to distill robustness to defend from adversarial attacks. Heo et al.'s paper [20] proposed a BSS attack for exploring the teacher's properties using adversarial examples to increase the student's clean input accuracy. This BSS can only produce one adversarial example, which insufficiently explores the full spectrum of possible teacher model perturbations [42, 34]. In this paper, our approach sufficiently explores teacher's properties (e.g., decision boundaries) using the teacher adversarial local distribution (TALD). The student is then regularized by TALD regularization to improve clean input accuracy.

3. Method

In this section, we introduce our teacher adversarial local distribution (TALD) regularization that can be used to improve performance of many previous knowledge distilaltion methods (e.g., KD, FitNet, CRD, SSKD, etc.). We denote the large classifier teacher model by T with parameters θ_T. The teacher is pre-trained and fixed. The student is a smaller model which needs to be trained with help from the teacher. The smaller student model is S parameterized by θ_S. Let input $\boldsymbol{x} \in \mathbb{R}^d$ be our d-dimensional clean input data in a space $\boldsymbol{X}$, and $(\boldsymbol{x}, y) \sim P_{\mathbb{D}}$ is our data-label distribution.

3.1. Teacher adversarial local distribution

We use adversarial examples, which are near decision boundaries, to explore teacher decision boundary properties called teacher adversarial examples. The student decision boundaries are then regularized by matching the teacher loss and student loss given these input examples. These adversarial inputs can be found by attacking the teacher model. The attack adds crafted perturbations within a ball constraint to clean input $\boldsymbol{x}$, which maximizes the teacher loss function ℓ to generate adversarial examples $\boldsymbol{x}_{adv}$. We denote the ball constraint of $\boldsymbol{x}_{adv}$ by $C_\epsilon(\boldsymbol{x}) = \{\boldsymbol{x}_{adv} \in \boldsymbol{X} : ||\boldsymbol{x}_{adv} - \boldsymbol{x}||_p \leq \epsilon\}$, where $\boldsymbol{x}_{adv}$ is adversarial example, and ϵ is a ball constraint radius with respect to a norm $||\cdot||_p$. The teacher attack can be defined by the maximization optimization in Eq. 1.

$$\boldsymbol{x}_{adv} = \underset{\boldsymbol{x}_{adv} \in C_\epsilon(\boldsymbol{x})}{\arg\max}\ \ell(\boldsymbol{x}_{adv}, \boldsymbol{x}; \theta_T), \tag{1}$$

where ℓ is the Kullback-Leibler divergence loss (D_{KL}) $D_{\mathrm{KL}}(T(\boldsymbol{x}_{adv}), T(\boldsymbol{x}))$ [50]. However, vanilla attack methods [50, 13, 31, 20] can only create one adversarial example, which could be insufficient to explore entire space of possible teacher model perturbations. In addition, the works [42, 34] illustrate even attacks with random initialization can also lie together and lose diversity that reduces the quality of adversarial examples.

We propose to improve the vanilla adversarial attack optimization (Eq. 1) with a teacher adversarial local distribution (TALD), which captures the distribution of all teacher adversarial examples around clean input $\boldsymbol{x}$ within the constraint $C_\epsilon(\boldsymbol{x})$, as shown in Eq. 2.

$$\begin{aligned} P_{\theta_T}(\boldsymbol{x}_{adv}|\boldsymbol{x}) :&= \frac{e^{\ell(\boldsymbol{x}_{adv}, \boldsymbol{x}; \theta_T)}}{\int_{C_\epsilon(\boldsymbol{x})} e^{\ell(\boldsymbol{x}'_{adv}, \boldsymbol{x}; \theta_T)} d\boldsymbol{x}'_{adv}} \\ &= \frac{e^{\ell(\boldsymbol{x}_{adv}, \boldsymbol{x}; \theta_T)}}{M(\boldsymbol{x}; \theta_T)}, \end{aligned} \tag{2}$$

where $P_{\theta_T}(\cdot|\boldsymbol{x})$ is the teacher conditional adversarial local distribution over $C_\epsilon(\boldsymbol{x})$, and a normalization function is $M(\boldsymbol{x}; \theta_T)$. Here we show that the TALD can sufficiently represent the entire space of possible teacher perturbations.

3.2. TALD approximation using multiple particle-based search

In this paper, we leverage a multiple particle-based search method named Stein Variational Gradient Descent (SVGD) [29] to find the TALD $P_{\theta_T}(\cdot|\boldsymbol{x})$ because finding the denominator $M(\boldsymbol{x}; \theta_T)$ term in the Eq. 2 is intractable. SVGD is a Bayesian inference algorithm that seeks a set of points (or particles) to approximate the target distribution using iterative gradient-based updates. It has a simple form that closely mimics the typical gradient descent for optimization. This makes SVGD highly flexible and scalable, and can be easily combined with various state-of-the-art techniques responsible for the success of gradient optimization. While Markov chain Monte Carlo (MCMC) is often slow and has difficulty reaching convergence, SVGD efficiently approximates the target distribution by using an off-the-shelf optimization solver and is easily applicable to large datasets. It also enforces diversity of particles and works without explicit parametric assumptions in its solution, demonstrating better than other particle-based SGLD [45] and parametric-based method (with strong assumptions such as the target distribution follows Gaussian distributions) [16].

We denote $\boldsymbol{x}^1_{adv}, \boldsymbol{x}^2_{adv}, \boldsymbol{x}^3_{adv}, \ldots, \boldsymbol{x}^k_{adv} \sim P_{\theta_T}(\cdot|\boldsymbol{x})$, where $\boldsymbol{x}^i_{adv}$ is a i^{th} teacher adversarial example (named *teacher adversarial particle*), and $K = |\{1, 2, \ldots, k\}|$ is the number of adversarial examples. Here we show that our method can sufficiently explore the teacher decision boundaries by using multiple adversarial particles compared to vanilla attacking methods [50, 13, 31, 20]. SVGD is used to find a set of teacher adversarial particles to approximate the teacher adversarial local distribution $P_{\theta_T}(\cdot|\boldsymbol{x})$. First, the particles $\{\boldsymbol{x}^1_{adv}, \boldsymbol{x}^2_{adv}, \boldsymbol{x}^3_{adv}, \ldots, \boldsymbol{x}^k_{adv}\}$ are initialized by adding uniform noises to $\boldsymbol{x}$ constrained within the $C_\epsilon(\boldsymbol{x})$. They are then iteratively updated as well as projected to $C_\epsilon(\boldsymbol{x})$ until reaching the termination conditions (line 4 in Alg. 1). The normalization function $M(\boldsymbol{x}; \theta_T)$ is estimated based on the number of particles (K), which is implicitly demonstrated in the mean operator of line 5 - Alg. 1. Moreover, the two terms of line 5 in Alg. 1 have different major roles: (i) the first one transports the adversarial particles more toward to the high density areas of $P_{\theta_T}(\cdot|\boldsymbol{x})$ and (ii) the second term prevents all particles from collapsing into local modes of $P_{\theta_T}(\cdot|\boldsymbol{x})$ (e.g., pushing the particles away for enhancing the particle diversity). We empirically use l_2 normalization ($norm_2$), and radial basis function kernel $F(\boldsymbol{x}', \boldsymbol{x}) = \exp\left\{\frac{-||\boldsymbol{x}' - \boldsymbol{x}||^2}{2\sigma^2}\right\}$ with σ=1e-3 in this paper. We show that our method in a generalization of previous attacks when $K = 1$ from an asymptotic analysis of adversarial local distribution approximation section in the supplementary material.

3.3. Teacher adversarial local distribution (TALD) regularization

In this section, we propose our Teacher Adversarial Local Distribution (TALD) regularization. Recall that we form the TALD which is approximated using the adversarial par-

```
Input: clean example $\boldsymbol{x} \sim P_{\mathbb{D}}$. Number of adversarial particles $K$. Radius $\epsilon$ of the constraint $C_\epsilon$. Normalization function $norm_p$. Initial noise factor $\tau$. Uniform noise $U(-\epsilon, \epsilon)$. Step size updating particles $\eta$. Number of iterations $L$. Kernel function $F$.
Initialise a set of $K$ particles and project to the $C_\epsilon(x)$ constraint $\{\boldsymbol{x}_{adv}^i \in \mathbb{R}^d, i \in \{1, 2, \dots, k\} | \boldsymbol{x}_{adv}^i = \prod_{C_\epsilon}(\boldsymbol{x} + \tau * U(-\epsilon, \epsilon))\}$;
for $l = 1$ to $L$ do
    for $i = 1$ to $K$ do
        $\boldsymbol{x}_{adv}^{i,(l+1)} = \prod_{C_\epsilon}\left(\boldsymbol{x}_{adv}^{i,(l)} + \eta * norm_p(\phi(\boldsymbol{x}_{adv}^{i,(l)}))\right)$;
        where $\phi(\boldsymbol{x}_{adv}) = \frac{1}{K}\sum_{j=1}^{K}\Big[ F(\boldsymbol{x}_{adv}^{j,(l)}, \boldsymbol{x}_{adv})\nabla_{\boldsymbol{x}_{adv}^{j,(l)}} \log P(\boldsymbol{x}_{adv}^{j,(l)}|\boldsymbol{x}) + \nabla_{\boldsymbol{x}_{adv}^{j,(l)}} F(\boldsymbol{x}_{adv}^{j,(l)}, \boldsymbol{x}_{adv})\Big]$;
    end
end
return $\{\boldsymbol{x}_{adv}^{1,(L)}, \boldsymbol{x}_{adv}^{2,(L)}, \dots, \boldsymbol{x}_{adv}^{k,(L)}\}$;
Output: Set of adversarial particles $\{\boldsymbol{x}_{adv}^1, \boldsymbol{x}_{adv}^2, \dots, \boldsymbol{x}_{adv}^k\} \sim P_{\theta_T}(\cdot|\boldsymbol{x})$
```

Algorithm 1: Stein Variational Gradient Descent solver to approximate the teacher adversarial local distribution $P_{\theta_T}(\cdot|\boldsymbol{x})$.

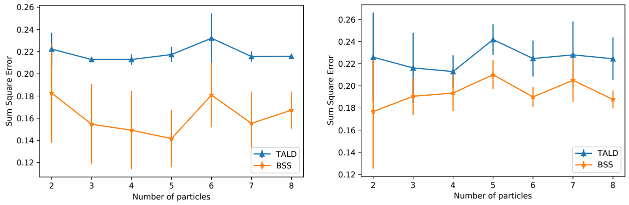
ticles generated by SVGD. We now illustrate how to use the teacher adversarial particles for knowledge distillation. We propose the TALD regularization term (ℓ_{TALD}) with respect to the student parameters θ_S at a position $\boldsymbol{x}$ with label y:

$$\ell_{TALD} := \min_{\theta_S} \mathbb{E}_{\boldsymbol{x}_{adv} \sim P_\theta(\cdot|\boldsymbol{x})} \Big[\|\ell^{CE}(T(\boldsymbol{x}_{adv}), y) - \ell^{CE}(S(\boldsymbol{x}_{adv}), y)\|_2^2 \Big], \tag{3}$$

where ℓ^{CE} is the cross-entropy loss function.

For each $\boldsymbol{x}$, SVGD samples K adversarial particles from the high density areas of $P_{\theta_T}(\cdot|\boldsymbol{x})$ to sufficiently explore the teacher decision boundaries, while the vanilla attacking methods [50, 13, 31, 20] only generate one adversarial example. We use these adversarial particles to regularize decision boundaries of the student by matching the cross-entropy loss between the teacher and student model, as shown in Eq. 3.

Here we show how to apply TALD regularization to existing knowledge distillation methods. The adversarial particles are generated with the differentiable teacher model and do not need additional learnable modules, as shown in Alg. 1. Therefore, we can easily combine to existing knowledge distillation losses, as shown in Eq. 4.



(a) resnet32x4

(b) wrn-40-2

Figure 1: Diversity comparison of our method and BSS with random initialization using sum of square error (SSE) using the pre-trained (a) resnet32x4 and (b) wrn-40-2 architectures. The figure illustrates the average of mean (point) and standard deviation (bar) of the three different inputs from CIFAR-100.

$$\min_{\theta_S} \mathbb{E}_{(\boldsymbol{x},y)\sim P_{\mathbb{D}}} \Big[\ell_S + \ell_{KD} + \ell_{AM} + \lambda \ell_{TALD} \Big], \tag{4}$$

where ℓ_S is the student cross-entropy loss $\ell^{CE}(S(\boldsymbol{x}), y)$. ℓ_{KD} is the original knowledge distillation loss proposed by Hinton et al. [21]. ℓ_{AM} can be an additional loss from other existing methods such as FitNet [40], CRD [43], etc.. λ is the weighted loss hyper-parameter[1].

4. Experiments

In this section, we conduct various experiments on CIFAR-100 [27] and ImageNet [14]. In Section 4.1, we compare the diversity between the adversarial particles generated by our method and adversarial examples from BSS with random initialization. We then show that TALD regularization can improve the performance of many existing methods in Section 4.2 and 4.3. We evaluate decision boundary similarity between the teacher and student in Section 4.4. The effect of the number particles to the student is studied in Section 4.5. Please refer to the supplementary material for all experimental settings.

4.1. Diversity of teacher adversarial particles vs. random initialization

Setting. We use pre-trained classifiers (e.g., resnet32x4 and wrn-40-1 architecture) on CIFAR-100 in this experiment. All pre-trained models are fixed. BSS [20] is an attack method that can generate one adversarial example at one run. We run BSS multiple times with random initialization to generate adversarial examples compared to the adversarial particles using our method. We set the same radius ball constraint ϵ, updating step η, and uniform noise

[1] Note that we ignore other weighted loss hyper-parameters.

factor τ initialization. Note that all adversarial examples and particles fool the classifiers.

Experimental setup. We randomly select three images from CIFAR-100 dataset. Given these inputs, we generate adversarial examples using BSS with random initializations. The adversarial particles are generated by our method with different numbers of particles, as shown in Fig. 1. We then calculate sum squared error (SSE) between these particles to evaluate their diversity. At each setting of the number of particles, we calculate the average of the means and standard deviations of SSE.

Result. Note that the advantages are illustrated in the Alg. 1 where the first and the second term of SVGD can sample in the high density areas and enforce diverse adversarial particles from the local distribution, respectively. Previous attack methods [50, 13, 31, 20] can generate multiple adversarial examples using random initialization but it can lie together and lose diversity [42]. Therefore, the adversarial particles from our method are diverse. In Fig. 1, our method has bigger SSE compared to BSS with random initialization because generated samples are more diverse.

4.2. TALD regularization with existing methods on CIFAR-100

Setting. In this experiment, we evaluate TALD regularization on model compression of a large network (teacher T) to a smaller one (student S). We use CIFAR-100 [27], which contains 50K training images with 500 images per class and 10K test images. We apply our TALD regularization to improve performance of existing methods using CIFAR-100. The existing methods is implemented from RepDistiller[2] and HSAKD[3] repositories. Our regularization is combined with these existing methods without changing parameter settings on CIFAR-100. We set the radius constraint $\epsilon = 0.3$, number of particles $K = 4$, and $\lambda = 0.01$.

Experimental setup. The goal of knowledge distillation is to improve performance of the student S by using the teacher knowledge. In this experiment, all teacher T models are pre-trained on CIFAR-100 and fixed. The accuracy of all models trained on CIFAR-100 with only ℓ_S is shown in Table 1. The TALD regularization is intensively evaluated on many existing methods such as KD [21], FitNet [40], AT [49], SP [44], CC [38], VID [1], RKD [36], PKT [37], AB [20], FT [24], NST [22], CRD [43], BSS [20], and HSAKD[11]. We setup various teacher-student neural network architectures for the same architecture style (Fig. 2) and across-architecture style (Fig. 3) knowledge distillation settings. BSS[20] is an attack proposed for the knowledge distillation task. Therefore, we compare BSS with KD to our TALD with KD[21], as shown in Fig. 4.

Architecture	Accuracy (%)
wrn-40-2	75.61
wrn-40-1	71.98
wrn-16-2	73.26
resnet56	72.34
resnet20	69.06
resnet32x4	79.42
resnet8x4	72.5
ShuffleNetV1	70.5
MobileNetV2	64.6
ResNet50	79.34

Table 1: Test accuracy (%) of different pre-trained model architectures on CIFAR-100. Note that all test accuracies are used from [43, 11].

Result. Recall that the proposed method is an additional regularization loss, which can combine with many existing methods. Our regularization explores the teacher decision boundaries using the teacher adversarial particles, then enforces decision boundary matching between the teacher and student loss. In Fig. 2, the teacher and student are from the same architectural style. When adding our TALD loss, we consistently improve test accuracy. In the context of transfer across very different teacher and student, we also increase the performance of existing methods in Fig. 3. Additionally, our method (KD+TALD) outperforms the adversarial approach BSS [20] for distillation (KD+BSS) shown in Fig. 4.

4.3. TALD regularization with existing methods on ImageNet

Setting. In this experiment, TALD regularization is evaluated on a large-scale ImageNet [14] dataset (1.2 million for training and 50K for validation images with 1K classes). We adopt the implementation of existing methods from Torchdistill[4][32], ResNet-34 as the teacher and ResNet-18 as the student. The ResNet-34 and ResNet-18 architectures are released by the PyTorch team. We keep all original settings of [32] and set our TALD following $\lambda = 0.001$, number of particles $K = 4$, and the radius constraint $\epsilon = 0.3$.

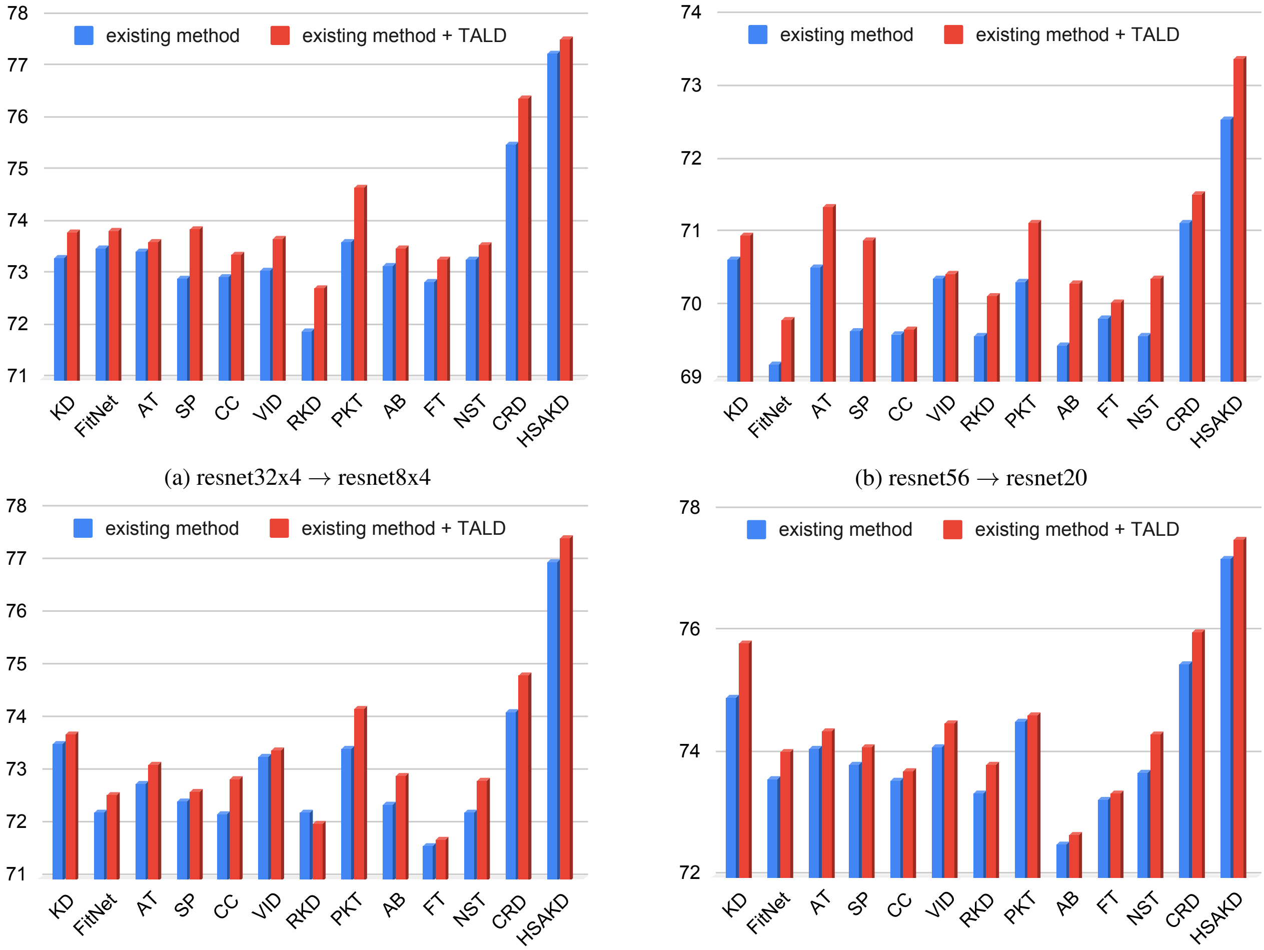
Experiment setup. We illustrate the performance of TALD regularization by compressing the teacher ResNet-34 to the student ResNet-18. The teacher is fixed and pre-trained with 73.31% accuracy. The base student trained on ImageNet without distillation methods achieves 69.75% accuracy. We combine our method to improve the implemented existing methods such as KD [21], AT [49], FT [24], CRD [43], and SSKD [47].

Result. We calculate the accuracy of students on 50K validation images. In Fig. 5, the all student accuracies are used from the implementation of the Torchdistill reposi-

[2] https://github.com/HobbitLong/RepDistiller

[3] https://github.com/winycg/HSAKD

[4] https://github.com/yoshitomo-matsubara/torchdistill



(a) resnet32x4 → resnet8x4

(b) resnet56 → resnet20

(c) wrn-40-2 → wrn-40-1

(d) wrn-40-2 → wrn-40-1

Figure 2: Test accuracy (%) of student networks on CIFAR-100 of a number of distillation methods from teacher to student (teacher → student). *Existing method* denotes a previous distillation method, while *existing method + TALD* is a combination of the respective existing method and our regularization. All student accuracies of existing methods are used from [43, 11].

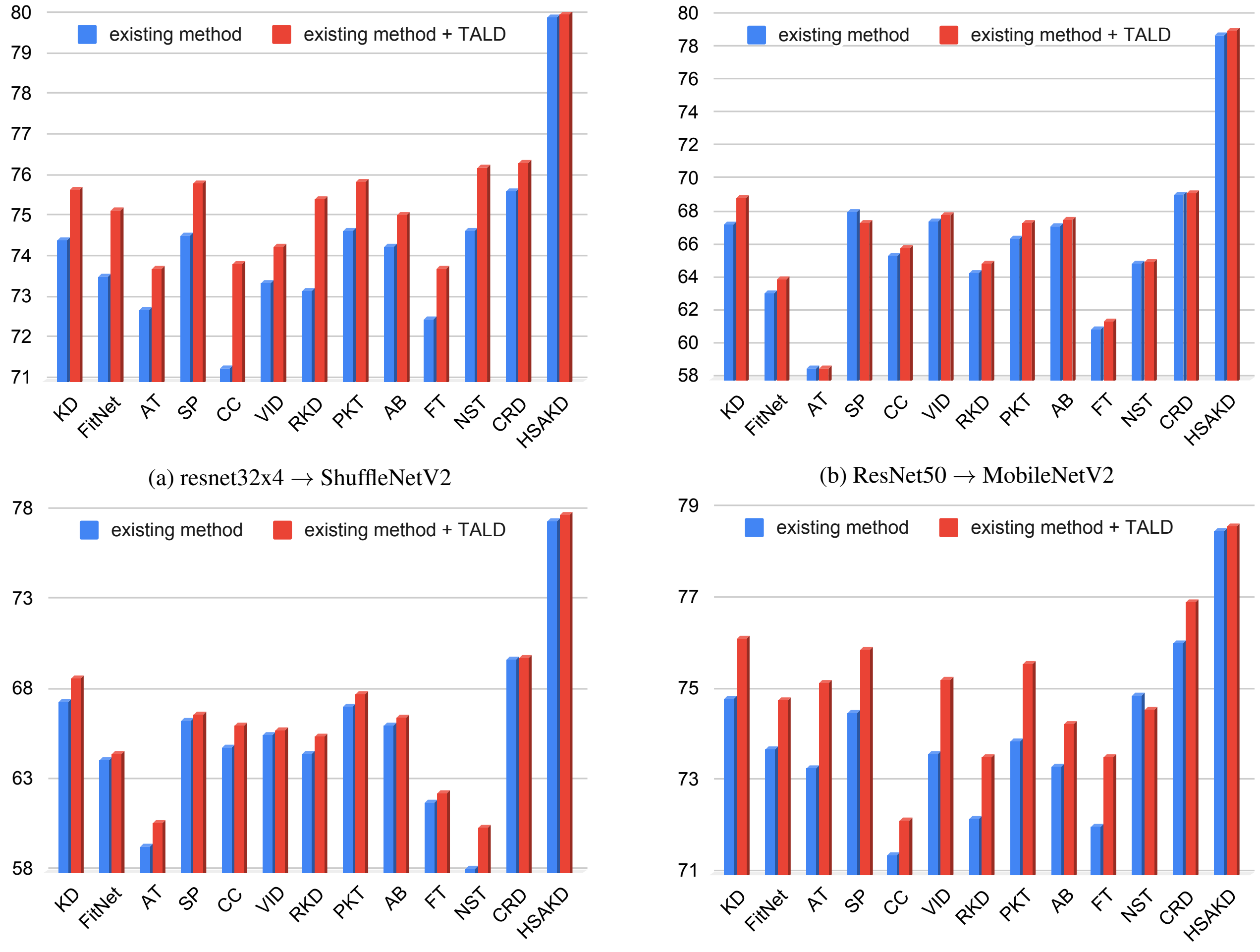
tory [32]. As can be seen, our TALD regularization can consistently improve the accuracy of the ResNet-18 student on top of respecting existing methods such as KD [21], AT [49], FT [24], CRD [43], and SSKD [47].

4.4. Decision boundary similarity evaluation

Metrics for similarity of decision boundaries. To verify our TALD regularization, we use metrics proposed by Heo et al. [20] to measure the similarity between the decision boundaries of two classifiers (e.g., teacher and student in the knowledge distillation task). The metrics are calculated using BSS attack. For each data point $\boldsymbol{x}$, BSS attacks the teacher and student to generate teacher $\boldsymbol{x}_{adv}^{T}$ and student $\boldsymbol{x}_{adv}^{S}$ adversarial example, respectively. We then obtain the perturbation vector of teacher ($\boldsymbol{v}^{T} = \boldsymbol{x}_{adv}^{T} - \boldsymbol{x}$) and student ($\boldsymbol{v}^{S} = \boldsymbol{x}_{adv}^{S} - \boldsymbol{x}$). Since the perturbation vector is obtained by the attacking path from clean sample $\boldsymbol{x}$ to model decision boundaries, we compare the *Magnitude Similarity (MagSim)* and *Angle similarity (AngSim)* of the two vectors. *MagSim* represents the similarity with respect to the distance from the clean sample $\boldsymbol{x}$ to the decision boundary, while *AngSim* reflects it with respect to the path direction from the clean sample $\boldsymbol{x}$ to the decision boundary. These two metrics have values in the range of [0,1] and higher values represent more similar decision boundaries. Please refer to the work [20] for more information.

Setup. We use pre-trained teachers and distilled students using CIFAR-100 from Section 4.2. Our baseline is KD without regularization. Our method is compared to KD + BSS, which uses adversarial examples to support student decision boundaries. We calculate *MagSim* and *AngSim*, as shown in Fig. 6.

Result. Recall that KD method does not have decision boundary regularization, while KD + BSS insufficiently ex-



(a) resnet32x4 → ShuffleNetV2

(b) ResNet50 → MobileNetV2

(c) vgg13 → MobileNetV2

(d) wrn-40-2 → ShuffleNetV1

Figure 3: Test accuracy (%) of student networks on CIFAR-100 of a number of distillation methods for transfer across very different teacher to student architectures (teacher → student). *Existing method* denotes a previous distillation method without TALD regularization term, while *existing method* + *TALD* is a combination of the respective existing method and our regularization. All student accuracies of existing methods are used from [43, 11].

plores the teacher perturbations [42]. In Fig. 6, our KD + TALD method can consistently improve decision boundary matching based on *MagSim* and *AngSim* metrics with various architectures such as wrn-40-2 → wrn-16-2, resnet56 → resnet20, and resnet32x4 → resnet8x4.

4.5. Teacher adversarial particle analysis

Setting. We study the number of teacher adversarial particles that affect the performance of the student on CIFAR-100. We perform the model compression task (teacher → student) on the same architecture style (teacher: resnet56 → student: resnet20) and very different architecture style (teacher: wrn-40-2 → student: ShuffleNetV1) with different knowledge distillation methods. The implementation adopts the RepDistill, and all parameters are kept similar to Section 4.2 settings except the number of particles K.

Experiment setup. We change the number of teacher adversarial particles K in $\{0, 1, 2, 4, 8\}$. When $K = 0$ implies that we do not use the TALD regularization. We study our method using different knowledge distillation methods with different K such as KD [21], AT [49] and SP [44] for resnet56 → resnet20, and KD [21], VID [1] and FT [24] for wrn-40-2 → ShuffleNetV1.

Result. Note that we approximate the teacher adversarial local distribution $P_{\theta_T}(\cdot|\boldsymbol{x})$ using the particles. Thus, by increasing the number of particles, we accordingly increase the regularization strength of the student model. Fig. 7 shows that the test accuracy can be improved by increasing K from 0 to 4. It is as expected that over regularization may hurt the performance when $K = 8$ on Fig. 7(a). However,

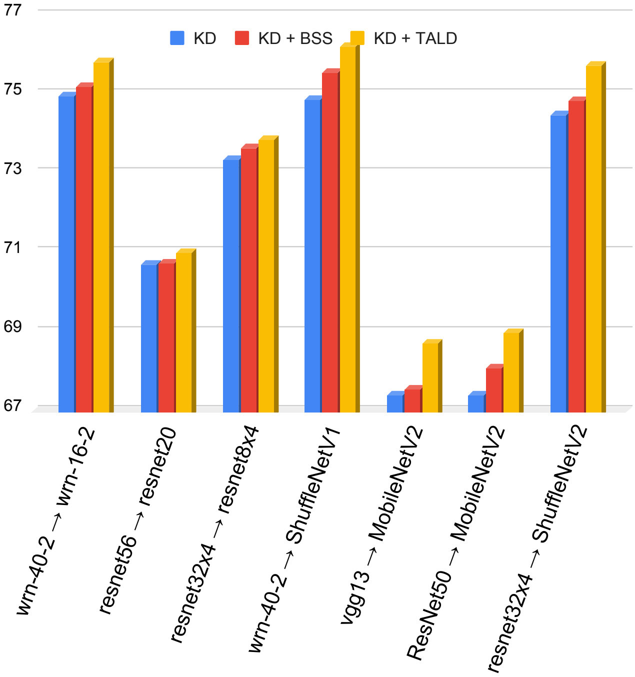


Figure 4: Test accuracy (%) of student networks on CIFAR-100 of KD, KD +BSS, and KD + TALD for transfer various teacher and student architectures (teacher → student).

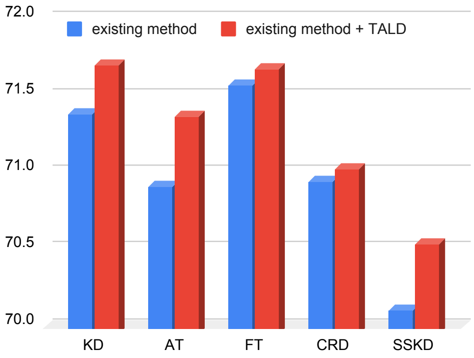
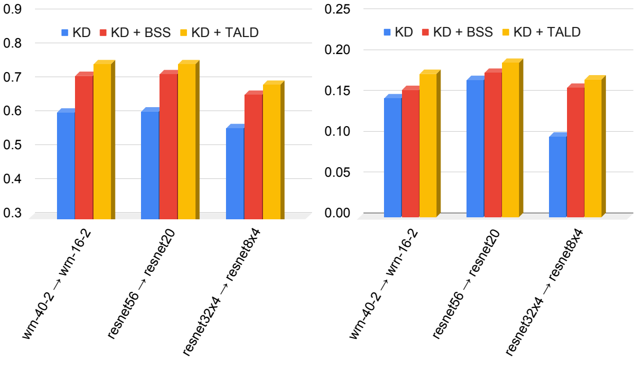


Figure 5: Accuracy (%) of ResNet-18 student on validation ImageNet dataset (ResNet-34 → ResNet-18). All student accuracies of existing methods are used from [32].

our method can still outperform existing methods without TALD regularization ($K = 0$) in these cases.

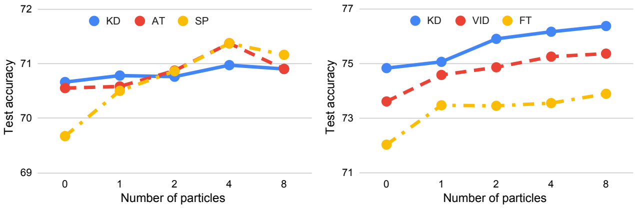
5. Conclusion and future work

In this paper, we have introduced a novel teacher adversarial local distribution (TALD) regularization that can adapt well to improve on many existing methods such as KD [21], FitNet [40], AT [49], SP [44], CC [38], VID [1], RKD [36], PKT [37], AB [20], FT [24], NST [22], CRD [43],and HSAKD[11]. In the proposed method, we form the teacher adversarial local distribution for exploring the teacher's properties (e.g., decision boundaries). Our strategy uses SVGD to estimate the adversarial local distribution using more diverse adversarial particles. We intensively conduct experiments on CIFAR-100 and ImageNet where the TALD consistently improves the performance of many existing knowledge distillation methods. By using a few adversarial particles, we improve the student performance at the cost of increasing the training time. In the future, we would like to reduce the TALD running time and use targeted attack perspective using TALD.



(a) MagSim (b) AngSim

Figure 6: Evaluation on decision boundary similarity between teacher and student (teacher → student) using *Magnitude Similarity (MagSim)* and *Angle similarity (AngSim)*. These two metrics have values in the range of [0,1] and higher values represent more similar decision boundaries.



(a) resnet56 → resnet20 (b) wrn-40-2 → ShuffleNetV1

Figure 7: Test accuracy (%) of the students when distilling from teacher to student (teacher → student) at different number of teacher adversarial particles $K \in \{0, 1, 2, 4, 8\}$. When $K = 0$ implies that we do not use TALD regularization.

Acknowledgements

This work was partially supported by the Australian Defence Science and Technology (DST) Group under the Next Generation Technology Fund (NGTF) scheme.

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of CVPR*, pages 9163–9171, 2019.

[2] Tao Bai, Jinnan Chen, Jun Zhao, Bihan Wen, Xudong Jiang, and Alex Kot. Feature distillation with guided adversarial contrastive learning. *arXiv preprint arXiv:2009.09922*, 2020.

[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[4] Anh Bui, Trung Le, He Zhao, Paul Montague, Olivier deVel, Tamas Abraham, and Dinh Phung. Improving adversarial robustness by enforcing local and global compactness. In *European Conference on Computer Vision*, pages 209–223. Springer, 2020.

[5] Anh Tuan Bui, Trung Le, Quan Hung Tran, He Zhao, and Dinh Phung. A unified wasserstein distributional robustness framework for adversarial training. In *International Conference on Learning Representations*, 2022.

[6] Anh Tuan Bui, Trung Le, He Zhao, Paul Montague, Olivier deVel, Tamas Abraham, and Dinh Phung. Improving ensemble robustness by collaboratively promoting and demoting adversarial robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6831–6839, 2021.

[7] Xiaoyu Cao and Neil Zhenqiang Gong. Mitigating evasion attacks to deep neural networks via region-based classification. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, pages 278–287, 2017.

[8] Alvin Chan, Yi Tay, and Yew-Soon Ong. What it thinks is important is important: Robustness transfers through input gradients. In *Proceedings of CVPR*, pages 332–341, 2020.

[9] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *Proceedings of AAAI*, volume 35, pages 7028–7036, 2021.

[10] Ting-Wu Chin, Cha Zhang, and Diana Marculescu. Renofeation: A simple transfer learning method for improved adversarial robustness. In *Proceedings of CVPR*, pages 3243–3252, 2021.

[11] Linhang Cai Chuanguang Yang, Zhulin An and Yongjun Xu. Hierarchical self-supervised augmented knowledge distillation. In *Proceedings of IJCAI*, pages 1217–1223, 2021.

[12] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[13] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of ICML*, pages 2206–2216. PMLR, 2020.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Processings of CVPR*, pages 248–255. Ieee, 2009.

[15] Qianggang Ding, Sifan Wu, Hao Sun, Jiadong Guo, and Shu-Tao Xia. Adaptive regularization of labels. *arXiv preprint arXiv:1908.05474*, 2019.

[16] Yinpeng Dong, Zhijie Deng, Tianyu Pang, Jun Zhu, and Hang Su. Adversarial distributional training for robust deep learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Proceedings of NeurIPS*, volume 33, pages 8270–8283, 2020.

[17] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of AAAI*, volume 34, pages 3996–4003, 2020.

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *Proceeding ICLR*, 2014.

[19] Ruifei He, Shuyang Sun, Jihan Yang, Song Bai, and Xiaojuan Qi. Knowledge distillation as efficient pre-training: Faster convergence, higher data-efficiency, and better transferability. In *Proceedings of CVPR*, pages 9161–9171, 2022.

[20] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge distillation with adversarial samples supporting decision boundary. In *Proceedings of the AAAI*, volume 33, pages 3771–3778, 2019.

[21] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[22] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

[23] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. *Proceedings of Advances in Neural Information Processing Systems*, 33:16199–16210, 2020.

[24] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Proceedings of NeurIPS*, 31, 2018.

[25] Seung Wook Kim and Hyo-Eun Kim. Transferring knowledge to smaller network with class-distance loss. In *Proceedings of ICLR*, 2017.

[26] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *Proceeding of ICLR*, 2017.

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

[28] Trung Le, Anh Tuan Bui, He Zhao, Paul Montague, Quan Tran, Dinh Phung, et al. On global-view based defense via adversarial attack and defense risk guaranteed bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 11438–11460. PMLR, 2022.

[29] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Proceedings of NeurIPS*, volume 29, 2016.

[30] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[31] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of ICLR*, 2018.

[32] Yoshitomo Matsubara. torchdistill: A modular, configuration-driven framework for knowledge distillation. In *International Workshop on Reproducible Research in Pattern Recognition*, pages 24–44. Springer, 2021.

[33] Rafael Muller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? In *Proceeding of NeurIPS*, 2019.

[34] Thanh Nguyen-Duc, Trung Le, He Zhao, Jianfei Cai, and Dinh Phung. Particle-based adversarial local distribution regularization. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 5212–5224. PMLR, 28–30 Mar 2022.

[35] Thanh Nguyen-Duc, He Zhao, Jianfei Cai, and Dinh Phung. Med-tex: Transfer and explain knowledge with less data from pretrained medical imaging models. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2022.

[36] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of CVPR*, pages 3967–3976, 2019.

[37] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of ECCV*, pages 268–284, 2018.

[38] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of ICCV*, pages 5007–5016, 2019.

[39] Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Proceedings of NeurIPS*, volume 32, 2019.

[40] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[41] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[42] Yusuke Tashiro, Yang Song, and Stefano Ermon. Diversity can be transferred: Output diversification for white-and black-box attacks. *Proceedings of NeurIPS*, 33, 2020.

[43] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

[44] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of ICCV*, pages 1365–1374, 2019.

[45] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of ICML*, pages 681–688, 2011.

[46] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of CVPR*, pages 501–509, 2019.

[47] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. In *Processings of ECCV*, pages 588–604. Springer, 2020.

[48] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of CVPR*, pages 3903–3911, 2020.

[49] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[50] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of ICML*, pages 7472–7482. PMLR, 2019.

[51] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of CVPR*, pages 11953–11962, 2022.

[52] Guorui Zhou, Ying Fan, Runpeng Cui, Weijie Bian, Xiaoqiang Zhu, and Kun Gai. Rocket launching: A universal and efficient framework for training well-performing light net. In *Proceedings of AAAI*, volume 32, 2018.