AGKD-BML: Defense Against Adversarial Attack by Attention Guided Knowledge Distillation and Bi-directional Metric Learning

Hong Wang[1]*, Yuefan Deng[1], Shinjae Yoo[2], Haibin Ling[1], Yuewei Lin[2]†
[1]Stony Brook University, Stony Brook, NY, USA
[2]Brookhaven National Laboratory, Upton, NY, USA
{hong.wang.2,yuefan.deng,haibin.ling}@stonybrook.edu, {sjyoo,ywlin}@bnl.gov

Abstract

While deep neural networks have shown impressive performance in many tasks, they are fragile to carefully designed adversarial attacks. We propose a novel adversarial training-based model by Attention Guided Knowledge Distillation *and* Bi-directional Metric Learning *(AGKD-BML). The attention knowledge is obtained from a weight-fixed model trained on a clean dataset, referred to as a teacher model, and transferred to a model that is under training on adversarial examples (AEs), referred to as a student model. In this way, the student model is able to focus on the correct region, as well as correcting the intermediate features corrupted by AEs to eventually improve the model accuracy. Moreover, to efficiently regularize the representation in feature space, we propose a bidirectional metric learning. Specifically, given a clean image, it is first attacked to its most confusing class to get the forward AE. A clean image in the most confusing class is then randomly picked and attacked back to the original class to get the backward AE. A triplet loss is then used to shorten the representation distance between original image and its AE, while enlarge that between the forward and backward AEs. We conduct extensive adversarial robustness experiments on two widely used datasets with different attacks. Our proposed AGKD-BML model consistently outperforms the state-of-the-art approaches. The code of AGKD-BML will be available at:* `https://github.com/hongw579/AGKD-BML`.

1 Introduction

Deep neural networks (DNNs) have achieved great breakthrough on a variety of fields, such as computer vision [22], speech recognition [17], and natural language processing [8]. However, their vulnerability against the so-called
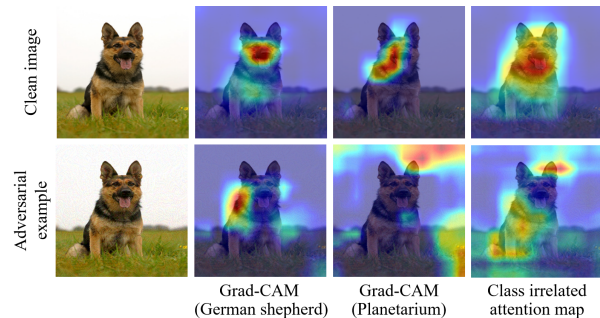
---

*This work was done during the research assistantship at BNL.
†Corresponding author.

Figure 1. A clean image ("German shepherd") and its adversarial example (incorrectly classified as "Planetarium") are in the first column. The class relevant attention maps (Grad-CAM) of correct and incorrect labels, and the class irrelevant attention maps are shown in the second, third and fourth columns, respectively. It shows that the adversarial perturbations corrupt the attention maps.

adversarial examples (AEs), which are the data with carefully designed but imperceptible perturbations added, has drawn significant attention [38]. The existing of AEs is a potential threat for the safety and security of DNNs in real-world applications. Thus, many efforts have been made to defend against adversarial attacks as well as improve the adversarial robustness of the machine learning model. In particular, adversarial training [16, 27]-based models are among the most effective and popular defending methods. Adversarial training solves a min-max optimization problem, in which the inner problem is to find the strongest AE within an ϵ−ball by maximizing the loss function, while the outer problem is to minimize the classification loss of the AE. Madry *et al.* [27] provided a multi-step projected gradient descent (PGD) model, which has become the standard model of the adversarial training. Following PGD, a number of recent works have been proposed to improve adversarial training from different aspects, *e.g.*, [6, 11, 28, 32, 35, 42, 49, 51].

However, the adversarial training-based models still suffer from relatively poor generalization on both clean and adversarial examples. Most of the existing adversarial train-

ing based models focus only on the on-training model that utilizes adversarial examples, which may be corrupted, but have not well explored the information from the model trained on clean images. In this work, we aim to improve the model adversarial robustness by distilling the attention knowledge and utilizing bi-directional metric learning.

The attention mechanism plays a critical role in human visual system and is widely used in a variety of application tasks [34, 53]. Unfortunately, one of our observations shows that the perturbations in the adversarial example (AE) will be augmented through the network, and thus significantly corrupts the intermediate features and attention maps. It is shown in the Figure 1, the AE confuses the model by letting it focus on different regions from the clean image. Intuitively, if we can transfer the knowledge of clean images from the teacher model to the student model to 1) obtain right attention information, and 2) correct the intermediate features corrupted by AE, we should be able to improve the model's adversarial robustness.

With this motivation, we propose an *Attention Guided Knowledge Distillation* (AGKD) module, which applies knowledge distillation (KD) [18] to efficiently transfer attention knowledge of the corresponding clean image from the teacher model to the on-training student model. Specifically, the teacher model is pre-trained on the original clean images and will be fixed during training, while the student model is the on-training model. The attention map of a clean image obtained from the teacher model is used to guide the student model to generate the attention map of the corresponding AE against the perturbations.

We further use t-distributed Stochastic Neighbor Embedding (t-SNE) to study the behavior of the AE in the latent feature space (see Figure 3), and observe that the representations of the AE are usually far away from their original class, similar as shown in [28]. While AGKD transfers information of clean image to the student model from the teacher model and thus provides the constraints on the similarity between the AE and its corresponding clean image, there is no constraint of samples from different classes taken into account. Previous works [24, 28, 52] proposed using metric learning to regularize the latent representations of different classes. Specifically, a triplet loss is utilized, in which latent representations of the clean image, its corresponding AE and an image from another class are considered as the *positive*, *anchor*, and *negative* example, respectively. However, this strategy only considers the *one-directional* adversarial attack, *i.e.*, from the clean image to its adversarial example, making it less efficient.

To address the above issue, we propose a *Bi-directional attack Metric Learning* (BML) to provide a more efficient and strong constraint. Specifically, the original clean image (*positive*) is first attacked to its *most confusing class*, which is the class that has the smallest loss other than the correct label, to get the *forward* adversarial example (*anchor*). Then, a clean image is randomly picked from the most confusing class and is attacked to the original image to get the *backward* adversarial example as the *negative*.

By integrating AGKD and BML, our **AGKD-BML** model outperforms the state-of-the-art models on two widely used datasets, CIFAR-10 and SVHN, under different attacks. In summary, our contribution is three-fold:

- An attention guided knowledge distillation module is proposed to transfer attention information of clean image to the student model, such that the intermediate features corrupted by adversarial examples can be corrected.

- A bidirectional metric learning is proposed to efficiently constrain the representations of the different classes in feature space, by explicitly shortening the distance between original image and its forward adversarial example, while enlarging the distance between the forward adversarial example and the backward adversarial example from another class.

- We conduct extensive adversarial robustness experiments on the widely used datasets under different attacks, the proposed AGKD-BML model outperforms the state-of-the-art approaches with both the qualitative (visualization) and quantitative evidence.

2. Related Works

Adversarial Attacks. Generally, there are two types of adversarial attacks: *white-box* attack where the adversary has full access to the target model, including the model parameters, and the *black-box* attack, where the adversary has almost no knowledge of the target model. For white-box attack, Szegedy *et al.* [38] discovered the vulnerability of deep networks against adversarial attacks. They used a box-constrained L-BFGS method to generate effective adversarial attacks. After that, several algorithms were developed to generate adversarial examples. As a one-step attack, the fast gradient sign method (FGSM) proposed in [16] uses the sign of the gradient to generate attacks, with ℓ_∞-norm bound. In [23], Kurakin *et al.* extended FGSM by applying it iteratively and designed basic iterative method (BIM). A variant of BIM was proposed in [12] by integrating momentum into it. DeepFool [29] tried to find the minimal perturbations based on the distance to a hyperplane and quantify the robustness of classifiers. In [31], the authors introduced a Jacobian-based Saliency Map Attack. The projected gradient descent (PGD) was proposed in [27] as a multi-step attack method. The CW attack, a margin-based attack, was proposed in [4]. Recently, Croce *et al.* introduced a parameter-free attack named AutoAttack [10],

which is an ensemble of two proposed parameter-free versions of PGD attacks and the other two complementary attacks, *i.e.*, FAB [9] and Square Attack [1]. It evaluates each sample based on its worst case over these four diverse attacks which includes both white-box and black-box ones. Besides the additive attacks, [14, 15, 20] show that even small geometric transformations, such as affine or projective transformation can fool a classifier. In addition to those attacks on the image input to the model, attempts are made to design adversarial patches that can fool the model in the physical world [13, 19, 23]. On the other side of the coin, adversarial attacks may also be used to improve the model performance [44, 25, 30].

Adversarial defense. Adversarial training-based models, which aim to minimize the classification loss to the strongest adversarial examples (maximal loss within a $\epsilon-$ball), are believed as one of the most effective and widely used defense methods. In practice, they iteratively generate adversarial examples for training. In [16], Goodfellow *et al.* generated the adversarial examples by FGSM, while Madry *et al.* [27] used the Projected Gradient Descent (PGD) attacks during adversarial training. Many variants based on adversarial training were proposed in recent years. For example, [35] computed the gradient for attacks and the gradient of model parameters at the same time, and significantly reduced the computation time. Adversarial logit paring [21] constraints distance between the logits from a clean image and its adversarial example, while [28] and [52] built a triplet loss between a clean image, its corresponding adversarial example and a negative sample. TRADES [51] optimized the trade-off between robustness and accuracy. In [41], the authors designed an adversarial training strategy with both adversarial images and adversarial labels. In [49], feature scattering is used in the latent space to generate adversarial examples and further improved the model's accuracy under different attacks. Xie *et al.* [45] proposed feature denoising models by adding denoise blocks into the architecture to defend the attack.

Most of the existing adversarial training-based models focus on the on-training model that utilizes adversarial examples, which may be corrupted, but have not explored the information from the model trained on clean images.

Other adversarial defense models. In [26, 46], the authors proposed to firstly detect and reject adversarial examples. Several methods proposed to estimate the clean image by using a generative model [36, 37, 47]. Cohen *et al.* [7] proposed to use randomized smoothing to improve adversarial robustness. There are also several works utilized large scale external unlabeled data to improve the adversarial robustness, *e.g.*, [5] and [39].

In this paper, we focus on improving the adversarial robustness of the model itself without using external data or pre-processing the testing data.

3 Proposed Method

In this section, we present the framework of our proposed AGKD-BML model in detail. As illustrated in Figure 2, AGKD-BML framework consists of two modules, *i.e.*, the attention guided knowledge distillation (AGKD) module and the bidirectional attack metric learning (BML) module. The AGKD module is used for distilling attention knowledge of the clean image to the student model, to obtain a better attention map for adversarial example, as well as correcting the corrupted intermediate features. The BML module efficiently regularizes the representation in feature space by using bidirectional metric learning. In the rest of this section, we first briefly introduce the standard adversarial training (AT) and (non-)targeted adversarial attack, and then describe the two modules of our proposed model and the integration of them.

3.1 Preliminaries

We first briefly describe the standard adversarial training (AT) [27]. Suppose we have a labeled C-class classification dataset $\mathcal{D} = \{(x, y)\}_{i=1}^{N}$ of N samples, where the label $y \in \{1, 2, \ldots, C\}$. There are two types of adversarial attacks, *i.e.*, the non-targeted attacks and the targeted attacks, which can be formulated as eq. (1) and (2), respectively:

$$\max_{\delta \in \Delta} \mathcal{L}(f_\theta(x + \delta), y) \tag{1}$$

$$\min_{\delta \in \Delta} \mathcal{L}(f_\theta(x + \delta), y_t) \tag{2}$$

where δ is the perturbation added to the image x, Δ provides an ℓ_∞-norm bound of the perturbation, and $f_\theta(\cdot)$ and $\mathcal{L}(\cdot)$ to denote the network with model parameters θ and the loss function, respectively. Non-targeted attacks maximize the loss function given the correct label y, while targeted attacks minimize the loss function given the target label y_t.

Standard AT uses non-targeted PGD (Projected Gradient Descent) attack [27] during training, which can be formulated as a min-max optimization problem:

$$\min_\theta \mathbb{E}_{x \in \mathcal{D}} \left[\max_{\delta \in \Delta} \mathcal{L}(f_\theta(x + \delta), y) \right] \tag{3}$$

In the objective function, the outer minimization is the update of the model parameters while the inner maximization is for generating adversarial attacks. Specifically, PGD is used to generate attacks, which is an iterative non-targeted attack with random start at the beginning. In this paper, following [41], we use targeted attacks during training with the most confusing class as target class.

3.2 Attention guided knowledge distillation

To distill attention information of clean images to the student model, we propose an attention guided knowledge
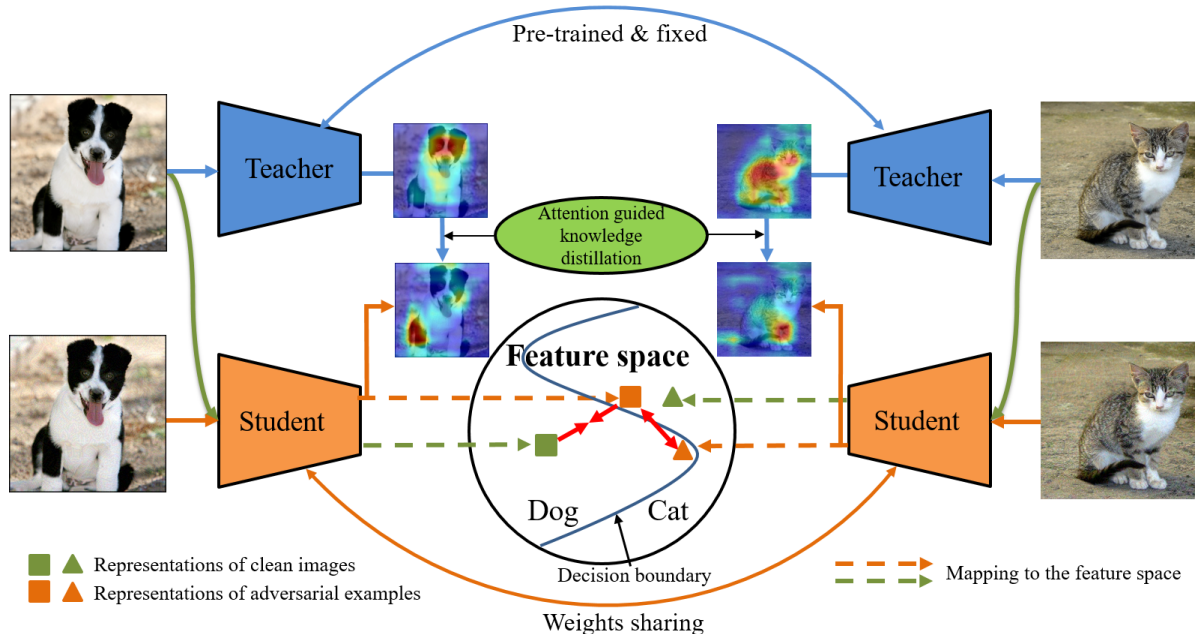
Figure 2. The framework of the proposed AGKD-BML model. Top-left is a clean image that belongs to "Dog", and bottom-left is its adversarial example (AE) targeted attack to its most confusing class "Cat". Similarly, top-right and bottom-right are a clean image of "Cat" and its AE targeted "Dog", respectively. "Teacher" is the model pre-trained on clean images and "Student" is the on-training model. The AE will fool the model by 1) focusing on incorrect regions, and 2) crossing the decision boundary in feature space. Attention guided knowledge distillation, illustrated as a green ellipse, is used for correcting the focus region. Bidirectional metric learning, illustrated as red arrows in the "Feature space", is used to pull the AEs back to their original classes. Better viewed in color.

distillation module. Figure 1 shows attention maps of a clean image ("German shepherd") and its adversarial example ("Planetarium"). As a class relevant attention map, the Grad-CAM [34] shows the focusing region related to a specific class. From the figure we can see that although the adversarial example degrades the attention map of original class, it hurts the attention map of target (incorrect) class much more largely, and makes the features of incorrect class overwhelm the correct class and dominate overall (as such makes the model mis-classified). We argue that only distilling the class relevant attention information has limited effects on correcting the features of the targeted class. Therefore, we propose to distill class irrelevant attention information (see section 3.2.1) of clean images. We provide more explanations and discussions to justify our choice in *supplementary materials.*

3.2.1 Class irrelevant attention map

We generate the class irrelevant attention map at the last convolutional layer. Specifically, we treat the backbone neural network until the last convolutional layer as a feature extractor, denoted by $\mathcal{F}(x)$ for a given image x, where $\mathcal{F}(x) \in \mathbb{R}^{C \times H \times W}$. We then produce an operator, denoted by $\mathcal{A}(\cdot)$, to map the feature map to the two-dimensional attention map, $\mathcal{A}(\mathcal{F}(x)) \in \mathbb{R}^{1 \times H \times W}$. In this paper, we sim-

ply pick the average pooling through the channel dimension (or identical weights 1×1 convolution) as $\mathcal{A}$.

3.2.2 Knowledge distillation

The knowledge distillation (KD) [18] utilizes a *student-teacher (S-T) learning framework* to transfer information learned from the teacher model to the student model. In this paper, we treat the model trained on the natural clean images by standard training as the teacher model and the one under adversarial training as the student model. The attention information is what we expect to transfer from the teacher model to the student model. As the teacher model is trained on the clean images with high testing accuracy, it is able to provide correct regions that model should focus on. Therefore, the attention map of the clean image extracted by the teacher model will transfer to the student model. The loss function of this attention guided knowledge distillation is written as:

$$\mathcal{L}_{KD}(x_t, x_s) = D(\mathcal{A}(\mathcal{F}_t(x_t)), \mathcal{A}(\mathcal{F}_s(x_s))) \qquad (4)$$

where x_t and x_s are input images of the teacher and student model, respectively, and $\mathcal{F}_t$ and $\mathcal{F}_s$ are feature extractors of the teacher and student models, respectively. $D(\cdot)$ is the distance function (*e.g.*, ℓ_1) to measure the similarity between these two attention maps. Given an adversarial example,

the AGKD guides the student model to focus on the same regions as its clean image.

3.3 Bidirectional attack metric learning

In our work, we use the targeted attack to obtain the adversarial examples. Let x^s refers to a samples with the label $y = s$, and x_t^s refers to an adversarial example of x^s with the target label $y_t = t$. In this paper, the forward adversarial example is targeted towards the most confusing class, which is defined as follow:

$$y_{mc} = \arg\min_{y \neq s} \mathcal{L}(f(x^s), y). \tag{5}$$

Given an original clean image x^s, we first generate the targeted adversarial example $x_{y_{mc}}^s$ towards its most confusing class. Then, we randomly select a sample $x^{y_{mc}}$ from the most confusing class, and generate its adversarial example $x_s^{y_{mc}}$ that targeted back to the original label s. We utilize x^s, $x_{y_{mc}}^s$, $x_s^{y_{mc}}$ as positive, anchor and negative sample, respectively. The triplet loss is defined as:

$$\begin{aligned} & \mathcal{L}_{tr}(x_a, x_p, x_n) \\ & = [d(E(x_a), E(x_p)) - d(E(x_a), E(x_n)) + m]_+, \end{aligned} \tag{6}$$

where x_a, x_p, x_n denote to positive, anchor and negative samples, respectively. $E(\cdot)$ is the representation from the penultimate layer of the model. $d(a, b)$ denotes the distance between two embeddings a and b, which is defined as the angular distance $d(a, b) = 1 - \frac{|a \cdot b|}{\|a\|_2 \|b\|_2}$, following [28]. m is the margin. Comparing to the previous metric learning based adversarial training, *e.g.*, [28] and [52] which only consider forward adversarial example, we consider both the forward and backward adversarial examples. Therefore, we name it the bidirectional metric learning.

By adding a ℓ_2-norm regularization on the embedding, the final BML loss function is written as:

$$\mathcal{L}_{BML} = \lambda_1 \mathcal{L}_{tr}(x_{y_{mc}}^s, x^s, x_s^{y_{mc}}) + \lambda_2 \mathcal{L}_{norm}, \tag{7}$$

where $\mathcal{L}_{norm} = \|E(x_{y_{mc}}^s)\|_2 + \|E(x^s)\|_2 + \|E(x_s^{y_{mc}})\|_2$ is the normalization term, and λ_1 and λ_2 are the trade-off weights for the two losses.

3.4 Integration of two modules

We integrate the attention guided knowledge distillation and bidirectional metric learning together to take the benefits from both modules. As we consider bidirectional adversarial attack, we have two clean/adversarial image pairs, $x^s/x_{y_{mc}}^s$ and $x^{y_{mc}}/x_s^{y_{mc}}$. For both pairs, we apply the AGKD from the attention map of the clean image obtained by teacher model to the student model, which can be formulated as:

$$\mathcal{L}_{AGKD} = \mathcal{L}_{KD}(x^s, x_{y_{mc}}^s) + \mathcal{L}_{KD}(x^{y_{mc}}, x_s^{y_{mc}}) \tag{8}$$

where the first term denotes the AGKD loss for the forward attack pair, *i.e.*, x^s and $x_{y_{mc}}^s$, while the second term denotes the backward attack pair, *i.e.*, $x^{y_{mc}}$ and $x_s^{y_{mc}}$,

By combining the standard cross entropy loss used in the traditional adversarial training, the BML loss, and the AGKD loss, the final total loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \mathcal{L}_{AGKD} + \mathcal{L}_{BML} \tag{9}$$

The overall procedure of AGKD-BML model is shown in Algorithm. 1.

Algorithm 1: AGKD-BML model

Input: Clean image set $\mathcal{D}$, epoch number N, batch size b, learning rate γ
Output: Network parameter θ

1 **for** epoch $= 1, ..., N$ **do**
2   **for** minibatch $\{x_i, y_i\}_{i=1}^b$ **do**
3   initialize $\mathcal{L}_{batch} = 0$;
4   **for** sample one x^s belongs to class s **do**
5   a. find its MC class y_{mc} by Eq. 5, sample one data $x^{y_{mc}}$ from class y_{mc} ;
6   b1. obtain $x_{y_{mc}}^s$ by attacking x^s to y_{mc} ;
7   b2. obtain $x_s^{y_{mc}}$ by attacking $x^{y_{mc}}$ to s ;
8   c. calculate $\mathcal{L}_{BML}$ by Eq. 7 ;
9   d. calculate $\mathcal{L}_{AGKD}$ by Eq. 8 ;
10   e. calculate $\mathcal{L}_{total}$ by Eq. 9 ;
11   f. update $\mathcal{L}_{batch} = \mathcal{L}_{batch} + \frac{1}{b}\mathcal{L}_{total}$.
12   update $\theta = \theta - \gamma \cdot \nabla_\theta \mathcal{L}_{batch}$.

13 **return** θ;

---

4 Experiments

4.1 Experimental settings

Dataset We evaluate our method on two popular datasets: CIFAR-10 and SVHN. *CIFAR-10* consists of 60k 3-channel color images with size of 32×32 in 10 classes, in which 50k images for training and 10k images for testing. *SVHN* is the street view house number dataset, which has 73257 images for training and 26032 images for testing. We evaluate model on a larger datasets: Tiny ImageNet, and the results are shown in ***supplementary materials.***

Comparison methods We use comparison methods include: (1) *undefended model (UM)*, where the model is trained by standard training; (2) *adversarial training (AT)* [27], which uses non-targeted PGD adversarial examples (AEs) for training; (3) *single-directional metric learning (SML)* [28]; (4) *Bilateral* [41], which generates AEs on both images and labels; (5) *feature scattering*

(FS) [49], where adversarial attacks for training are generated with feature scattering in the latent space; (6) and (7) utilize the *channel-wise activation suppressing (CAS)* [3] on TRADES [50] and MART [43], respectively, which showed the superior compared to the original version. Note that Bilateral, FS generate AEs by using single-step attacks *in training*, while AGKD-BML uses 2-steps attacks, "AT" and "SML" use 7-step attacks, and "TRADES+CAS" and "MART+CAS" use 10-step attacks. To fairly compare to these multi-step attack models, we also train a 7-step attack variant of AGKD-BML, referred as to "AGKD-BML-7". We test the models with various attacks including FGSM [16], BIM [23], PGD [27], CW [4], MIM [12] with different attack iterations. We also evaluate the models in a per-sample manner using AutoAttack (AA) [10], which is an ensemble of four diverse attacks. Finally, we also test the black-box adversarial robustness of the model.

Implementation details Following [27] and [28], we use Wide-ResNet (WRN-28-10) [48], and set the initial learning rate γ as 0.1 for CIFAR-10 and 0.01 for SVHN. We use the same learning rate decay points as [41] and [49], where decay schedule $[100, 150]$ for CIFAR-10 and $[60, 90]$ for SVHN, with 200 epochs in total. "AGKD-BML-7" has the learning rate that decays at 150 epochs and the training stops at 155 epochs, following the suggestions in [43, 33]. In training phase, the perturbation budget $\epsilon = 8$ and label smoothing equals to 0.5 following [49]. In the AGKD module, we adopt ℓ_1 norm to measure the similarity between attention maps. For the BML module, parameters are the same as [28], *i.e.*, margin $m = 0.03$, $\lambda_1 = 2$ and $\lambda_2 = 0.001$.

4.2 Evaluation of adversarial robustness

We evaluate our model's adversarial robustness and report the comparisons in Table 1. The results on "clean" images are used as a baseline for evaluating how much the accuracy of the defenders will drop as increasing the adversarial robustness. It is shown in Table 1, AGKD-BML overall outperforms the comparison methods on CIFAR-10. AGKD-BML also shows better adversarial robustness on SVHN dataset with a large margin.

Interestingly, in Table 1, we observed that AGKD-BML showed different superiors to different attacks, *i.e.*, AGKD-BML trained on 7-step attack has higher performance than that trained on 2-step attack against AA, but much lower performance against the regular attacks, *e.g.* PGD and CW. The reason for this phenomenon is, in our opinion, that compared to the regular attacks, AA is an ensemble of four different types of attack, including white-box and black-box ones, which requires the generalization capability of defense against different types of attack. The generation of the 7-step attack significantly increases the diversity of

AEs used for training and thus, it improves the robustness against AA with some sacrifice on accuracy against regular attacks. On the other hand, the generation of the 2-step attack focuses more on the regular attacks but less diverse, which makes it has lower performance against AA. As an empirical defense method, we argue the model trained by small-number-step attack is still useful in some scenarios that the adversarial attacks are known. We provide more results of AGKD-BML model trained on large-number-step attacks against AA in ***supplementary materials.***

4.3 Ablation study

We analyze the ablation effect of each component of AGDK-BML on CIFAR-10 dataset. The quantitative and qualitative results are shown in Table 2 and Figure 3, respectively. "UM", "AT" and "SML" are the same models described above. "BML" denotes the *bidirectional* metric learning without using any knowledge distillation. "AGKD" denotes the model applied attention map guided knowledge distillation without any metric learning. In Figure 3, we provide the t-SNE plots to show the sample representations in feature space. The triangle points with different colors represent the clean images in different classes, while the red circle points are the AEs under PGD-20 attack. We show AEs from two classes (*i.e.*, deer and frog).

"UM" shows how the adversarial attacks behave if a model dose not have any defense. A simple one step attack FGSM drops UM's accuracy to $\sim30\%$, while the multi-step attacks, *e.g.*, PGD-20 and CW-20, drop its accuracy to 0%. It is also visualized in the first column of Figure 3, where all the AEs locate far from their original class, and fit into the distributions of other classes. As a standard benchmark defense model, "AT" provides a baseline for improvements on both single-step and multi-step attacks.

The Effect of Bidirectional Metric Learning "SML" and "BML" both apply metric learning to constrain the clean image and its AE to keep a short distance, while push away the images in different classes. The difference between them is that SML only considers forward attacks and BML considers both forward and backward attacks. In the second and third columns of Figure 3, we can see that the SML does pull many of the AEs back to their original class, i.e., purple in the first row and cyan in the second row. However, one of the side effects of SML is that it makes classes confusing for clean images and thus may make a significant accuracy drop on clean images. In contrast, BML keeps better separations between different classes, and has much less amount of AEs located far away compared to SML. It demonstrates the benefit of the bidirectional strategy.

The Effect of Attention Guided Knowledge Distillation Utilizing "AGKD" alone is able to obtain a good accuracy, which is better than BML. By integrating both AGKD

Table 1. Evaluation results on CIFAR-10 and SVHN, under different widely used attacks. For CIFAR-10 dataset, we grouped the models by small-number or larger-number steps attack *in training*. "Bilateral" and "FS" use one-step attack, AGKD-BML uses two-step attack, and others use large-number steps with the step numbers show followed by the model names. "AGKD-BML-7" is a veriant of AGKD-BML that uses 7-step attack for training. The best accuracy for each group is illustrated as bold, and the overall best accuracy is highlighted.

CIFAR-10								
Attacks(steps)	clean	FGSM	BIM(7)	PGD (20)	PGD (100)	CW (20)	CW (100)	AA [10]
UM	95.99%	31.39%	0.38%	0%	0%	0%	0%	0%
Bilateral [41]	91.2%	70.7%	-	57.5%	55.2%	56.2%	53.8%	29.35%
FS [49]	90.0%	78.4%	-	70.5%	68.6%	62.4%	60.6%	36.64%
AGKD-BML	91.99%	76.69%	73.81%	71.02%	70.72%	63.67%	62.55%	37.07%
AT-7 [27]	86.19%	62.42%	54.99%	45.57%	45.22%	46.26%	46.05%	44.04%
SML-7 [28]	86.21%	58.88%	52.60%	51.59%	46.62%	48.05%	47.39%	47.41%
TRADES+CAS-10 [3]	85.83%	65.21%	-	55.99%	-	67.17%	-	48.40%
MART+CAS-10 [3]	86.95%	63.64%	-	54.37%	-	63.16%	-	48.45%
AGKD-BML-7	86.25%	70.06%	64.97%	57.30%	56.88%	53.36%	52.95%	50.59%

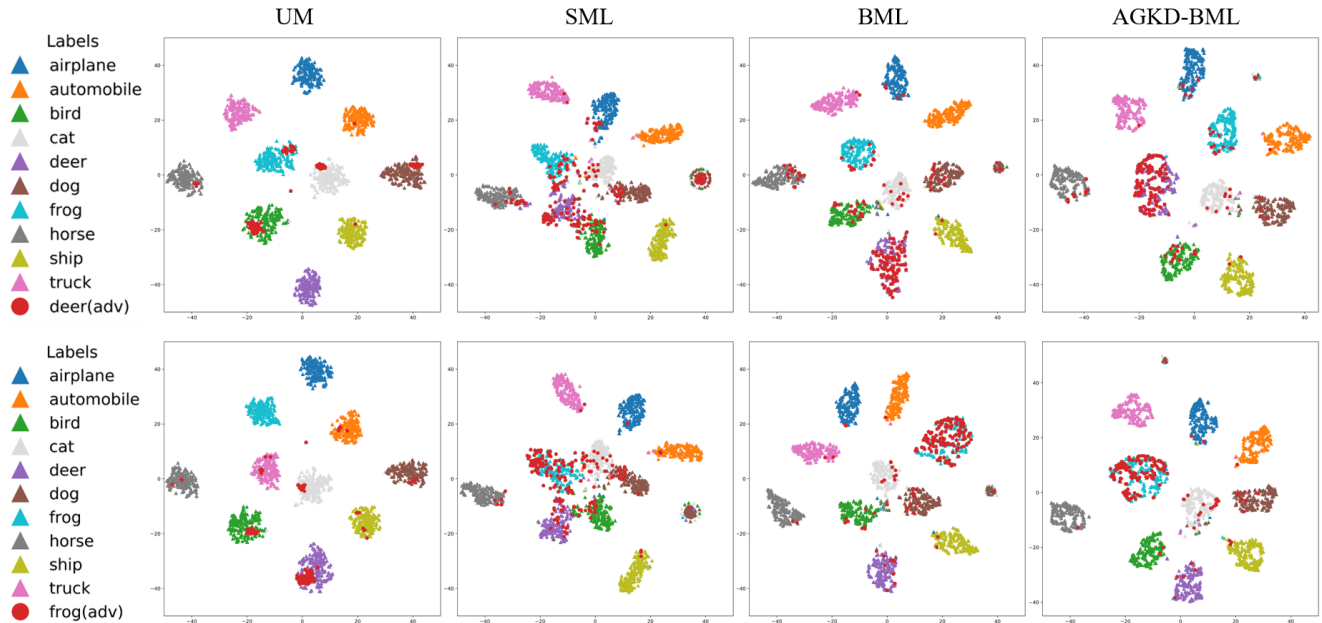| SVHN								
Attacks(steps)	clean	FGSM	BIM(10)	PGD (20)	PGD (100)	CW (20)	CW (100)	MIM (40)
UM	96.36%	46.33%	1.54 %	0.33%	0.22%	0.37%	0.24%	5.39%
Bilateral [41]	94.1%	69.8%	-	53.9%	50.3%	-	48.9%	-
FS [49]	96.2%	83.5%	-	62.9%	52.0%	61.3%	50.8%	-
AT-7 [27]	91.55%	67.13%	54.03%	45.64%	44.02%	47.14%	45.66%	52.13%
SML-7 [28]	83.95%	70.28%	57.58%	51.91%	49.81%	51.25%	49.31%	43.80%
TRADES+CAS-10 [3]	91.69%	70.79%	-	55.26%	-	60.10%	-	-
MART+CAS-10 [3]	93.05%	70.30%	-	51.57%	-	53.38%	-	-
AGKD-BML	95.04%	89.32%	75.06%	74.94%	69.23%	69.85%	62.22%	76.86%

Figure 3. t-SNE plots for illustrating the sample representations in feature space. The triangle points with different colors represent the clean images in different classes, and the red circle points are the adversarial examples under PGD-20 attack. Best viewed in color.

and BML, the proposed AGKD-BML obtain the best performance in terms of both quantitative and qualitative results. In the fourth column of Figure 3, AGKD-BML pulls most of the AEs back to their original class, while keeps better separation between classes than BML does. We also provide the attention maps of the AEs obtained by the trained models in Figure 4. Compared to AT, AGDK-BML obtains better attention maps which are more identical to the ones

Table 2. Ablation study on CIFAR-10 dataset.

	FGSM	PGD (20)	CW (20)
UM	31.39%	0%	0%
AT [27]	62.42%	45.57%	46.26%
SML [28]	58.88%	51.59%	48.05%
BML	71.08%	60.51%	56.53%
AGKD	75.57%	65.93%	60.71%
AGKD-BML	76.69%	71.02%	63.67%

Table 3. Black-box adversarial robustness.

AT [27]	FS [49]	Bilateral [41]	SML [28]	AGKD-BML
85.4%	88.9%	89.9%	86.4%	**90.75%**

of clean images. This suggests that the AGKD does help on correcting the representation of AEs in feature space.
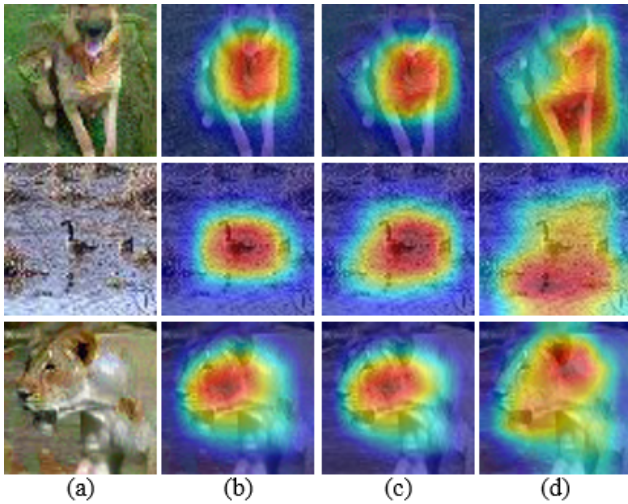
Figure 4. (a) Sample adversarial examples. (b) Attention maps obtained from clean images, which are treated as ground truth. (c) and (d) are the attention maps obtained by AGDK-BML and AT.

Different attack iterations and budgets We evaluate model robustness under different PGD attack iterations, and different attack budgets (ϵ) with a fixed attack iteration of 20. It is shown in Figure 5 that AGKD-BML consistently outperforms two comparison methods, *i.e.*, feature scatter (FS) [49] and standard AT, on all numbers of attack iterations up to 100 and all attack budgets up to $\epsilon = 20$. Moreover, AGKD-BML also shows more robust to large attack budgets as the accuracy drops are significantly less than the other two comparison methods.
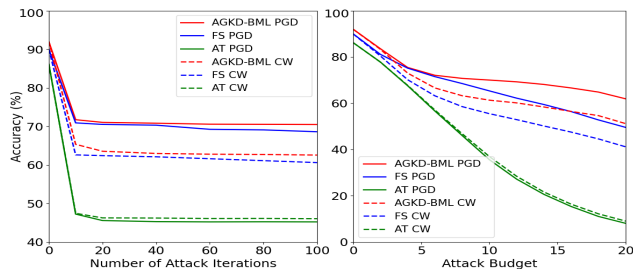

Figure 5. The accuracy under different attack iterations and perturbation budgets (ϵ).

4.4 Black box adversarial robustness

To evaluate the black-box adversarial robustness, i.e., the adversary has no knowledge about the model, we generate an AE for each clean image in CIFAR-10 testing set by using natural models under PGD-20 attack with $\epsilon = 8$. Then the AGKD-BML model, as well as the comparison models, are tested on the generated adversarial example data. As demonstrated in Table 3, AGKD-BML model achieves the best accuracy among the models suggesting that AGKD-BML is robust to the black-box attacks as well.

4.5 Discussion

Based on the analysis in [2], we claim that the robustness of our model is not from gradient obfuscation for the following reasons: 1) In table 1, iterative attacks are stronger than one-step attack (FGSM). 2) Figure 5 shows that the accuracy monotonically declines under attacks with more steps or increasing perturbation budgets. 3) Table 3 shows that black-box attacks have a lower success rate (higher accuracy) than white-box attacks. 4) We evaluate our model against a gradient-free attack [40] and the accuracy is 88.67%, which is higher than gradient-based attacks (71.02% for PGD20).

5 Conclusion

We proposed a novel adversarial training based model, named as *AGKD-BML*, that integrates two modules, *i.e.*, the attention guided knowledge distillation module and the bidirectional metric learning module. The first module transfers attention knowledge of the clean image from a teacher model to a student model, so as to guide student model for obtaining better attention map, as well as correcting the intermediate features corrupted by adversarial examples. The second module efficiently regularizes the representation in the feature space, by shortening the representation distance between original image and its forward adversarial example, while enlarging the distance between the forward and backward adversarial examples. Extensive adversarial robustness experiments on two popular datasets with various attacks show that our proposed AGKD-BML model consistently outperforms the state-of-the-art approaches.

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *ECCV*, pages 484–501. Springer, 2020.

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, pages 274–283, 2018.

[3] Yang Bai, Yuyuan Zeng, Yong Jiang, Shu-Tao Xia, Xingjun Ma, and Yisen Wang. Improving adversarial robustness via channel-wise activation suppressing. In *ICLR*, 2021.

[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, 2017.

[5] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi. Unlabeled data improves adversarial robustness. *NeurIPS*, 2019.

[6] Minhao Cheng, Qi Lei, Pin-Yu Chen, Inderjit Dhillon, and Cho-Jui Hsieh. Cat: Customized adversarial training for improved robustness. *arXiv:2002.06789*, 2020.

[7] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *ICML*, pages 1310–1320, 2019.

[8] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, pages 160–167, 2008.

[9] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *ICML*, pages 2196–2205, 2020.

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

[11] Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. In *ICLR*, 2020.

[12] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, pages 9185–9193, 2018.

[13] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *CVPR*, pages 1625–1634, 2018.

[14] Alhussein Fawzi and Pascal Frossard. Manitest: Are classifiers really invariant? In *BMVC*, pages 106.1–106.13, 2015.

[15] Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, and Andrew Ng. Measuring invariances in deep networks. In *NeurIPS*, pages 646–654, 2009.

[16] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[17] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

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

[19] Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *CVPR*, pages 720–729, 2020.

[20] Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Geometric robustness of deep networks: analysis and improvement. In *CVPR*, pages 4441–4449, 2018.

[21] Harini Kannan, Alexey Kurakin, and Ian J. Goodfellow. Adversarial logit pairing. *arXiv:1803.06373*, 2018.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012.

[23] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *ICLR Workshop*, 2017.

[24] Pengcheng Li, Jinfeng Yi, Bowen Zhou, and Lijun Zhang. Improving the robustness of deep neural networks via adversarial training with triplet loss. *arXiv:1905.11713*, 2019.

[25] Ping Liu, Yuewei Lin, Zibo Meng, Lu Lu, Weihong Deng, Joey Tianyi Zhou, and Yi Yang. Point adversarial self-mining: A simple method for facial expression recognition. *IEEE Transactions on Cybernetics*, pages 1–12, 2021.

[26] Jiajun Lu, Theerasit Issaranon, and David Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. In *ICCV*, pages 446–454, 2017.

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

[28] Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, and Baishakhi Ray. Metric learning for adversarial robustness. In *NeurIPS*, pages 480–491, 2019.

[29] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, pages 2574–2582, 2016.

[30] Pingbo Pan, Ping Liu, Yan Yan, Tianbao Yang, and Yi Yang. Adversarial localized energy network for structured prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5347–5354, 2020.

[31] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE European symposium on security and privacy*, pages 372–387, 2016.

[32] Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. Bit-flip attack: Crushing neural network with progressive bit search. In *ICCV*, pages 1211–1220, 2019.

[33] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *ICML*, pages 8093–8104. PMLR, 2020.

[34] 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 *ICCV*, pages 618–626, 2017.

[35] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *NeurIPS*, pages 3358–3369, 2019.

[36] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *ICLR*, 2018.

[37] Bo Sun, Nian-hsuan Tsai, Fangchen Liu, Ronald Yu, and Hao Su. Adversarial defense by stratified convolutional sparse coding. In *CVPR*, pages 11447–11456, 2019.

[38] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

[39] Jonathan Uesato, Jean-Baptiste Alayrac, Po-Sen Huang, Robert Stanforth, Alhussein Fawzi, and Pushmeet Kohli. Are labels required for improving adversarial robustness? *NeurIPS*, 2019.

[40] Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning*, pages 5025–5034. PMLR, 2018.

[41] Jianyu Wang and Haichao Zhang. Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks. In *ICCV*, pages 6629–6638, 2019.

[42] Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. In *ICML*, 2019.

[43] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*, 2019.

[44] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020.

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

[46] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv:1704.01155*, 2017.

[47] Jianhe Yuan and Zhihai He. Ensemble generative cleaning with feedback loops for defending adversarial attacks. In *CVPR*, pages 581–590, 2020.

[48] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, pages 87.1–87.12, 2016.

[49] Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. In *NeurIPS*, pages 1831–1841, 2019.

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

[51] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019.

[52] Yaoyao Zhong and Weihong Deng. Adversarial learning with margin-based triplet embedding regularization. In *ICCV*, pages 6549–6558, 2019.

[53] Bolei Zhou et al. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.