This WACV 2023 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Inducing Data Amplification Using Auxiliary Datasets in Adversarial Training

Saehyung Lee* Hyungyu Lee
Department of Electric and Computer Engineering
Seoul National University, Seoul 08826, South Korea
{halo8218, rucy74}@snu.ac.kr

Abstract

Several recent studies have shown that the use of extra in-distribution data can lead to a high level of adversarial robustness. However, there is no guarantee that it will always be possible to obtain sufficient extra data for a selected dataset. In this paper, we propose a biased multi-domain adversarial training (BiaMAT) method that induces training data amplification on a primary dataset using publicly available auxiliary datasets, without requiring the class distribution match between the primary and auxiliary datasets. The proposed method can achieve increased adversarial robustness on a primary dataset by leveraging auxiliary datasets via multi-domain learning. Specifically, data amplification on both robust and non-robust features can be accomplished through the application of BiaMAT as demonstrated through a theoretical and empirical analysis. Moreover, we demonstrate that while existing methods are vulnerable to negative transfer due to the distributional discrepancy between auxiliary and primary data, the proposed method enables neural networks to flexibly leverage diverse image datasets for adversarial training by successfully handling the domain discrepancy through the application of a confidence-based selection strategy. The pre-trained models and code are available at: `https://github.com/Saehyung-Lee/BiaMAT`.

1. Introduction

The usefulness of adversarial examples in training deep neural networks (DNNs) demonstrates that the method through which these structures perceive the world is markedly different from that employed by humans. Many approaches [1] have been proposed to bridge the gap in adversarial robustness between humans and DNNs. Among these, training based on the use of adversarial examples as training data is considered as the most effective method to improve the robustness of DNNs. Unfortunately, as demonstrated by Schmidt et al. [34], the sample complexity of adversarially robust generalization is substantially higher than that of standard generalization. To address this issue, several recent studies [4, 38] leveraged extra (in-distribution) unlabeled data and developed methods for improving the sample complexity of robust generalization. However, although such methods enable state-of-the-art adversarial robustness, they are not always capable of obtaining extra in-distribution data for any selected data distribution. In this paper, we propose a biased multi-domain adversarial training (BiaMAT) method to improve the adversarial robustness of a classifier on a primary dataset based on the use of publicly available (labeled) auxiliary datasets. The proposed method yields the desired effect based on the following assumption:

Figure 1. Visualization of robust and non-robust features [14, 21]

Assumption 1. *A common robust and non-robust feature space exists between the primary and auxiliary datasets.*

Figure 1 shows that robust features [21] exhibit human-perceptible patterns. We may assume that if two datasets are similar from a human perspective, then they share robust features. However, non-robust features are imperceptible to humans, thus, we cannot determine whether Assumption 1 is correct by a human perception. Fortunately, recent studies [28, 24] have provided empirical evidence in support of the presence of a common non-robust feature space among diverse image datasets. Therefore, unlike existing state-of-the-art methods [4, 38], which employ in-distribution data, under BiaMAT, the distribution of the auxiliary dataset and the corresponding primary dataset can differ. For example,

*Correspondence to: Saehyung Lee halo8218@snu.ac.kr.

by applying BiaMAT, we can leverage CIFAR-100 [22], Places365 [44], or ImageNet [8, 12] as an auxiliary dataset for adversarial training on CIFAR-10 [22].

The proposed method achieves an inductive transfer between adversarial training on the primary dataset (referred to as the "primary task") and auxiliary dataset (referred to as "auxiliary task"). In other words, BiaMAT learns primary and auxiliary tasks in parallel within the framework of multi-domain learning [27], and the inductive bias provided by the auxiliary tasks is transferred to the primary task through a common hidden structure. This mechanism can be considered to be an increase in the size of the training dataset [5]. In addition, based on studies that have demonstrated the presence of non-robust features [40, 21], we classify the effects of adversarial training into two types and demonstrate the usefulness of the proposed method irrespective of the type considered. In particular, we dissociate the compound effect of the proposed method into the effects of *non-robust feature regularization* and *robust feature learning* and assess the contribution of each through the use of the expectation of random labels [5, 24]. Our experimental results on the CIFAR datasets and ImageNet demonstrate that BiaMAT can effectively use training signals generated from various auxiliary datasets. Furthermore, we show that while existing methods are vulnerable to negative transfer due to the distributional discrepancy between auxiliary and primary data, the proposed method enables neural networks to flexibly leverage diverse image datasets for adversarial training by successfully dealing with domain discrepancy through the application of a confidence-based selection strategy.

2. Biased multi-domain adversarial training

2.1. Method

Existing state-of-the-art methods leveraging extra unlabeled data [4, 38] first remove out-of-distribution (OOD) data from a given auxiliary dataset, and then use the remaining data for training with pseudo-labels. Therefore, they are only effective when a given auxiliary dataset contains a large number of in-distribution data, and are suboptimal in terms of data utility. To maximize the data utility, we leverage given auxiliary data via multi-domain learning [13]. Multi-domain learning is a strategy for improving the performance of tasks that solve the same problem across multiple domains by sharing information across these domains. In standard settings, domains typically share semantic features. For example, in the Office dataset [33], data that belong to the same class (*e.g.*, keyboard) are separated into different domains (*e.g.*, amazon, webcam). In adversarial settings, by contrast, it is possible to find evidence for tighter-than-expected relationships between different datasets [28]. In particular, the use of domain-agnostic adversarial examples [28] and robust training methods that leverage different datasets [6, 24] demonstrates that common adversarial spaces can exist across different datasets. Therefore, our proposed method expands the range of related domains relative to that considered under standard settings with the goal of *maximizing the adversarial robustness of the classifier on one primary dataset*. In this respect, BiaMAT differs from standard multi-domain learning, for which the primary goal is increasing the average performance over multiple domains.

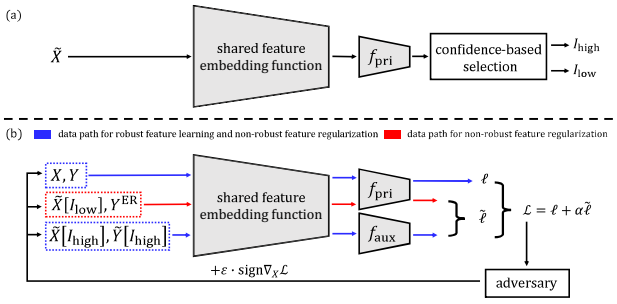

Figure 2. **Overview of BiaMAT.** We use a shared feature embedding function for both primary and auxiliary tasks, while prediction functions (f_{pri} and f_{aux}) are separated for each task. (a) At the first stage of BiaMAT, a confidence-based selection strategy classify given auxiliary data $\tilde{X}$ into two groups: auxiliary data that share robust and non-robust features with the primary dataset ($\tilde{X}[I_{\text{high}}]$); and that share only non-robust features with the primary dataset ($\tilde{X}[I_{\text{low}}]$). (b) We adversarially train the model on $\tilde{X}[I_{\text{high}}]$ and the primary data X using their ground truth labels ($\tilde{Y}[I_{\text{high}}]$ and Y) to achieve robust feature learning and non-robust feature regularization (indicated in blue). For $\tilde{X}[I_{\text{low}}]$, however, we use the expectation of random labels y^{ER} to induce data amplification only for non-robust feature regularization (indicated in red)

We classify given auxiliary data into two types: (i) auxiliary data that share robust and non-robust features [21] with the primary dataset; and (ii) that share only non-robust features with the primary dataset. We do not consider auxiliary data sharing only robust features with the primary dataset based on studies demonstrating the presence of a common non-robust feature space among diverse image datasets [36, 24]. As previous multi-domain learning studies observed [5, 27], using (i) will be beneficial to the primary task performances. By contrast, (ii) can help the primary task by regularizing the shared non-robust features but, at the same time, it can suppress the advantages of multi-domain learning by generating an inductive bias toward extraneous robust features–an effect, called "negative transfer". Therefore, we need a method that can acquire training signals for non-robust features without achieving inductive bias for robust features from (ii). In Sec. 2.2, we theoretically demonstrate that for each of the two data types, multi-domain learning can improve the adversarial robustness of the primary task in a simple Gaussian model. Especially, in the theoretical case of infinite batch size, we demonstrate that

the adversarial training on (ii) with random labels can improve adversarial robustness while avoiding negative transfer. Based on our theoretical analysis, we propose the use of the expectation of random labels (y^{ER}) on (ii). Interestingly, the use of y^{ER} is identical to the recently proposed OAT method [24], which uses OOD data to improve adversarial robustness. We compare our method with OAT in Appendix D. To classify given auxiliary data into the two types, our proposed method utilizes a confidence-based data selection strategy. Confidence scores are used in many research fields, including semi-supervised learning [37, 32] and OOD detection [18]. For example, in the case of semi-supervised learning, high-confidence unlabeled data are used with their pseudo-labels, while the remaining unlabeled data are not used. By contrast, in our proposed method, the use of y^{ER} naturally separates the auxiliary data according to their robust features, and adversarial training algorithms for each of the two data types are applied. We provide the detailed description of the confidence-based selection strategy in Sec. 2.4.

To sum up, we consider a multi-domain learning problem on a primary dataset $D \subset \mathcal{X} \times \mathcal{Y}$ and an auxiliary dataset $\tilde{D} \subset \mathcal{X} \times \tilde{\mathcal{Y}}$. $\tilde{D}$ is divided into two data subsets $\tilde{D}_{\text{high}}$ (high confidence) and $\tilde{D}_{\text{low}}$ (low confidence) through the confidence-based selection strategy. The hypotheses of the tasks are denoted by $h_{\text{pri}} : \mathcal{X} \to \mathcal{Y}$ and $h_{\text{aux}} : \mathcal{X} \to \tilde{\mathcal{Y}}$, where $h_{\text{pri}} = f_{\text{pri}} \circ g$ and $h_{\text{aux}} = f_{\text{aux}} \circ g$. Here, f_{pri} and f_{aux} are the prediction functions that output class probabilities for the primary and auxiliary datasets, respectively, and g is the shared feature embedding function. The loss function for D is defined as $\ell = \mathbb{E}_{(\boldsymbol{x},y)\sim D}\left[\ell_{\text{adv}}(\boldsymbol{x}, y; h_{\text{pri}}, S)\right]$, where ℓ_{adv} is the adversarial loss, and S represents the set of perturbations an adversary can apply. Any existing adversarial losses [25, 43] can be employed for ℓ_{adv}. In addition, the loss for $\tilde{D}$ is defined as $\tilde{\ell} = \frac{1}{|\tilde{D}|}\sum_{(\tilde{\boldsymbol{x}},\tilde{y})\sim\tilde{D}_{\text{high}}} \ell_{\text{adv}}(\tilde{\boldsymbol{x}}, \tilde{y}; h_{\text{aux}}, S) + \frac{1}{|\tilde{D}|}\sum_{(\tilde{\boldsymbol{x}},\tilde{y})\sim\tilde{D}_{\text{low}}} \ell_{\text{adv}}(\tilde{\boldsymbol{x}}, y^{\text{ER}}; h_{\text{pri}}, S)$. Our goal is to attain a small adversarial loss on the primary dataset. Thus, the proposed method minimizes the following loss:

$$\mathcal{L} = \ell + \alpha\tilde{\ell}, \quad \text{where } \alpha \in [0, 1]. \tag{1}$$

α is a hyperparameter that biases the multi-domain learning toward the primary task. Although we describe BiaMAT at the dataset level, we actually apply BiaMAT at the mini-batch level. Figure 2 provides an overview of BiaMAT.

2.2. Theoretical motivation

We analyze the proposed method from the perspective of *non-robust feature regularization* and *robust feature learning*, which are the two effects of adversarial training. In particular, we define a simple Gaussian model to demonstrate how the proposed method induces training data amplification using an auxiliary dataset that satisfies Assumption 1.

Preliminary. Tsipras et al. [40] described the effect of adversarial training by constructing a classification task through which training examples $(\boldsymbol{x}, y) \in \mathbb{R}^{d+1} \times \{\pm 1\}$ are drawn from a distribution, as follows:

$$\begin{gathered} y \overset{u.a.r}{\sim} \{-1, +1\}, \quad x_1 = \begin{cases} +y & \text{w.p. } p \\ -y & \text{w.p. } 1-p \end{cases}, \\ x_2, \ldots, x_{d+1} \overset{i.i.d.}{\sim} \mathcal{N}(\eta y, 1), \end{gathered} \tag{2}$$

where x_1 is a robust feature that robustly correlates to the label ($p \geq 0.5$), and the remaining features $x_2, \ldots, x_{d+1}$ are non-robust features that are vulnerable to adversarial attacks ($0 < \eta < \ell_\infty$-bound). For this data distribution, the authors demonstrated that the following linear classifier could attain a standard accuracy arbitrarily close to 100%, although it is susceptible to adversarial attacks:

$$f(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}_{\text{unif}}^\top \boldsymbol{x}), \text{ where } \boldsymbol{w}_{\text{unif}} = \left[0, \frac{1}{d}, \ldots, \frac{1}{d}\right]. \tag{3}$$

Notably, the following lemma indicates the importance of adversarial training:

Lemma 1. (Tsipras et al.) *Adversarial training results in a classifier that assigns zero weight to non-robust features* $x_2, \ldots, x_{d+1}$.

Lemma 1 shows that adversarial training (i) lowers the sensitivity of the classifier to non-robust features and (ii) achieves a certain level of classification accuracy by learning robust features. We refer to (i) and (ii) as *non-robust feature regularization* and *robust feature learning*, respectively.

Setup and overview. Given a shared feature embedding function $g : \mathcal{X} \to \mathcal{Z}$, we define primary and auxiliary data models in the feature space $\mathcal{Z}$, sampled from each of the following distributions:

$$\begin{gathered} \text{(Primary)} \quad y \overset{u.a.r}{\sim} \{-1, +1\}, \quad z_1 \sim \mathcal{N}(y, u^2), \\ z_2, \ldots, z_{d+1} \overset{i.i.d.}{\sim} \mathcal{N}(\eta y, 1), \\ \text{(Auxiliary)} \quad \tilde{y} = \text{sign}(\gamma) \cdot y, \quad \tilde{z}_1 \sim \mathcal{N}(y|\gamma|, v^2), \\ \tilde{z}_2, \ldots, \tilde{z}_{d+1} \overset{i.i.d.}{\sim} \mathcal{N}(\eta y|\gamma|, 1), \end{gathered} \tag{4}$$

where $\gamma \in [-1, 1]$ is the correlation coefficient between the two tasks. We use the accent (tilde) to represent variables associated with the "auxiliary task". In Eq. (4), z_1 is a robust feature that robustly correlates to the label, whereas the other features $z_2, \cdots, z_{d+1}$ are non-robust features that are vulnerable to adversarial attacks ($0 < \eta < \ell_\infty$-bound). If the two datasets are highly correlated in terms of robust and non-robust features, from Eq. (3), it is evident that the following linear classifiers can achieve a high standard accuracy on the

primary and auxiliary datasets, respectively, although they have low adversarial robustness:

$$\begin{aligned} \text{(Primary)} \quad p(y \mid \boldsymbol{z}) &= \frac{1-y}{2} + y\sigma(\boldsymbol{w}^\top \boldsymbol{z}), \\ \text{(Auxiliary)} \quad p(\tilde{y} \mid \tilde{\boldsymbol{z}}) &= \frac{1-\tilde{y}}{2} + \tilde{y}\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}), \end{aligned} \tag{5}$$

where $\boldsymbol{w} = \left[0, \frac{1}{d}, \ldots, \frac{1}{d}\right]$, and $\sigma(\cdot)$ denotes a sigmoid function. To study the effect of adversarial training on the auxiliary task on the non-robust feature regularization for the primary task, we derive the gradients of the primary and auxiliary adversarial losses with respect to the non-robust features, which are then back-propagated through the shared feature embedding function g. In addition, we demonstrate how the use of random labels enables us to dissociate the compound effect of the proposed method into the effects of non-robust feature regularization and robust feature learning.

Non-robust feature regularization. First, we generate the adversarial feature vector of our classification model ($\tilde{\boldsymbol{z}}^{\text{adv}} = g(\tilde{\boldsymbol{x}} + \boldsymbol{\delta}) : \boldsymbol{\delta} \in S$, where $\tilde{\boldsymbol{x}} \in \mathcal{X}$ denotes the auxiliary input vector). The objective function of the adversary to deceive our model is the cross-entropy loss [25].

Lemma 2. *Let $i \in \{2, \cdots, d+1\}$ and $\eta < \lambda < 1$. Then, the expectation of the adversarial feature vector against the auxiliary task is*

$$\mathbb{E}\left[\tilde{z}_1^{\text{adv}}\right] = y, \quad \mathbb{E}\left[\tilde{z}_i^{\text{adv}}\right] = (\eta - \lambda)y. \tag{6}$$

Proof is in Appendix A. The stochastic gradient descent to the cross-entropy loss on $\tilde{\boldsymbol{z}}^{\text{adv}}$ is applied to update our classification model. In particular, by deriving the auxiliary loss gradient with respect to the adversarial feature vector, we determine the training signals that are generated from the auxiliary task and transferred to the primary task through the shared feature embedding function.

Theorem 1. *Let $\ell(;\boldsymbol{w})$ and $\tilde{\ell}(;\gamma\boldsymbol{w})$ be the loss functions of the primary and auxiliary tasks, respectively, and $t = \frac{1}{2}(y+1)$. When the auxiliary data are closely related to the primary data from the perspective of robust and non-robust features, i.e., $|\gamma| = 1$, the expectation of the gradient of $\tilde{\ell}$ with respect to $\tilde{z}_i^{\text{adv}} : i \in \{2, \cdots, d+1\}$ is*

$$\mathbb{E}\left[\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i^{\text{adv}}}\right] = \frac{1}{d}\mathbb{E}\left[\sigma(\boldsymbol{w}^\top \boldsymbol{z}^{\text{adv}}) - t\right] = \mathbb{E}\left[\frac{\partial \ell}{\partial z_i^{\text{adv}}}\right]. \tag{7}$$

The theoretical results in the cases of $|\gamma| < 1$ (weak correlation) are discussed in Appendix A. From Lemma 2 and Theorem 1, for $i \in \{2, \cdots, d+1\}$, it can be seen that $\text{sign}\left(\mathbb{E}\left[\tilde{z}_i^{\text{adv}}\right]\right) = \text{sign}\left(\mathbb{E}\left[\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i^{\text{adv}}}\right]\right)$. That is, the application of a gradient descent guides the shared feature embedding function to pay less attention to non-robust features. In addition, Theorem 1 shows that if the auxiliary task is closely related to the primary task in terms of non-robust features, the training signals obtained from the auxiliary adversarial loss and back-propagated to the shared feature embedding function have the same effect as those of the primary task from the perspective of non-robust feature regularization. Therefore, this can be considered data amplification for non-robust feature regularization.

Robust feature learning. If $|\gamma| = 1$ and the weight value for the robust feature z_1 is non-zero, clearly, the auxiliary task on $\tilde{\boldsymbol{z}}^{\text{adv}}$ can induce data amplification for robust feature learning as well as non-robust feature regularization. However, when the auxiliary dataset contains extraneous robust features, the learning on $\tilde{\boldsymbol{z}}^{\text{adv}}$ may lead to negative transfer, which suppresses the advantages of multi-domain learning. To prevent inductive transfer between tasks, Caruana [5] shuffled the class labels among all samples in the auxiliary dataset. Similarly, to avoid negative transfer, the use of random labels can be considered in our case. To investigate the effect of adversarial training on the shuffled auxiliary dataset, we replace the true labels in the auxiliary data defined in Eq. (4) with random labels. That is, we define the auxiliary feature–label pairs, $(\tilde{\boldsymbol{z}}, q) \in \mathbb{R}^{d+1} \times \{\pm 1\}$, sampled from a distribution as follows:

$$\begin{gathered} \tilde{z}_1 \sim \mathcal{N}(y|\gamma|, v^2), \ \tilde{z}_2, \ldots, \tilde{z}_{d+1} \overset{i.i.d.}{\sim} \mathcal{N}(\eta y|\gamma|, 1), \\ q \overset{u.a.r}{\sim} \{-1, +1\}. \end{gathered} \tag{8}$$

For the case in which the auxiliary task is adversarially trained on $(\tilde{\boldsymbol{z}}, q)$ pairs with the cross-entropy loss, the following theorem can be proven:

Theorem 2. *Let $\tilde{\ell}(;\gamma\boldsymbol{w})$ be the loss function of the auxiliary task. Then, if $|\gamma| = 1$, with high probability, the signs of $\tilde{z}_i^{\text{adv}} : i \in \{2, \cdots, d+1\}$ and the auxiliary loss gradient with respect to $\tilde{z}_i^{\text{adv}}$ are*

$$\text{sign}\left(\tilde{z}_i^{\text{adv}}\right) = -\gamma q = \text{sign}\left(\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i^{\text{adv}}}\right). \tag{9}$$

Because the gradient with respect to $\tilde{z}_i^{\text{adv}}$ is of the same sign as $\tilde{z}_i^{\text{adv}}$ with high probability, the application of a gradient descent makes the shared feature embedding function to refrain from using non-robust features, thereby enabling the model to achieve non-robust feature regularization. Conversely, the shuffled dataset cannot provide any robust features because the use of random labels completely eliminates the relationship between images and labels. To further investigate the effect of adversarial training on the shuffled auxiliary dataset with regard to robust feature learning, we assign a positive number to the weight (defined in Eq. (5))

Table 1. Accuracy (under Autoattack [10]) comparison of the models adversarially trained [25] on CIFAR-10 using either $\tilde{y}$ or y^{ER} for the auxiliary dataset. The best result for each auxiliary dataset is indicated in bold; auxiliary datasets that produce better results when used with y^{ER} are indicated in red

$\tilde{y}/y^{\text{ER}}$	Auxiliary dataset				Baseline [25]
	SVHN	CIFAR-100	Places365	ImageNet	
$\tilde{y}$	47.44	48.48	48.88	**50.33**	48.53
y^{ER}	**48.53**	**49.89**	**49.24**	49.81	

corresponding to the robust feature z_1 and derive the training signals that are sent to the shared feature embedding function as follows:

Theorem 3. *Let $\tilde{\ell}(;\gamma \boldsymbol{w})$ be the loss function of the auxiliary task. Then, if $|\gamma| = 1$ and $w_1 > 0$, with high probability, the signs of $\tilde{z}_1^{\text{adv}}$ and the auxiliary loss gradient with respect to $\tilde{z}_1^{\text{adv}}$ are*

$$\text{sign}(\tilde{z}_1^{\text{adv}}) = y, \quad \text{sign}\left(\frac{\partial \tilde{\ell}}{\partial \tilde{z}_1^{\text{adv}}}\right) = -\gamma q. \tag{10}$$

Assuming that the classification model is still vulnerable to adversarial examples, $\left|\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i^{\text{adv}}}\right|$ is independent of q because an adversary can always yield a large loss regardless of q. Hence, in the theoretical case of an infinite batch size, the adversarial training on the shuffled auxiliary dataset will not affect the robust feature learning for the primary task because y and q are independent of each other, and q is sampled uniformly at random. In practice, however, the minibatch gradient descent is employed to train DNNs, and thus, unfavorable training signals can be generated from the auxiliary task on the shuffled dataset in terms of robust feature learning. To resolve this issue, we use the expectation of random labels instead of the one-hot random labels. Furthermore, to close the gap between the theory ($|\gamma| = 1$) and practice, we use a shared prediction function for the primary and the shuffled auxiliary data. In other words, we assign $y^{\text{ER}} = [\frac{1}{c}, \dots, \frac{1}{c}]$, where c is the number of the classes in the primary dataset, to the auxiliary data to observe only the effectiveness of non-robust feature regularization while excluding the contribution of robust feature learning.

2.3. Empirical evidence

To empirically demonstrate the arguments developed above, we conduct a test confirming the following two statements: (i) When $\tilde{D}$ has a weak relationship with D in terms of robust features, the use of y^{ER} ($I_{\text{high}} = \emptyset$ and $I_{\text{low}} = \{0, 1, \cdots, |\tilde{D}| - 1\}$ in Fig. 2) results in better adversarial robustness than that produced by the use of $\tilde{y}$ ($I_{\text{high}} = \{0, 1, \cdots, |\tilde{D}| - 1\}$ and $I_{\text{low}} = \emptyset$ in Fig. 2), and (ii) when $\tilde{D}$ is closely related to D in terms of robust features, the use of y^{ER} results in worse adversarial robustness than that produced by the use of $\tilde{y}$. Table 1 lists the results of executing the test using CIFAR-10 as the primary dataset D. Here, we use SVHN [29], CIFAR-100, and Places365 as auxiliary datasets that are weakly related to CIFAR-10 in terms of robust features, based on the previous OOD detection studies [18, 35]. In addition, we use ImageNet as an auxiliary dataset that is closely related to CIFAR-10 from the perspective of robust features [17, 7]. As shown, for SVHN, CIFAR-100, and Places365, the use of y^{ER} leads to better adversarial robustness than that induced by the use of $\tilde{y}$, even though image-label mappings in the auxiliary datasets are disrupted. These results demonstrate that the robust feature learning induced by using all the image–label pairs ($\tilde{x}$–$\tilde{y}$) in each of the SVHN, CIFAR-100, and Places365 datasets is detrimental to the primary task on CIFAR-10. By contrast, for ImageNet, the fact that blocking robust feature learning using y^{ER} leads to less performance improvement indicates that beneficial inductive transfer in terms of robust feature learning can be achieved from the auxiliary task on ImageNet. That is, ImageNet shares a large number of robust features as well as non-robust features with CIFAR-10.

2.4. A confidence-based selection strategy

Our analysis demonstrate that when using an auxiliary dataset for the primary task, the optimal algorithm to be applied varies according to the relationship between the two datasets in terms of robust features. In real world scenario, however, the auxiliary dataset may contain both favorable and unfavorable robust features for the primary task. Hence, we introduce a sample-wise selection strategy in our proposed method. The selection strategy is required to classify given auxiliary data into two groups based on the robust features, and the robust features exhibit human-perceptible patterns as shown in Fig. 1. Therefore, the objective of the selection strategy is consistent with that of existing OOD detection methods [18, 35]. Hendrycks et al. [18] proposed an OOD detection method using the confidence scores of the query data. To be specific, they trained models to give OOD samples a uniform posterior. Interestingly, the use of y^{ER} is naturally connected to their proposed method. That is, the use of y^{ER} achieves non-robust feature regularization without resulting in negative transfer while at the same time lowering the confidence scores of data irrelevant to the primary task in terms of robust features. On this basis, our selection strategy classify given auxiliary data samples based on their confidence scores. The proposed confidence-based method first (i) trains a classifier from scratch on the primary dataset; (ii) after a few epochs (warm-up), sets up a threshold using a hyperparamter $\pi \in \mathbb{R}^+$ and the mean confidence of the sampled primary data to sort out the auxiliary data samples that are likely to cause negative transfer; (iii) selects the lower-

Algorithm 1 Biased multi-domain adversarial training (BiaMAT) with the confidence-based selection strategy

Require: Primary dataset D, auxiliary dataset $\tilde{D}$, model parameter $\boldsymbol{\theta}$, training iterations K, warmup iterations K_w, learning rate τ, hyperparameters $\alpha \in \mathbb{R}^+$ and $\pi \in \mathbb{R}^+$

for $k = 1$ **to** K_w **do**
  /* Warm-up training on D */
  Sample a minibatch $B \sim D$
  $\mathcal{L} \leftarrow \mathbb{E}_{(\boldsymbol{x},y)\sim B}\left[\ell_{\text{adv}}(\boldsymbol{x}, y; h_{\text{pri}}, S)\right]$
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \tau \cdot \nabla_\theta \mathcal{L}$
end for
/* Confidence threshold */
$\omega \leftarrow \pi \cdot \mathbb{E}_{(\boldsymbol{x},y)\sim B}\left[\max h_{\text{pri}}(\boldsymbol{x})\right]$
for $k = K_w + 1$ **to** K **do**
  Sample a minibatch pair $B \sim D$ and $\tilde{B} \sim \tilde{D}$
  $\ell \leftarrow \mathbb{E}_{(\boldsymbol{x},y)\sim B}\left[\ell_{\text{adv}}(\boldsymbol{x}, y; h_{\text{pri}}, S)\right]$
  $\ell_{\text{high}}, \ell_{\text{low}} \leftarrow 0, 0$
  /* Confidence-based selection strategy */
  for $\tilde{\boldsymbol{x}}, \tilde{y}$ **in** $\tilde{B}$ **do**
    if $\max h_{\text{pri}}(\tilde{\boldsymbol{x}}) < \omega$ **then**
      $\ell_{\text{low}} \mathrel{+}= \frac{1}{|\tilde{B}|}\ell_{\text{adv}}(\tilde{\boldsymbol{x}}, y^{\text{ER}}; h_{\text{pri}}, S)$
    else
      $\ell_{\text{high}} \mathrel{+}= \frac{1}{|\tilde{B}|}\ell_{\text{adv}}(\tilde{\boldsymbol{x}}, \tilde{y}; h_{\text{aux}}, S)$
    end if
  end for
  $\mathcal{L} \leftarrow \ell + \alpha(\ell_{\text{low}} + \ell_{\text{high}})$
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \tau \cdot \nabla_\theta \mathcal{L}$
end for
Output: adversarially robust classifier $h_{\text{pri}} = f_{\text{pri}} \circ g$

than-threshold auxiliary data in each training batch based on their confidence scores for the primary classes; (iv) uses the low confidence data with y^{ER} and the remaining auxiliary data with $\tilde{y}$. The pseudo-code is provided in Algorithm 1.

3. Experimental results and discussion

3.1. Experimental setup

Datasets. We complement our analysis with experiments conducted on the CIFAR datasets and ImageNet. ImageNet is resized [8] to dimensions of 64×64 and then randomly divided into datasets that contain 100 and 900 classes, which are termed ImgNet100 and ImgNet900, respectively. SVHN [29], Places365, and ImageNet are used as auxiliary datasets. Auxiliary data that do not fit the input size of the classifier are resized to the primary data size. For instance, when CIFAR-10 is used as the primary dataset, Places365 is down-sampled to a dimension of 32×32, and ImageNet32x32 is leveraged.

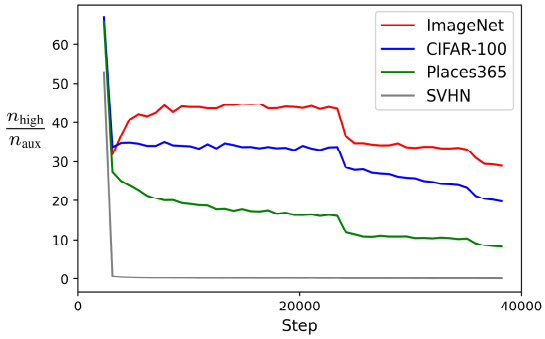

Figure 3. Ratio $\frac{n_{\text{high}}}{n_{\text{aux}}}$ for each auxiliary dataset with respect to the primary task on CIFAR-10.

Adversarial attack methods. Fast gradient sign method (FGSM) [15] is an one-step attack using the sign of the gradient. Madry et al. [25] proposed an iterative application of the FGSM method (PGD). Carlini & Wagner (CW) [3] attack is a targeted attack that maximize the logit of a target class and minimize that of ground-truth. Autoattack (AA) [10] is an ensemble attack that consists of two PGD extensions, one white-box attack [9], and one black-box attack [2]. We focus on the ℓ_∞-robustness, the most common robustness scenario considered in the field of heuristic defenses [25, 43, 23].

Implementation details. In our experiments, we adopt the adversarial training methods proposed by Madry et al. [25] and Zhang et al. [43] as the baseline methods, denoted by AT and TRADES, respectively. On the CIFAR datasets, we use WRN28-10 [42] and WRN34-10 for AT and TRADES, respectively. Although increasing the number of training epochs is expected to lead to higher adversarial robustness because of the use of additional data, owing to the high-computational complexity of adversarial training, we restrict the training of BiaMAT to 100 or 110 epochs with a batch size of 256 (128 primary and 128 auxiliary data samples, respectively). To evaluate adversarial robustness, we apply multiple attacks, including PGD, CW, and AA, with an ℓ_∞-bound with the same setting as that used in the training. PGD and CW with K iterations are denoted by PGD^K and CW^K, respectively, and the unperturbed test set is denoted by Clean. We consistently select the best checkpoint [41] to measure the adversarial robustness of the model on the test set. Further details regarding the model implementation, including an ablation study on choosing different values of π, are summarized and discussed in Appendix E.

3.2. Adversarial robustness under various attacks

Table 2 summarizes the improvements in the adversarial robustness of the models obtained from the application of BiaMAT. The proposed method can freely use various auxiliary datasets as it avoids negative transfer through the

Table 2. Performance improvements (accuracy %) on CIFAR-10, CIFAR-100, and ImgNet100 following application of the proposed method using various datasets. The best results within each baseline method (AT and TRADES) are indicated in bold

Primary dataset	Method	Auxiliary dataset	Clean	PGD^{100}	CW^{100}	AA
CIFAR-10	AT	-	87.37	50.87	50.93	48.53
	AT+BiaMAT (ours)	SVHN	87.34	51.90	51.40	48.61
		CIFAR-100	87.22	55.93	52.09	50.08
		Places365	87.76	57.00	51.70	49.48
		ImageNet	**88.75**	**57.63**	**53.04**	**50.78**
	TRADES	-	85.85	56.62	55.16	53.93
	TRADES+BiaMAT (ours)	SVHN	85.49	56.86	55.21	53.94
		CIFAR-100	87.02	58.69	56.85	55.48
		Places365	87.18	59.15	56.36	55.24
		ImageNet	**88.03**	**59.80**	**58.01**	**56.64**
CIFAR-100	AT	-	62.59	26.80	26.07	24.13
	AT+BiaMAT (ours)	Places365	63.44	32.61	28.53	26.49
		ImageNet	**64.05**	**33.74**	**29.78**	**27.65**
	TRADES	-	62.04	32.53	30.07	28.82
	TRADES+BiaMAT (ours)	Places365	64.58	34.38	30.72	29.24
		ImageNet	**65.82**	**36.36**	**33.42**	**31.87**
ImgNet100	AT	-	66.60	35.46	31.90	29.54
	AT+BiaMAT (ours)	Places365	**70.04**	**40.52**	33.24	30.64
		ImgNet900	68.00	40.18	**35.00**	**32.88**
	TRADES	-	56.16	27.90	22.98	21.90
	TRADES+BiaMAT (ours)	Places365	57.80	29.30	24.14	23.06
		ImageNet	**58.76**	**31.26**	**25.98**	**24.98**

application of the confidence-based selection strategy; in fact, a comparison of Tab. 1 and Tab. 2 demonstrates that the proposed method effectively overcomes negative transfer and achieves only beneficial training signals for the primary task from the auxiliary task. To observe how the confidence-based selection strategy works while the model is being trained through the application of the proposed method, we define a ratio $\frac{n_{\text{high}}}{n_{\text{aux}}}$, where n_{aux} and n_{high} denote the amount of auxiliary data and the higher-than-threshold auxiliary data within each training batch, respectively. That is, the ratio represents the percentage of data used for robust feature learning as well as non-robust feature regularization for an auxiliary dataset. Figure 3 shows the plot of the ratio $\frac{n_{\text{high}}}{n_{\text{aux}}}$ at $\pi = 0.55$ (defined in Algorithm 1) during the training of the AT+BiaMAT models using various auxiliary datasets on CIFAR-10. As shown, the confidence-based selection strategy successfully filters out data that are likely to induce negative transfer for the primary task. In other words, a relatively high percentage of ImageNet data are used for robust feature learning, and each of the SVHN, CIFAR100, and Places365 datasets are mostly used with y^{ER}, which is consistent with the results listed in Tab. 1. Additional analysis is in Appendix F.

We conduct several experiments to further investigate the proposed method. (Appendix B) To observe the effects of the use of more auxiliary datasets, we train a BiaMAT model using a combination of two auxiliary datasets; the results show that the use of more auxiliary datasets does not always lead to further improvements in adversarial robustness. In other words, the relationship (in terms of robust and non-robust features) between the primary and auxiliary datasets is more important to BiaMAT than the number of auxiliary datasets. (Appendix C) To further demonstrate that robust feature learning can be achieved from the auxiliary task in BiaMAT, we construct robust datasets [21] from the AT and AT+BiaMAT models and normally train models from scratch on each robust dataset (D^{AT} and D^{BiaMAT}); the results show that D^{BiaMAT} results in more robust models than those trained on D^{AT}, implying that BiaMAT enables DNNs to learn better robust features via inductive transfer between adversarial training on the primary and auxiliary datasets.

Table 3. Comparison (accuracy %) of the effectiveness of BiaMAT with the semi-supervised [4] and pre-training [17] methods on CIFAR-10.

Method	Auxiliary dataset	Clean	PGD^{100}	CW^{100}	AA
TRADES (baseline)	-	85.85	56.62	55.16	53.93
Hendrycks *et al*. [17]	CIFAR-100	80.21	45.68	44.52	42.36
	ImageNet	87.11	57.16	55.43	55.30
Carmon *et al*. [4]	CIFAR-100	82.61	54.32	51.64	50.81
	Places365	83.95	56.72	53.95	52.81
	ImageNet	85.42	57.46	54.66	53.79
	ImageNet-500k	86.02	59.49	56.43	55.63
TRADES+BiaMAT (ours)	CIFAR-100	**87.02**	**58.69**	**56.85**	**55.48**
	Places365	**87.18**	**59.15**	**56.36**	**55.24**
	ImageNet	**88.03**	**59.80**	**58.01**	**56.64**

3.3. Comparison with other related methods

Carmon et al. [4] proposed a semi-supervised learning technique where the training dataset is augmented with unlabeled in-distribution data; the main difference between this and BiaMAT is the distribution of additional data. For instance, Carmon et al. collected the in-distribution data of CIFAR-10 from 80 Million Tinyimages dataset [39] and used the unlabeled data with pseudo-labels. Therefore, no assumptions are required regarding the classes of the primary and auxiliary datasets in our scenario, but the semi-supervised method is ineffective when the primary and auxiliary datasets do not share the same class distributions. To demonstrate this, we assign pseudo-labels to the auxiliary data using a pre-trained classifier and configure each training batch (for TRADES) such that it contains the same amount of primary and pseudo-labeled data, as in [4]. In particular, we sort ImageNet based on the confidence in the CIFAR-10 classes and select the top 50k (or top 5k) samples for each class in CIFAR-10 (or CIFAR-100); this is denoted as ImageNet-500k. As shown in Tab. 3, the Carmon et al. [4] method exhibits lower effectiveness than the proposed method. Specifically, the results obtained using CIFAR-100 and Places365 demonstrate that the semi-supervised method is vulnerable to negative transfer because of the considerable domain discrepancy between the primary and auxiliary datasets.

Hendrycks et al. [17] demonstrated that ImageNet pre-training can improve adversarial robustness on the CIFAR datasets. However, the pre-training method is effective only when a dataset that has a distribution similar to that of the primary data and a sufficiently large number of samples is used. To demonstrate this, we adversarially pre-train the CIFAR-100 and ImageNet [17] models and then adversarially fine-tune them on CIFAR-10. The results in Tab. 3 demonstrate that the pre-training method is ineffective when leveraging datasets that do not satisfy the conditions mentioned above. In other words, because the effect achieved by the pre-training method arises from the reuse of features pre-trained on a dataset that contains a large quantity of data with a distribution similar to that of the primary dataset, CIFAR-100 is not suitable for application of the CIFAR-10 task. Conversely, BiaMAT avoids such negative transfer through the application of a confidence-based strategy. That is, these results emphasize the high compatibility of the proposed method with a variety of datasets. Additional experimental results, including comparisons with OAT [24] or a generative model-based method [16], can be found in Appendix D.

4. Conclusions and future directions

In this study, we develop BiaMAT, a method that uses publicly available (labeled) auxiliary datasets to reduce the large gap between training and test errors in adversarial training. Our theoretical and empirical analysis demonstrate that the effectiveness of BiaMAT can be attributed to two factors: non-robust feature regularization and robust feature learning. In particular, we show that while existing methods are vulnerable to negative transfer due to the distributional discrepancy between auxiliary and primary data, BiaMAT can successfully overcome negative transfer through the application of a confidence-based selection strategy. In this study, however, the application of any method that can improve the performance of multi-domain learning is not considered. In addition, there is room for improvement in the effectiveness of BiaMAT with regard to the strategy used to avoid negative transfer. In future work, therefore, we will develop algorithms in which additional techniques, such as the use of adaptive weighting strategies [26], are implemented.

Acknowledgements: This work was supported by the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2022.

References

[1] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.

[2] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. *arXiv preprint arXiv:1912.00049*, 2019.

[3] 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.

[4] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 11190–11201, 2019.

[5] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[6] Alvin Chan, Yi Tay, and Yew-Soon Ong. What it thinks is important is important: Robustness transfers through input gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[8] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

[9] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. *arXiv preprint arXiv:1907.02044*, 2019.

[10] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *arXiv preprint arXiv:2003.01690*, 2020.

[11] Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[13] Mark Dredze, Alex Kulesza, and Koby Crammer. Multi-domain learning by confidence-weighted parameter combination. *Machine Learning*, 79(1-2):123–149, 2010.

[14] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019.

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[16] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34, 2021.

[17] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019.

[18] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[19] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

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

[21] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[23] Saehyung Lee, Hyungyu Lee, and Sungroh Yoon. Adversarial vertex mixup: Toward better adversarially robust generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[24] Saehyung Lee, Changhwa Park, Hyungyu Lee, Jihun Yi, Jonghyun Lee, and Sungroh Yoon. Removing undesirable feature contributions using out-of-distribution data. In *International Conference on Learning Representations*, 2021.

[25] 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.

[26] YUREN MAO, Weiwei Liu, and Xuemin Lin. Adaptive adversarial multi-task representation learning. In *Proceedings of Machine Learning and Systems 2020*, pages 1832–1841. 2020.

[27] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4293–4302, 2016.

[28] Muhammad Muzammal Naseer, Salman H Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-domain transferability of adversarial perturbations. In *Advances in Neural Information Processing Systems*, pages 12885–12895, 2019.

[29] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[30] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 512–523. Curran Associates, Inc., 2020.

[31] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. In *International Conference on Learning Representations*, 2021.

[32] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised

learning. In *International Conference on Learning Representations*, 2021.

[33] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[34] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pages 5014–5026, 2018.

[35] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021.

[36] Ali Shafahi, Parsa Saadatpanah, Chen Zhu, Amin Ghiasi, Christoph Studer, David Jacobs, and Tom Goldstein. Adversarially robust transfer learning. In *International Conference on Learning Representations*, 2020.

[37] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020.

[38] Robert Stanforth, Alhussein Fawzi, Pushmeet Kohli, et al. Are labels required for improving adversarial robustness? *arXiv preprint arXiv:1905.13725*, 2019.

[39] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

[40] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[41] Eric Wong, Leslie Rice, and Zico Kolter. Overfitting in adversarially robust deep learning. In *Proceedings of Machine Learning and Systems 2020*, pages 5304–5315. 2020.

[42] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[43] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482, Long Beach, California, USA, 09–15 Jun 2019. PMLR. `https://github.com/yaodongyu/TRADES`.

[44] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.