# Mitigating Adversarial Vulnerability through Causal Parameter Estimation by Adversarial Double Machine Learning

Byung-Kwan Lee*, Junho Kim*, Yong Man Ro†
Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea
{leebk, arkimjh, ymro}@kaist.ac.kr

## Abstract

*Adversarial examples derived from deliberately crafted perturbations on visual inputs can easily harm decision process of deep neural networks. To prevent potential threats, various adversarial training-based defense methods have grown rapidly and become a de facto standard approach for robustness. Despite recent competitive achievements, we observe that adversarial vulnerability varies across targets and certain vulnerabilities remain prevalent. Intriguingly, such peculiar phenomenon cannot be relieved even with deeper architectures and advanced defense methods. To address this issue, in this paper, we introduce a causal approach called Adversarial Double Machine Learning (ADML), which allows us to quantify the degree of adversarial vulnerability for network predictions and capture the effect of treatments on outcome of interests. ADML can directly estimate causal parameter of adversarial perturbations per se and mitigate negative effects that can potentially damage robustness, bridging a causal perspective into the adversarial vulnerability. Through extensive experiments on various CNN and Transformer architectures, we corroborate that ADML improves adversarial robustness with large margins and relieve the empirical observation.*

## 1. Introduction

Along with the progressive developments of deep neural networks (DNNs) [17, 10, 2], an aspect of AI safety comes into a prominence in various computer vision research [45, 64, 12, 19]. Especially, adversarial examples [50, 15, 30] are known as potential threats on AI systems. With deliberately crafted perturbations on the visual inputs, adversarial examples are hardly distinguishable to human observers, but they easily result in misleading decision process of DNNs. Such adversarial vulnerability provokes weak reliability of inference process of DNNs and discourages AI adoption to the safety critical areas [54, 44].

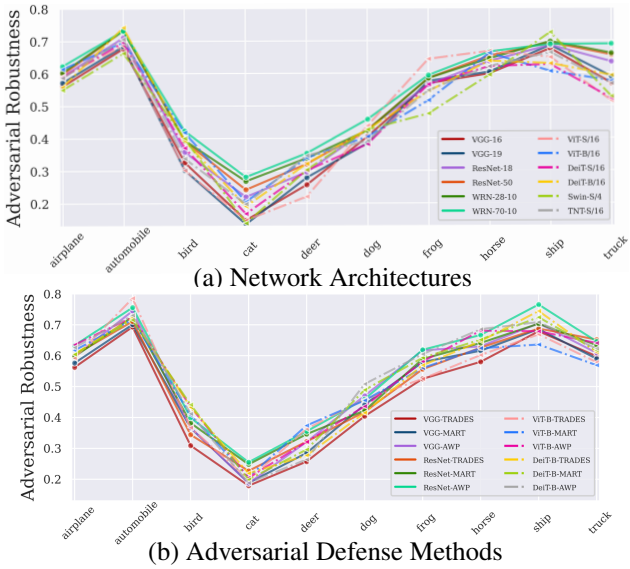
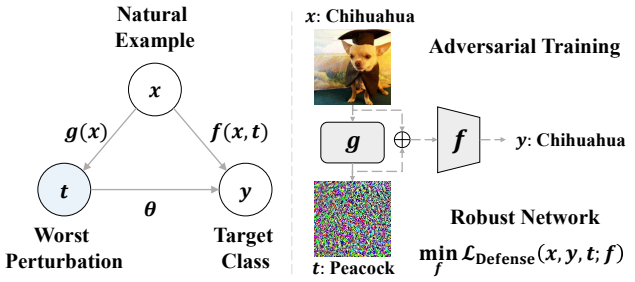



Figure 1. The comparison of adversarial robustness along target classes with respect to (a) Network Architectures and (b) Adversarial Defense Methods on CIFAR-10 [27]. Note that, the distribution of adversarial robustness is consistent along both criteria.

In order to achieve robust and trustworthy DNNs from adversarial perturbation, previous methods [33, 28, 3, 62, 55, 58, 8] have delved into developing various adversarial attack and defense algorithms in the sense of cat-and-mouse game. As a seminal work, Madry *et al.* [33] have paved the way for obtaining robust network through adversarial training (AT) regarded as an ultimate augmentation training [52] with respect to adversarial examples. Based on its effectiveness, various subsequent works [62, 55, 58, 63, 39, 25] have investigated it to further enhance adversarial robustness.

Although several AT-based defense methods have become a de facto standard due to their competitive adversarial robustness, we found an intriguing property of the current defense methods. As in Figure 1, we identify that the adversarial robustness for the each target class significantly varies with a large gap, and this phenomenon equally happens in the course of (a) network architectures and (b)

*Equal contribution. † Corresponding author.




(a) AT Causal Diagram (b) Training Robust Network

Figure 2. Overview of canonical adversarial training procedure for robust network and its causal diagram.

various AT-based defense methods. In addition, we would like to point out that the robustness of particular target is still severely vulnerable than others even with advanced architectures [10, 51, 31] and defense methods [62, 55, 58]. We argue that such phenomenon is derived from the current learning strategies of AT-based defense methods that lacks of understanding causal relations between the visual inputs and predictions. When considering AT methods as the ultimate augmentation [52], current methods rely solely on strengthening the correlation between adversarial examples and target classes through canonical objectives that improve robustness. To fundamentally address such vulnerability and understand the causal relation, we need to quantify the degree of vulnerability (*i.e.,* causal parameter) and should mitigate its direct effects to the network predictions.

Accordingly, we investigate the AT-based defense methods in a causal viewpoint and propose a way of precisely estimating causal parameter between adversarial examples and their predictions, namely *Adversarial Double Machine Learning (ADML)*. We first represent a causal diagram of AT-based methods and interpret it as a generating process of robust classifiers as illustrated in Figure 2. Regarding standard adversarial training [33] as an optimizing procedure for the robust network parameters $f$ with respect to the worst perturbations $t$, we can instantiate a generation $g$[1] as an adversarial attack of projected gradient descent (PGD) [33] for the given clean examples $x$.

Then, our research question is how to quantitatively compute the causal parameter $\theta$ between the perturbations $t$ and target classes $y$, and identify the causal effects on outcome of our interests. Through double machine learning (DML) [5], widely studied as a powerful causal estimator [4, 7, 13, 20, 21] for the given two regression models, we can establish an initial research point of estimating causal parameter of adversarial perturbation with theoretical background. However, it is difficult to directly estimate $\theta$ in the high-dimensional manifolds, especially for DNNs. In this paper, we shed some lights on identifying causal parameter of the perturbations while theoretically bridging the gap between causal inference and adversarial robustness. Then, by minimizing the magnitude of the estimated causal parameter, we essentially lessen negative causal effects of adversarial vulnerability, and consequently acquire robust network with the aforementioned phenomenon alleviated.

To corroborate the effectiveness of ADML on adversarial robustness, we set extensive experiments with four publicly available datasets [27, 29, 9]. Our experiments include various convolutional neural network architectures (CNNs), as well as Transformer architectures that have drawn great attention in both vision and language tasks [53, 10, 66, 65] yet relatively lack of being studied in adversarial research.

Our contributions can be summarized as follows:

- We present an empirical evidence that despite the recent advances in AT-based defenses, fundamentally adversarial vulnerability still remains across various architectures and defense algorithms.
- Bridging a causal perspective into adversary, we propose Adversarial Double Machine Learning (ADML), estimating causal parameter in adversarial examples and mitigating its causal effects damaging robustness.
- Through extensive experiments and analyses on various CNN and Transformer architectures, we corroborate intensive robustness of our proposed method with the phenomenon alleviated.

## 2. Background and Related Work

**Notation.** We deal with DNNs for classification as in Figure 2, represented by $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ denotes image and probability space, respectively. Let $x \in \mathcal{X}$ denote clean images and $y \in \mathcal{Y}$ indicate (one-hot) target classes corresponding to the images. Adversarial examples $x_{\text{adv}}$ are generated by adversarial perturbations $t$ through DNNs, such that $x_{\text{adv}} = x + t$. Here, the perturbations are carefully created through the following formulation:

$$\max_{\|t\|_\infty \leq \gamma} \mathcal{L}_{\text{CE}}(f(x+t), y), \tag{1}$$

where $\mathcal{L}_{\text{CE}}$ represents a pre-defined loss such as cross-entropy for classification task. We regard adversarial perturbations $t$ as $l_\infty$ perturbation within $\gamma$-ball (*i.e.,* perturbation budget). Here, $\|\cdot\|_\infty$ describes $l_\infty$ perturbation magnitude.

### 2.1. Adversarial Training

After several works [50, 15, 24] have found that human-imperceptible adversarial examples easily break network predictions, Madry *et al.* [33] have thrown a fundamental question: "How can we make models robust to adversarial examples with security guarantee?". To answer it, they have used the concept of empirical risk minimization (ERM) serving as a recipe to obtain classifiers with small population risk. Thanks to its reliable guarantee, they have consolidated it on the purpose of adversarial defense and

[1] Selecting $g$ as proper perturbations varies according to domain specific tasks (*e.g.,* rotations, translations [11], or spatial deformations [59]).

accomplished the yardstick of adversarial training. The key factor of its achievement is regarding adversarial training as min-max optimization in a perspective of saddle point problem, which can be written as follows:

$$\min_{f} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[ \max_{\|t\|_\infty \leq \gamma} \mathcal{L}_{\text{CE}} \left( f(x+t), y \right) \right], \qquad (2)$$

where $\mathcal{D}$ denotes a set of data samples $(x, y)$. Here, they have presented an adversarial attack based on PGD to powerfully behave inner-maximization on Eq. (2), which is an ultimate first-order adversary with a multi-step variant of fast gradient sign method [28] by adding a random perturbation around the clean images $x$.

According to its impact, various adversarial training methods [33, 62, 55, 58] have grown exponentially and become de facto standards robustifying DNNs against adversarial perturbation. Zhang *et al.* [62] have pointed out the trade-off between clean accuracy and adversarial robustness, and reduced the gap between clean errors and robust errors. Wang *et al.* [55] have claimed that all of clean images are used to perform both inner-maximization and outer-minimization process in Eq. (2), irrespective of whether they are correctly classified or not. Thus, they have focused on misclassified clean images prone to be easily overlooked during adversarial training and demonstrated their significant impacts on the robustness by incorporating an explicit regularizer for them. Wu *et al.* [58] have studied loss landscapes with respect to network parameters and shown a positive correlation between the flatness of the parameter loss landscapes and the robustness. In the end, they have presented a double-perturbation mechanism where clean images are perturbed, while network parameters are simultaneously perturbed as well.

On the other hand, we plunge into investigating where adversarial vulnerability comes from and observe that the vulnerability varies along target classes, and it significantly deteriorates network predictions. Further, we find that this phenomenon commonly happens across various network architectures and advanced defense methods. To relieve such peculiarity, we deploy double machine learning (DML) that helps to capture how treatments (*i.e.,* adversarial perturbations) affect outcomes of our interests (*i.e.,* network predictions), which is one of the powerful causal inference methods. After we concisely explicate the necessary background of DML, we will bridge it to the adversary in Sec. (3).

## 2.2. Double Machine Learning

In data science and econometrics, one of the fundamental problems is how to measure causality between treatments $t$ and outcomes of our interest $y$ among high-dimensional observational data samples (see Figure 2) to identify data generating process. At a first glance, it seems simple to compute their causality, but we should keep in mind the possibility for the existence of covariates $x$ affecting both treatments and outcomes. In other words, for example, if one may want to know the causal effects of drug dosage $t$ to blood pressure changes $y$, one needs to collect observational data with respect to a variety of patients characteristics and their clinical histories $x$, so as not to fall into biased environment. In reality, though, it is impossible to collect observational data including all covariates concerning treatments and outcomes, so it is not an easy problem to catch genuine causality under the unknown covariates $x$. Therefore, there has been a growing demand for robustly predicting the unbiased causal relation, despite with the limited data samples.

Recently, the advent of double machine learning (DML) [5] enables us to clarify the causality between treatments $t$ and outcomes $y$, when two regression models are given. The formulation of initial DML can be written as:

$$\begin{aligned} y &= f(x) + \theta t + u, \quad && (\mathbb{E}[u \mid x, t] = 0) \\ t &= g(x) + v, && (\mathbb{E}[v \mid x] = 0) \end{aligned} \qquad (3)$$

where $\theta \in \mathbb{R}$ denotes causal parameter representing causal relation between $t \in \mathbb{R}^d$ and $y \in \mathbb{R}^d$. In addition, $f$ indicates one regression model projecting covariates to outcome domain, and $g$ denotes another regression model generating treatments $t$. In the sense that two regression models $f$ and $g$ are not main interest of DML, they are called as nuisance parameters to estimate the causal parameter $\theta$. Note that, early DML assumes the problem setup is proceeded in partially linear settings as a shape of Robinson-style [42] described in Eq. (3), where "partially" literally means that treatments $t \in \mathbb{R}^d$ are linearly connected to outcome $y \in \mathbb{R}^d$, while covariates $x$ are not. In addition, it is supposed that the conditional expected error of $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$ equals to zero vector $0 \in \mathbb{R}^d$.

To obtain the causal parameter $\theta$, Chernozhukov *et al.* [5] have provided a solution of estimating the causal parameter such that $\hat{\theta} = (y - \mathbb{E}[y \mid x]) \cdot v / \|v\|^2$ which satisfies Neyman-orthogonality [35, 34]. It makes $\hat{\theta}$ invariant to their erroneous of two nuisance parameters with the variance of causal parameter reduced. Furthermore, they have addressed a chronic problem that $\theta$ is only accessible when the two nuisance parameters are in a class of Donsker condition, where deep neural networks are not included in that condition. They have theoretically demonstrated *sample-splitting* plus *cross-fitting* can effectively relax Donsker condition and allow a broad array of modern ML methods [5] to compute unbiased causal parameter $\theta$.

Following the principle, they first split the data samples $\{\mathcal{D}_1, \mathcal{D}_2\} \sim \mathcal{D}$ and divided the process of causal inference into two steps: (a) training two nuisance parameters $f$ and $g$ with $\mathcal{D}_1$, (b) estimating unbiased $\theta$ with $\mathcal{D}_2$. Here, data samples $\mathcal{D}_2$ used to estimate unbiased causal parameters should not be overlapped with $\mathcal{D}_1$ utilized to train the nuisance parameters. To make copious combinations, they

swapped the role of partitioned data samples $\mathcal{D}_1 \leftrightharpoons \mathcal{D}_2$ or repeatedly split $\mathcal{D}$. Subsequently, they have performed cross-fitting (*e.g.,* $k$-fold cross validation) by averaging the estimated causal parameters from various split samples.

Along with the success of initial DML in partially linear settings, numerous variants [4, 7, 32, 14, 13, 26, 6] have emerged, and they have extended its initial nature to non-parametric settings with continuous treatments $t$ in order to capture more complicated non-linear causal relations in a debiased state. A non-parametric formulation [4] represents a more general problem setup of DML as follows:

$$\begin{aligned} y &= f(x,t) + u, \quad (\mathbb{E}[u \mid x,t] = 0) \\ t &= g(x) + v, \quad\quad (\mathbb{E}[v \mid x] = 0) \end{aligned} \tag{4}$$

where there is no explicit term for causal parameter $\theta$ exhibiting causal relation between treatments $t$ and outcomes $y$, compared to Eq. (3). Colangelo *et al.* [7] have introduced a way of estimating causal parameter $\theta$ applicable to non-parametric settings with high-dimensional continuous treatments $t \in \mathcal{T}$, which can be written as:

$$\hat{\theta} = \frac{\partial}{\partial t}\mathbb{E}[y \mid \text{do}(\mathcal{T}{=}t)]. \tag{5}$$

They have utilized do-operator [37] commonly used in graphical causal models and intervened on treatments $t$ to compute an interventional expectation $\mathbb{E}[y|\text{do}(\mathcal{T}{=}t)]$. It represents the expected outcome averaged from all the possible covariates for the given fixed treatments $t$, such that $\mathbb{E}[y|\text{do}(\mathcal{T}{=}t)] = \sum_{x\in\mathcal{X}} \mathbb{E}[y|x,t]p(x)$. Specifically, they have estimated causal parameter $\theta$ by measuring how much the interventional expectation shifted, once they change the treatments slightly. Since the most important property of DML is Neyman-Orthogonality helping to robustly estimate the causal parameter, the interventional expectation should be also modified to satisfy the property [7, 23] of its invariance to nuisance parameters $f$ and $g$. Its formulation can be written as follows (see details in Appendix A):

$$\mathbb{E}[y \mid \text{do}(\mathcal{T}{=}t)] = \mathbb{E}_{\mathcal{D}_t}\left[f(x,t) + \frac{y - f(x,t)}{p(\mathcal{T}{=}t \mid x)}\right], \tag{6}$$

where $\mathcal{D}_t$ denotes a set of observational covariates and outcome samples for a fixed $t \in \mathcal{T}$ such that $(x,y) \sim \mathcal{D}_t$, a sub-population of $\mathcal{D}$. Note that, $p(\mathcal{T}{=}t|x)$ is related to treatment generator $g$. Here, differentiating Eq. (6) enables us to acquire unbiased causal parameter in non-parametric settings with non-linear causal relation.

In brief, DML captures unbiased causal relation between treatments $t$ and outcomes $y$ even with finite data samples, of which theoretical ground is (a) Neyman-Orthogonality for robustly estimated causal parameter despite undesirable outputs of nuisance parameters, and (b) sample-splitting plus cross-fitting for debiased causal parameters.

# 3. Adversarial Double Machine Learning

## 3.1. Adversarial Data Generating Process

In general deep learning schemes, we have clean visual images $x \in \mathbb{R}^{hwc}$ and their corresponding target classes $y \in \mathbb{R}^d$ in our hand as a format of dataset, where $h$, $w$, $c$ denotes image resolution of height, width, channel, repectively, and $d$ denotes the number of classes. Thus, we do not need additional data generating process. For adversarial training, on the other hand, we need another data, which are adversarial perturbations generated from data samples $(x,y)$ as in Eq. (1). They are normally created by PGD [33] at every training iteration to make DNNs $f$ robust through min-max optimization game according to Eq. (2).

Though, the more iterations of adversarial training, the fewer perturbations that impair network predictions. In other words, not all of the perturbations can corrupt network predictions among newly generated perturbations. Hence, we do not consider all of the perturbations as treatments but selectively define them as worst perturbations $t$ breaking network predictions, such that it satisfies $y \neq f(x+t)$, where we call $x_{\text{adv}} = x + t$ as worst examples. This is because our major goal is to catch actual adversarial vulnerability of DNNs, so that we do not tackle the perturbations incapable of harming network predictions.

To access such worst perturbation, we choose perturbation generator $g$ as an adversarial attack of PGD according to standard adversarial training [33]. In addition, we pick the worst perturbations $t$ damaging network predictions among adversarial perturbations from the generator $g$. In this way, we perform adversarial data generating process.

## 3.2. Adversarial Problem Setup

In the nature of adversarial training, the worst perturbations $t$ are explicitly injected to clean images $x$ such that $x_{\text{adv}} = x + t$, and these combined images are propagated into DNNs $f$. Here, through this formulation as: $f(x{+}t) = f(x,t)$, we connect DNNs for adversarial training and a nuisance parameter $f$ for non-parametric DML in Eq. (4). Fortunately, once we use Taylor expansion (with scalar-valued function for better understanding) and decompose $f$ by its input component as: $f(x{+}t) = f(x) + \sum_{i=1}^{\infty} t^i f^{(i)}(x)/i!$, where $f^{(i)}$ indicates $i$-th order derivative function, we can also express partially linear settings described in Eq. (3). That is, since adversarial examples start from the concept of "additive noise", both settings can exist at the same time in the scheme of adversarial training. From this reason, we build *Adversarial Double Machine Learning (ADML)*:

$$\begin{aligned} y &= f(x+t) = f(x) + \theta\bar{t} + u, \quad (\mathbb{E}[u \mid x,t] = 0) \\ t &= g(x) + v, \quad\quad\quad\quad\quad\quad (\mathbb{E}[v \mid x] = 0) \end{aligned} \tag{7}$$

where $\bar{t}$ indicates Taylor-order matrix: $[t, t^2, \cdots]^T$ and $\theta$ represents Taylor-coefficient matrix $[\frac{f^{(1)}(x)}{1!}, \frac{f^{(2)}(x)}{2!}, \cdots]$ (see

strict mathematical verification in Appendix B).

Here, we explain what the conditional expected error of $u$ and $v$ in Eq. (7) means in adversarial training. The former $\mathbb{E}[u|x,t]=0$ implies the nature of adversarial training, which can be viewed as an ultimate augmentation robustifying DNNs, when infinite data population of $x$ and $t$ is given. Thus, it means network predictions become invariant in the end, despite the given worst perturbations. To implement it practically, we replace it with a mild assumption as $\mathbb{E}[u|x,g(x)]=0$ (see Appendix C) that signifies PGD-based perturbations are used to perform adversarial training. This is because we cannot always acquire worst perturbations $t$ using only PGD. For the latter $\mathbb{E}[v|x]=0$, it represents PGD has capability of producing worst perturbations deviating network predictions during the training.

### 3.3. Estimating Adversarially Causal Parameter

Aligned with Eq. (3), an explicit term of $\theta$ is regarded as the causal parameter in our problem setup of ADML. We can now interpret that $\theta$ is a causal factor to spur adversarial vulnerability, since its magnitude easily catalyzes the deviations from network predictions of clean images. Here, if it is possible to directly compute $\theta$ over all data samples, we can finally handle adversarial vulnerability.

Favorably, ADML follows both partially linear and nonparametric settings due to the concept of additive noise, thus we can employ the way of estimating causal parameter as in Eq. (5). The following formulation represents the estimated causal parameter $\hat{\theta}$ in ADML (see Appendix D). Note that, as we emphasized, *sample-splitting* plus *cross-fitting* must be applied to estimate unbiased causal parameter.

$$\hat{\theta}=\mathbb{E}_{\mathcal{D}_t}\left[-\left(\frac{1}{p(\mathcal{T}{=}t\mid x)}-1\right)\frac{\partial}{\partial t}f(x+t)\right], \qquad (8)$$

where $\frac{\partial}{\partial t}f(x+t)$ indicates an input gradient for network predictions with respect to $t$, and $p(\mathcal{T}{=}t|x)$ represents a distribution of worst perturbation given clean images $x$. Here, we cannot directly handle this distribution due to the presence of multiple unknown parameters required to define it. For that reason, we instead approximate it with the sharpening technique by incorporating the information on attacked confidence such that $p(\mathcal{T}{=}t|x) \approx \mathbb{E}_{t'|x}[p(y_a|x,t')]$ (see Appendix E), where $y_a$ denotes attacked classes for the given worst perturbations $t$. It implicitly means that the higher the attacked confidence, the higher the probability of finding worst perturbations.

Aligned with the previous analysis [47] that show increasing magnitude of input gradient increases adversarial vulnerability, the magnitude of our causal parameter $|\hat{\theta}|$ also becomes huge due to $|\hat{\theta}| \propto |\frac{\partial}{\partial t}f(x+t)|$. In parallel, Qin *et al.* [38] show the more ambiguous confident, the lower robustness (high vulnerability), and interestingly, the magnitude of our causal parameter also $|\hat{\theta}|$ becomes large due to $|\hat{\theta}| \propto |1/\mathbb{E}_{t'|x}[p(y_a|x,t')]|$.

**Algorithm 1** ADML

**Require:** Data Samples $\mathcal{D}$, Network $f$

**for** $(x,y) \sim \mathcal{D}$ **do** ▷ Cross-Fitting
  $t' \leftarrow g(x)$ ▷ PGD Attack
  $(x_1,y_1,t'_1),(x_2,y_2,t'_2) \sim \text{Split}(x,y,t')$
  $\mathcal{L}_a \leftarrow \mathcal{L}_{\text{Defense}}(x_1,y_1,t'_1;f)$ ▷ Mild Assumption
  $(x_{t_2},y_{t_2},t_2) \leftarrow \text{Select}(y_2 \neq f(x_2+t'_2))$ ▷ Worst
  $j^* \leftarrow \arg\max_j f_j(x_{t_2}+t_2)$ ▷ $j$ : Class Index
  $\tau \leftarrow \frac{1}{f_{j^*}(x_{t_2}+t_2)}-1$ ▷ Balancing Ratio
  $\mathcal{L}_b \leftarrow \tau\mathcal{L}_{\text{CE}}(f(x_{t_2}+t_2),y_{t_2})+\mathcal{L}_{\text{CE}}(f(x_{t_2}),y_{t_2})$
  $\mathcal{L}_{\text{ADML}} \leftarrow \mathcal{L}_a+\mathcal{L}_b$ ▷ ADML Loss
  $w_f \leftarrow w_f-\alpha\frac{\partial}{\partial w_f}\mathcal{L}_{\text{ADML}}$ ▷ Weight Update ($\alpha$: lr)
**end for**

Bringing such factors at once, $\hat{\theta}$ represents a weighted measurement of attacked confidence and their input gradients. Comprehensively, we can revisit that the network predictions of worst examples are easily flipped due to the following adversarial vulnerability: (a) ambiguous confidence around classification boundaries, or (b) high gradient magnitude amplifying the leverage of the perturbations. To improve the adversarial robustness of DNNs, it is essential to minimize the negative effects of causal parameters, which are combinatorial outcomes of the gradient and confidence.

### 3.4. Mitigating Adversarial Vulnerability

By deploying ADML, we propose a way of estimating causal parameter representing the degree of adversarial vulnerability that disturbs to predict the target classes. Then, our final goal is essentially to lessen its direct causal effect from adversarial perturbations in order to achieve robust networks. In detailed, alleviating their causal effect derived from $\hat{\theta}$ is the process of comprehensive reconstruction to focus more on vulnerable samples as we reflect their attacked confidence and gradients effects altogether.

Accordingly, the very first way is naively reducing the magnitude of $\hat{\theta}$ to suppress adversarial vulnerability damaging the robustness. However, calculating $\hat{\theta}$ and minimizing its magnitude at every iteration is computationally striking because input gradient has huge dimension of $\mathbb{R}^{dhwc}$ and getting its gradient inevitably needs to compute second-order gradient with its tremendous dimension. We instead approximate the partial derivative $\frac{\partial}{\partial t}\mathbb{E}[y|\text{do}(\mathcal{T}{=}t)]$ and minimize its magnitude, which can be written as:

$$\min_f |\hat{\theta}| \approx \left|\frac{\mathbb{E}[y \mid \text{do}(\mathcal{T}{=}t)]-\mathbb{E}[y \mid \text{do}(\mathcal{T}{=}0)]}{t-0}\right|, \qquad (9)$$

where network parameters of DNNs $f$ are only dependent on the numerator, thus we engross the numerator only. Lastly, we redesign $\mathbb{E}[y|\text{do}(\mathcal{T}{=}t)]$ into the form of loss

| | Method | CIFAR-10 | | | | | | | CIFAR-100 | | | | | | | Tiny-ImageNet | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Clean | BIM | PGD | $CW_\infty$ | AP | DLR | AA | Clean | BIM | PGD | $CW_\infty$ | AP | DLR | AA | Clean | BIM | PGD | $CW_\infty$ | AP | DLR | AA |
| VGG-16 | AT | 78.8 | 49.4 | 48.1 | 46.8 | 46.4 | 46.4 | 46.3 | **53.9** | 26.0 | 25.0 | 24.1 | 23.7 | 23.8 | 23.7 | **56.5** | 26.6 | 25.4 | 25.5 | 24.7 | 24.6 | 24.6 |
| | +ADML | **80.9** | **61.8** | **61.7** | **59.8** | **55.0** | **54.8** | **54.5** | 52.2 | **31.0** | **30.8** | **29.9** | **27.8** | **27.6** | **27.3** | 55.4 | **35.5** | **34.9** | **32.9** | **32.7** | **32.4** | **32.2** |
| | TRADES | 79.5 | 48.6 | 47.6 | 45.7 | 46.4 | 46.3 | 46.3 | **53.3** | 25.5 | 24.8 | 23.5 | 23.6 | 23.7 | 23.2 | **56.1** | 28.3 | 27.3 | 26.2 | 25.8 | 25.8 | 25.7 |
| | +ADML | **81.0** | **62.6** | **62.4** | **59.0** | **55.3** | **55.1** | **54.9** | 52.2 | **31.2** | **31.1** | **29.8** | **27.6** | **27.3** | **27.3** | 55.4 | **36.6** | **36.0** | **34.0** | **33.6** | **33.4** | **33.3** |
| | MART | 78.3 | 51.9 | 50.6 | 48.8 | 48.9 | 48.8 | 48.7 | **52.6** | 26.6 | 26.0 | 24.4 | 24.3 | 24.3 | 24.2 | **55.7** | 27.8 | 26.6 | 26.0 | 25.7 | 25.7 | 25.6 |
| | +ADML | **80.4** | **62.4** | **62.2** | **60.3** | **55.7** | **55.3** | **55.2** | 51.6 | **31.6** | **31.0** | **29.6** | **27.8** | **27.5** | **27.3** | 55.1 | **36.2** | **35.8** | **34.9** | **34.3** | **33.9** | **33.7** |
| | AWP | 77.2 | 53.9 | 52.6 | 50.1 | 51.4 | 51.1 | 51.0 | 52.1 | 30.2 | 29.3 | 27.5 | 28.7 | 28.5 | 28.3 | **56.9** | 31.6 | 31.0 | 29.4 | 29.9 | 29.8 | 29.7 |
| | +ADML | **80.2** | **64.7** | **64.6** | **61.8** | **58.0** | **57.7** | **57.5** | **52.3** | **33.8** | **33.5** | **31.3** | **29.7** | **29.5** | **29.0** | 54.5 | **36.6** | **36.1** | **34.6** | **34.5** | **33.8** | **33.7** |
| ResNet-18 | AT | 83.1 | 53.3 | 51.9 | 50.8 | 49.9 | 49.7 | 49.5 | **59.1** | 27.1 | 26.3 | 25.4 | 25.3 | 25.3 | 25.1 | **61.5** | 31.0 | 30.1 | 29.5 | 28.8 | 28.9 | 28.8 |
| | +ADML | **84.5** | **61.1** | **60.8** | **58.5** | **56.7** | **56.2** | **55.6** | 56.4 | **31.6** | **30.9** | **29.5** | **28.4** | **28.2** | **27.8** | 57.3 | **35.9** | **35.5** | **33.5** | **34.7** | **34.7** | **34.6** |
| | TRADES | 83.3 | 53.0 | 52.0 | 50.9 | 50.9 | 50.8 | 50.7 | **58.5** | 27.6 | 26.8 | 26.2 | 25.9 | 25.9 | 25.8 | **60.4** | 32.0 | 31.0 | 30.2 | 29.7 | 29.6 | 29.5 |
| | +ADML | **84.0** | **62.3** | **61.9** | **59.5** | **56.7** | **56.6** | **55.9** | 57.3 | **31.7** | **31.6** | **30.0** | **28.8** | **28.5** | **28.0** | 58.4 | **37.0** | **36.0** | **33.3** | **35.0** | **34.1** | **34.0** |
| | MART | 82.5 | 54.1 | 52.8 | 51.3 | 51.5 | 50.8 | 50.8 | **58.1** | 27.7 | 26.7 | 25.5 | 25.4 | 25.1 | 25.0 | **60.6** | 31.0 | 30.2 | 29.8 | 29.0 | 29.0 | 29.0 |
| | +ADML | **84.1** | **63.3** | **62.9** | **58.7** | **56.8** | **56.4** | **56.2** | 57.4 | **32.1** | **31.7** | **30.2** | **28.9** | **28.8** | **28.5** | 57.1 | **36.6** | **36.0** | **33.6** | **35.1** | **33.9** | **33.8** |
| | AWP | 81.3 | 56.3 | 55.5 | 53.6 | 54.2 | 54.0 | 54.0 | **57.9** | 31.4 | 30.7 | 29.0 | 30.0 | 30.0 | 29.8 | **61.4** | 34.7 | 34.1 | 32.4 | 33.3 | 33.2 | 33.1 |
| | +ADML | **84.0** | **64.6** | **64.5** | **61.4** | **60.5** | **59.9** | **59.7** | 56.2 | **33.9** | **32.9** | **30.7** | **31.1** | **30.5** | **30.3** | 59.8 | **38.6** | **38.1** | **35.9** | **37.7** | **36.7** | **36.6** |
| WideResNet-28-10 | AT | 86.7 | 55.4 | 53.4 | 53.4 | 51.3 | 51.3 | 51.2 | **61.9** | 28.8 | 27.4 | 27.1 | 26.0 | 26.0 | 25.9 | **64.8** | 32.7 | 31.2 | 31.1 | 30.1 | 30.1 | 30.0 |
| | +ADML | **87.5** | **61.7** | **60.7** | **58.8** | **56.4** | **56.3** | **55.8** | 58.9 | **32.9** | **32.6** | **31.3** | **29.6** | **29.2** | **29.1** | 62.1 | **43.5** | **43.1** | **41.1** | **41.9** | **40.4** | **40.2** |
| | TRADES | 86.0 | 55.3 | 53.7 | 53.6 | 51.6 | 51.6 | 51.4 | **61.9** | 29.1 | 28.5 | 27.8 | 26.7 | 26.8 | 26.7 | **64.2** | 32.5 | 31.6 | 31.4 | 30.0 | 29.9 | 29.8 |
| | +ADML | **88.5** | **62.9** | **61.9** | **59.6** | **57.6** | **57.6** | **56.6** | 61.6 | **33.3** | **33.0** | **31.5** | **30.0** | **30.1** | **29.6** | 63.1 | **43.5** | **42.9** | **41.0** | **41.4** | **40.3** | **40.2** |
| | MART | 86.4 | 56.0 | 54.3 | 53.4 | 51.6 | 51.6 | 51.5 | **61.6** | 28.3 | 26.7 | 26.4 | 25.3 | 25.4 | 25.3 | **64.2** | 32.9 | 31.8 | 31.8 | 30.7 | 30.5 | 30.5 |
| | +ADML | **88.3** | **62.1** | **60.9** | **59.6** | **56.7** | **56.6** | **56.2** | 59.6 | **33.0** | **32.9** | **31.6** | **30.3** | **29.9** | **29.6** | 61.8 | **42.8** | **42.6** | **40.3** | **40.7** | **39.0** | **38.9** |
| | AWP | 85.9 | 60.2 | 58.9 | 57.2 | 56.9 | 56.9 | 56.8 | 62.4 | 33.0 | 32.2 | 31.1 | 30.9 | 30.9 | 30.8 | **65.4** | 36.9 | 36.0 | 35.1 | 34.8 | 34.8 | 34.7 |
| | +ADML | **88.2** | **67.5** | **67.4** | **64.2** | **63.5** | **63.2** | **63.1** | **62.5** | **39.7** | **39.3** | **36.9** | **37.6** | **37.1** | **36.8** | 64.9 | **44.6** | **44.2** | **41.9** | **43.3** | **42.1** | **42.0** |
| WideResNet-70-10 | AT | 88.1 | 56.6 | 54.8 | 55.0 | 52.8 | 53.0 | 52.8 | **64.1** | 28.4 | 27.3 | 27.4 | 26.0 | 26.4 | 25.6 | **65.3** | 34.9 | 33.4 | 33.9 | 32.2 | 32.2 | 32.1 |
| | +ADML | **88.9** | **61.5** | **61.4** | **61.0** | **56.5** | **56.3** | **56.0** | 63.3 | **30.0** | **29.2** | **28.8** | **26.9** | **26.7** | **26.4** | 61.0 | **37.7** | **37.4** | **36.9** | **33.8** | **33.0** | **32.9** |
| | TRADES | 87.7 | 56.3 | 54.7 | 55.0 | 53.4 | 53.3 | 53.3 | 63.3 | 28.7 | 27.8 | 27.9 | 26.6 | 26.2 | 26.0 | **65.7** | 34.4 | 32.6 | 33.0 | 31.5 | 31.5 | 31.4 |
| | +ADML | **89.1** | **63.9** | **63.3** | **62.7** | **59.0** | **59.6** | **59.0** | **63.4** | **31.2** | **30.8** | **30.3** | **27.5** | **27.1** | **27.0** | 61.8 | **40.2** | **39.5** | **38.8** | **36.1** | **35.5** | **35.4** |
| | MART | 88.0 | 57.4 | 55.5 | 55.4 | 52.8 | 52.8 | 52.6 | **63.2** | 28.7 | 27.5 | 27.5 | 25.8 | 26.3 | 25.6 | **65.4** | 33.8 | 32.5 | 32.4 | 31.3 | 31.3 | 31.2 |
| | +ADML | **88.5** | **61.7** | **61.3** | **60.8** | **56.7** | **56.8** | **56.6** | 62.3 | **30.1** | **30.0** | **29.4** | **29.3** | **29.1** | **28.7** | 63.2 | **41.8** | **41.0** | **40.2** | **37.7** | **36.6** | **36.5** |
| | AWP | 86.6 | 61.8 | 60.6 | 59.9 | 59.1 | 59.4 | 59.2 | 65.2 | 33.3 | 33.3 | 31.7 | 31.5 | 30.3 | 30.0 | **66.7** | 40.7 | 40.0 | 40.0 | 39.1 | 39.0 | 38.9 |
| | +ADML | **89.4** | **67.0** | **66.9** | **66.1** | **63.4** | **63.6** | **63.1** | **65.3** | **41.9** | **41.8** | **40.9** | **38.9** | **38.0** | **37.6** | 65.8 | **45.0** | **44.5** | **43.3** | **43.5** | **43.1** | **42.8** |

Table 1. Comparing adversarial robustness of various defence methods whether to the inclusion of ADML for CIFAR-10 [27], CIFAR-100 [27], Tiny-ImageNet [29] trained with VGG-16 [48], ResNet-18 [17], WideResNet-28-10 [61], and WideResNet-70-10 [61].

function used in deep learning and finally construct the objective function for ADML, of which formulation can be written as follows (see details in Appendix F):

$$\min_f \mathbb{E}_{\mathcal{D}_t}[\tau \mathcal{L}_{\text{CE}}(f(x+t), y)] + \mathbb{E}_{\mathcal{D}_0}[\mathcal{L}_{\text{CE}}(f(x), y)], \quad (10)$$

where we denote $\tau = \frac{1}{p(\mathcal{T}=t|x)} - 1$ as balancing ratio. The current AT-based defenses use an equal weight "$1/n$" to loss for all data samples: $\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{\text{Defense}}(x_i, y_i, t_i; f)$ because they presume all of perturbations have equal causal effect (successful attack) to change targets without realizing vulnerable samples. Whereas, ADML uses the balancing ratio $\tau$ to adaptively focus on vulnerable samples by reweighting the loss. To realize ADML, we describe Algorithm 1 to explain further details, where $\mathcal{L}_{\text{Defense}}(x, y, t; f)$ indicates a main body of the loss function for AT-based defenses.

# 4. Experiment

## 4.1. Implementation Details

**Datasets & Networks.** We conduct comprehensive experiments on various datasets and networks. For datasets, we use CIFAR-10 [27], CIFAR-100 [27], and two larger datasets: Tiny-ImageNet [29] and ImageNet [9]. For networks, four CNN architectures: [48, 17, 61] and four Transformer architectures: [10, 51] are used.

**Adversarial Attacks.** We adaptively set perturbation budget $\gamma$ of adversarial attacks depending on the classification difficulty of the four datasets: $8/255$ equally for CIFAR-10 [27] and CIFAR100 [27], $4/255$ for Tiny-ImageNet [29], and $2/255$ for ImageNet [9]. We prepare three standard attacks: BIM [28], PGD [33], $CW_\infty$ [3], and three advanced attacks: AP (Auto-PGD: step size-free), DLR (Auto-DLR: shift and scaling invariant), AA (Auto-Attack: parameter-free), all of which are introduced by Francesco *et al.* [8]. PGD, AP, DLR have 30 steps with random starts, where PGD has step size $2.3 \times \frac{\gamma}{30}$, and AP, DLR both have momentum coefficient $\rho = 0.75$. $CW_\infty$ uses PGD-based gradient clamping for $l_\infty$ with CW objective [3] on $\kappa = 0$.

**Adversarial Defenses.** We use four defense baselines with a standard baseline: AT [33] and three advanced defense baselines: TRADES [62], MART [55], AWP [58]. To fairly validate experiments, a perturbation generator, PGD [33] is equivalently used to generate adversarial examples for which we use the budget $8/255$ and set 10 steps with $2.3 \times \frac{\gamma}{10}$ step size in training. Especially, adversarially training Tiny-ImageNet [29] and ImageNet [9] is a computational burden, thus we employ fast adversarial training [57] with FGSM [15]. For training CNNs, we use SGD [43] with a learning rate of 0.5 scheduled by Cyclic [49] in 120 epochs and use early stopping to prevent overfitting [41]. For training Transformers, we use SGD [43] with a learn-

| | Method | CIFAR-10 | | | | | | | CIFAR-100 | | | | | | | Tiny-ImageNet | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Clean | BIM | PGD | CW | AP | DLR | AA | Clean | BIM | PGD | $CW_\infty$ | AP | DLR | AA | Clean | BIM | PGD | $CW_\infty$ | AP | DLR | AA |
| ViT-S/16 | AT | 83.5 | 49.9 | 47.3 | 46.6 | 44.9 | 44.8 | 44.7 | 59.9 | 26.6 | 25.8 | 25.1 | 24.6 | 24.5 | 24.5 | 73.8 | 35.8 | 34.1 | 33.5 | 31.9 | 31.9 | 31.8 |
| | +ADML | **88.1** | **56.8** | **55.1** | **53.8** | **51.3** | **50.7** | **50.7** | **62.7** | **32.4** | **30.7** | **29.4** | **28.4** | **27.9** | **27.7** | **75.6** | **47.7** | **46.6** | **45.6** | **45.3** | **42.7** | **42.6** |
| | TRADES | 85.0 | 51.0 | 49.4 | 48.6 | 48.1 | 48.0 | 47.8 | 59.5 | 27.3 | 26.6 | 26.3 | **25.8** | 25.9 | 25.7 | 72.9 | 38.8 | 37.8 | 37.4 | 36.1 | 36.1 | 36.0 |
| | +ADML | **87.9** | **57.6** | **56.2** | **55.1** | **52.7** | **52.0** | **51.9** | **63.2** | **35.0** | **34.7** | **33.8** | **31.6** | **31.3** | **31.2** | **75.3** | **49.1** | **48.0** | **47.1** | **45.6** | **43.0** | **43.0** |
| | MART | 85.7 | 52.4 | 49.7 | 48.9 | 46.7 | 46.7 | 46.6 | 60.9 | 28.6 | 27.9 | 27.5 | 26.5 | 26.5 | 26.4 | **77.6** | 38.6 | 37.2 | 36.8 | 35.2 | 35.2 | 35.2 |
| | +ADML | **88.0** | **57.5** | **56.1** | **54.7** | **51.9** | **51.4** | **51.4** | **62.7** | **34.4** | **32.7** | **31.7** | **30.0** | **29.3** | **29.2** | 76.5 | **48.9** | **47.7** | **46.3** | **46.0** | **42.9** | **42.9** |
| | AWP | 84.9 | 54.2 | 52.3 | 51.6 | **50.1** | **49.9** | **49.9** | 61.1 | 29.6 | 29.0 | 28.1 | 27.7 | 27.7 | 27.6 | **78.1** | 41.0 | 39.3 | 38.7 | 37.3 | 37.3 | 37.2 |
| | +ADML | **88.0** | **57.0** | **54.3** | **53.3** | **50.1** | 49.6 | 49.5 | **64.4** | **34.8** | **34.1** | **33.0** | **31.3** | **30.6** | **30.5** | 76.6 | **47.5** | **46.4** | **44.9** | **44.7** | **42.5** | **42.4** |
| ViT-B/16 | AT | 87.0 | 52.8 | 50.8 | 50.4 | 47.8 | 47.7 | 47.7 | 63.3 | 30.4 | 29.6 | 29.2 | 28.6 | 28.3 | 28.3 | 72.4 | 40.1 | 37.7 | 37.8 | 34.4 | 34.3 | 34.5 |
| | +ADML | **89.9** | **56.1** | **54.9** | **54.1** | **51.6** | **51.4** | **51.2** | **67.1** | **38.1** | **36.1** | **35.4** | **34.3** | **33.4** | **33.1** | **79.0** | **50.2** | **49.7** | **48.4** | **48.5** | **46.9** | **46.8** |
| | TRADES | 85.3 | 53.8 | 52.4 | 51.6 | 50.9 | 50.8 | 50.8 | 65.7 | 32.6 | 31.5 | 31.0 | 29.9 | 30.0 | 30.0 | 73.2 | 43.3 | 41.2 | 41.8 | 39.1 | 39.0 | 39.4 |
| | +ADML | **88.9** | **58.6** | **57.3** | **56.1** | **54.7** | **54.4** | **54.3** | **69.4** | **38.9** | **37.9** | **37.0** | **34.8** | **34.6** | **34.7** | **79.4** | **54.0** | **52.2** | **51.6** | **48.2** | **47.3** | **47.2** |
| | MART | 87.4 | 53.3 | 50.6 | 50.5 | 48.3 | 48.4 | 48.2 | 65.7 | 31.9 | 30.8 | 30.2 | 29.2 | 29.2 | 29.1 | 79.3 | 41.7 | 40.0 | 39.6 | 36.8 | 36.8 | 37.1 |
| | +ADML | **89.6** | **57.0** | **55.6** | **54.3** | **51.8** | **51.5** | **51.3** | **68.9** | **35.4** | **33.5** | **32.9** | **30.5** | **30.1** | **30.2** | **80.1** | **50.7** | **50.1** | **49.0** | **48.9** | **47.4** | **47.0** |
| | AWP | 87.4 | 54.9 | 52.9 | 51.9 | 49.8 | 49.8 | 49.7 | 66.8 | 33.4 | 31.7 | 31.3 | 30.3 | 30.2 | 30.2 | 78.7 | 45.0 | 42.2 | 42.2 | 39.5 | 39.4 | 39.7 |
| | +ADML | **90.4** | **59.1** | **56.7** | **56.6** | **53.8** | **53.6** | **52.8** | **70.0** | **37.7** | **36.1** | **35.6** | 33.3 | **32.8** | **32.9** | **79.3** | **50.8** | **49.7** | **48.0** | **48.0** | **46.0** | **45.8** |
| DeiT-S/16 | AT | 83.5 | 49.3 | 47.8 | 46.7 | 45.3 | 45.3 | 45.2 | 59.5 | 29.2 | 28.5 | 27.6 | 27.5 | 27.5 | 27.4 | **75.7** | 37.4 | 35.6 | 34.7 | 33.1 | 33.0 | 33.0 |
| | +ADML | **87.7** | **56.8** | **56.0** | **54.9** | **52.3** | **51.9** | **51.9** | **63.7** | **34.4** | **32.9** | **31.7** | **31.4** | **30.5** | **30.4** | 74.7 | **44.9** | **43.6** | **42.2** | **40.8** | **39.5** | **39.4** |
| | TRADES | 84.1 | 50.6 | 49.3 | 48.8 | 48.0 | 48.0 | 48.0 | 61.8 | 29.4 | 28.8 | 27.8 | 28.1 | 28.0 | 28.0 | 74.8 | 39.0 | 38.0 | 37.4 | 36.4 | 36.4 | 36.3 |
| | +ADML | **87.9** | **57.8** | **56.5** | **55.3** | **53.7** | **53.0** | **53.2** | **66.2** | **37.2** | **36.4** | **35.5** | **33.2** | **32.3** | **32.3** | **76.3** | **45.2** | **44.9** | **43.8** | **39.5** | **38.7** | **38.4** |
| | MART | 84.2 | 52.3 | 50.0 | 49.1 | 47.8 | 47.6 | 47.5 | 59.8 | 31.0 | 30.6 | 29.3 | 29.7 | 29.6 | 29.6 | 74.6 | 40.1 | 39.1 | 38.4 | 37.7 | 37.6 | 37.6 |
| | +ADML | **87.5** | **57.5** | **55.6** | **55.0** | **52.6** | **52.3** | **52.2** | **65.3** | **37.0** | **35.6** | **34.7** | **32.4** | **30.7** | **30.7** | **75.2** | **45.3** | **44.2** | **43.3** | **42.8** | **38.4** | **38.4** |
| | AWP | 82.3 | 53.5 | 52.3 | 51.5 | 50.5 | 50.4 | 50.4 | 60.7 | 31.8 | 31.4 | 30.2 | 31.0 | 30.0 | 30.0 | **75.4** | 41.7 | 40.9 | 39.8 | 39.0 | 39.1 | 39.0 |
| | +ADML | **86.7** | **55.9** | **53.2** | **52.6** | **50.6** | **50.5** | **50.5** | **64.7** | **39.4** | **38.1** | **36.8** | **35.7** | **34.5** | **34.5** | **75.4** | **49.4** | **47.6** | **46.5** | **45.7** | **42.8** | **42.8** |
| DeiT-B/16 | AT | 84.6 | 51.5 | 49.5 | 48.4 | 47.2 | 47.1 | 47.0 | 64.9 | 30.3 | 29.1 | 28.4 | 27.5 | 27.4 | 27.4 | **79.1** | 38.6 | 36.3 | 36.1 | 34.4 | 34.3 | 34.0 |
| | +ADML | **89.7** | **55.7** | **54.8** | **53.8** | **50.0** | **49.7** | 49.7 | **65.7** | **35.4** | **34.4** | **33.4** | **32.2** | **30.3** | **30.3** | 77.6 | **46.9** | **45.6** | **44.7** | **44.9** | **40.5** | **40.5** |
| | TRADES | 85.4 | 52.8 | 51.7 | 50.6 | 50.2 | 50.2 | 50.2 | 64.8 | 30.0 | 29.3 | 28.6 | 28.4 | 28.3 | 28.3 | 78.3 | 43.1 | 41.6 | 40.5 | 40.5 | 40.5 | 40.4 |
| | +ADML | **90.2** | **61.4** | **60.4** | **59.4** | **58.3** | **57.6** | **57.6** | **68.6** | **40.5** | **39.9** | **38.4** | **37.4** | **36.8** | **36.7** | **80.8** | **45.4** | **43.5** | **42.7** | **43.2** | **42.9** | **42.7** |
| | MART | 83.9 | 54.7 | 53.2 | 52.0 | 51.0 | 50.9 | 50.7 | 64.5 | 31.9 | 31.1 | 30.5 | 30.2 | 30.1 | 30.0 | 75.8 | 44.6 | 43.3 | 42.6 | 42.6 | 42.6 | 42.5 |
| | +ADML | **89.6** | **60.3** | **60.2** | **58.9** | **55.1** | **55.0** | **55.0** | **65.3** | **39.6** | **38.5** | **37.1** | **35.4** | **34.6** | **34.6** | **77.8** | **47.7** | **46.1** | **44.6** | **45.5** | **44.8** | **44.7** |
| | AWP | 83.3 | 54.1 | 53.1 | 52.4 | 51.8 | 51.6 | 51.5 | 65.4 | 32.3 | 31.5 | 30.4 | 30.2 | 30.1 | 30.1 | 76.6 | 42.8 | 41.4 | 40.7 | 42.8 | 42.8 | 42.7 |
| | +ADML | **88.9** | **59.0** | **57.3** | **56.4** | **54.0** | **53.8** | **53.7** | **69.7** | **39.4** | **38.3** | **37.3** | **35.2** | **34.4** | **34.4** | **80.3** | **51.2** | **50.2** | **48.9** | **49.4** | **48.1** | **48.0** |

Table 2. Comparing adversarial robustness of various defence methods whether to the inclusion of ADML for CIFAR-10 [27], CIFAR-100 [27], Tiny-ImageNet [29] trained with ViT-S/16 [10], ViT-B/16 [10], DeiT-S/16 [51], and DeiT-B/16 [51].

| | SS+CF | W | NW | CIFAR-10 PGD | CIFAR-10 $CW_\infty$ | CIFAR-10 DLR | CIFAR-10 AA | Tiny-ImageNet PGD | Tiny-ImageNet $CW_\infty$ | Tiny-ImageNet DLR | Tiny-ImageNet AA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| VGG-16 | ✓ | ✓ | ✗ | **61.7** | **59.8** | **54.8** | **54.4** | **34.9** | **32.9** | **32.4** | **32.2** |
| | ✓ | ✓ | ✓ | 52.3 | 49.4 | 48.9 | 48.8 | 25.4 | 24.9 | 24.2 | 24.2 |
| | ✓ | ✗ | ✓ | 48.0 | 47.1 | 45.8 | 45.8 | 26.0 | 25.8 | 24.5 | 24.4 |
| | ✗ | ✓ | ✗ | 52.6 | 50.0 | 49.4 | 49.4 | 28.1 | 27.0 | 26.4 | 26.4 |
| ResNet-18 | ✓ | ✓ | ✗ | **60.8** | **58.5** | **56.2** | **55.6** | **35.5** | **33.5** | **34.7** | **34.6** |
| | ✓ | ✓ | ✓ | 51.7 | 50.1 | 49.5 | 49.2 | 34.9 | 32.4 | 32.7 | 32.7 |
| | ✓ | ✗ | ✓ | 50.8 | 50.5 | 49.4 | 49.1 | 30.0 | 29.5 | 29.0 | 28.9 |
| | ✗ | ✓ | ✗ | 53.6 | 51.7 | 51.4 | 51.0 | 31.7 | 29.7 | 30.2 | 30.1 |

Table 3. Ablation study for the effects of sample-splitting plus cross-fitting (SS+CF) and Worst(W) / Non-Worst(NW) examples.

ing rate of $0.001$ on the equal experimental setup of CNNs, where $224 \times 224$ resolution is applied for all datasets and pretrained parameters on ImageNet-1k models are utilized.

**Training ADML.** After the completion of standard adversarial training [33], we apply AT-based defense methods to line 4 in Algorithm 1 for ADML. We then optimize adversarially trained CNNs in $10$ epochs using SGD [43] with a learning rate of $0.001$ scheduled by Cyclic [49], which allows empirically sufficient convergence to robustness. In addition, adversarially trained Transformers are also optimized with ADML using a learning rate of $0.0001$ on the equal experimental setup of CNNs. Note that, we set sample-splitting ratio in half (see Appendix G) for each batch, and cross-fitting is satisfied during training iterations.

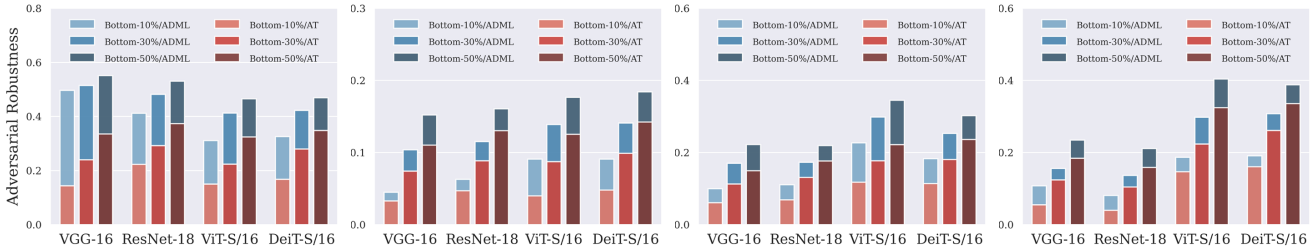
| | | PORT | +ADML | Gowal | +ADML | HAT | +ADML | SCORE | +ADML | Wang | +ADML |
|---|---|---|---|---|---|---|---|---|---|---|---|
| C-10 | Clean | 87.0 | **88.1** | 86.0 | **87.4** | 88.2 | **89.5** | 88.0 | **89.9** | 91.4 | **91.8** |
| | AA | 60.6 | **66.4** | 60.7 | **66.8** | 61.0 | **67.5** | 61.1 | **68.0** | 64.0 | **70.5** |
| C-100 | Clean | **65.9** | 65.8 | 59.2 | **59.9** | 62.2 | **62.4** | 62.0 | **62.3** | 68.1 | **68.2** |
| | AA | 31.2 | **37.9** | 30.8 | **37.7** | 31.2 | **37.3** | 31.2 | **37.1** | 35.7 | **41.1** |

Table 4. Comparing adversarial robustness of TRADES [62] using synthetic images: DDPM [18] and EDM [22] whether to the inclusion of ADML for CIFAR-10/100 with WRN-34-10: PORT [46] and WRN-28-10: Gowal [16], HAT [39], SCORE [36], Wang [56].

## 4.2. Robustness Validation on ADML

**Adversarial Robustness.** Based on our experimental setups, we have conducted enormous validations of adversarial robustness on CNNs in Table 1 and Transformers in Table 2. As shown in these tables, employing ADML on AT-based defense methods: AT [33], TRADES [62], MART [55], AWP [58] enables to largely improve adversarial robustness, compared with that of each defense method baseline. Bai *et al.* [1] have argued that Transformers cannot show noticeable adversarial robustness than CNNs, but we want to point out that the robustness of Transformers can be remarkably improved, especially in larger datasets.

**Ablation Study.** In Table 3, we conduct ablation studies on the effect of sample-splitting plus cross-fitting on robustness and the effect of considering treatments as worst examples, non-worst examples, or both on robustness, either. According to the results, only considering treatments




(a) CIFAR-10 [27] (b) CIFAR-100 [27] (c) Tiny-ImageNet [29] (d) ImageNet [9]

Figure 3. Cumulative distribution with averaged adversarial robustness for bottom-k classes against PGD [33] on four benchmark datasets. Note that, 10%, 30%, and 50% of $k$ values are applied, and perturbation budget is set to $[8/255, 8/255, 4/255, 2/255]$ on each dataset.

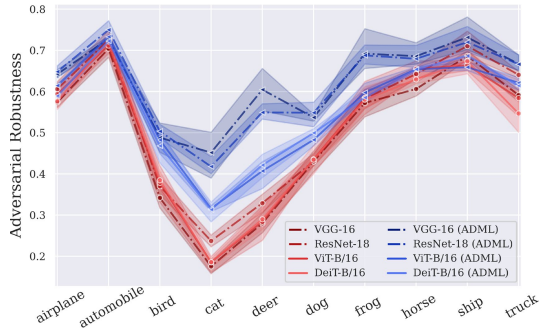



Figure 4. Distribution of adversarial robustness across whole classes on CIFAR-10. Four methods: AT [33], TRADES [62], MART [55], AWP [58] are integrated on each architecture.

as worst examples can catch actual adversarial vulnerability, thereby improving robustness much more than others.

**Utilizing Synthetic Images.** Recently, several works [46, 16, 39, 36, 56] have employed TRADES [62] utilizing synthetic images: DDPM [18] and EDM [22] to improve adversarial robustness based on the insight that data augmentation such as CutMix [60] can improve robustness [40]. To further investigate the benefits of ADML, we experiment ADML combined with TRADES on the synthetic images. Table 4 shows ADML can further improve the robustness even on synthetic images, demonstrating its efficacy.

### 4.3. Causal Analysis on ADML

**Adversarial Vulnerability.** To validate the alleviation of adversarial vulnerability existing in certain classes as in Figure 1, we evaluate the averaged adversarial robustness for the cumulative distribution of bottom-$k$ classes with respect to the network prediction. We set the $k$ value as 10%, 30%, and 50%. As in Figure 3, we can observe that AT shows noticeable vulnerability in bottom-$k$ classes, and such tendency pervades in four different datasets and architectures. If we successfully mitigate direct causal parameter of adversarial perturbations on each class, we expect apparent improvements of robustness for bottom-$k$ classes. As in the figure, we can observe the notable robustness of ADML in the vulnerable bottom-$k$ classes and corroborate its effectiveness to alleviate aforementioned phenomenon existing in current AT-based defenses. Further infographic is illustrated in Figure 4 for the integrated distribution of baselines [33, 62, 55, 58] and their corresponding ADML adoptions on each architecture, and it shows further adversarial robustness in general (Additional results in Appendix H).

| Networks | CIFAR10 | | | | Tiny-ImageNet | | | |
|---|---|---|---|---|---|---|---|---|
| | $\rho_{10}$ | $\rho_{30}$ | $\rho_{50}$ | $\rho_{\text{Avg}}$ | $\rho_{10}$ | $\rho_{30}$ | $\rho_{50}$ | $\rho_{\text{Avg}}$ |
| ResNet-18 | 53.98 | 56.89 | 52.97 | 67.33 | 4.93 | 5.39 | 5.54 | 6.49 |
| WRN-28-10 | 63.60 | 68.70 | 68.27 | 77.86 | 6.72 | 5.91 | 6.60 | 10.74 |
| ViT-B/16 | 76.78 | 85.45 | 80.03 | 84.33 | 1.54 | 1.73 | 1.91 | 3.08 |
| DeiT-B/16 | 69.36 | 74.46 | 68.63 | 69.71 | 1.15 | 1.13 | 1.22 | 2.05 |

Table 5. Relative ratio of causal parameter (%) in CIFAR-10 and Tiny-ImageNet with four architectures. Note that $k$ is set to 10, 30, 50, and Avg indicates average on whole classes in each dataset.

**Causal Parameter.** By deploying ADML, we present a way of mitigating the magnitude of causal parameter $|\theta|$. To numerically calculate $|\theta|$, we employ Eq. (8) and measure the average of $|\theta_{\text{ADML}}|$ for ADML with respect to the bottom-$k$ and whole classes, respectively. By dividing $|\theta_{\text{ADML}}|$ with $|\theta_{\text{AT}}|$, we can obtain relative ratio of causal parameter, $\rho_k := 100 \times |\theta_{\text{ADML}}|/|\theta_{\text{AT}}|$ of adversarial examples in bottom-$k$ classes. This ratio indicates that relative intensity of causal parameter compared to that of AT [33]. As in Table 5, we can observe that ADML shows less intensity of the causal parameter than AT, which means less causal effects of adversarial perturbations on target classes. From combinatorial results of preceding robustness comparison in Sec. 4.2, we corroborate that ADML indeed mitigate the intrinsic causal parameter and alleviate empirical observation in Figure 1, thus results in adversarial robustness.

## 5. Conclusion

In this paper, we observe adversarial vulnerability varies across targets and still pervades even with deeper architectures and advanced defense methods. To fundamentally address it, we build causal perspective in adversarial examples and propose a way of estimating causal parameter representing the degree of adversarial vulnerability, namely Adversarial Double Machine Learning (ADML). By minimizing causal effects from the estimated vulnerability, ADML can mitigate the empirical phenomenon as well as solidly improve adversarial robustness with other methods.

**Acknowledgments.** This work was partially supported by two funds: IITP grant funded by the Korea government (MSIT) (No.2022-0-00984) and Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD190031RD).

## References


[1] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, 2021.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229, 2020.

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy*, pages 39–57. IEEE Computer Society, 2017.

[4] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

[5] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[6] Victor Chernozhukov, Whitney Newey, Victor M Quintas-Martinez, and Vasilis Syrgkanis. Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests. In *International Conference on Machine Learning*, pages 3901–3914. PMLR, 2022.

[7] Kyle Colangelo and Ying-Ying Lee. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036*, 2020.

[8] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, 2020.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 248–255. Ieee, 2009.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[11] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. 2018.

[12] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1625–1634, 2018.

[13] Nitai Fingerhut, Matteo Sesia, and Yaniv Romano. Coordinated double machine learning. In *International Conference on Machine Learning*, pages 6499–6513, 2022.

[14] Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

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

[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*, 34:4218–4233, 2021.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

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

[19] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *2018 IEEE Symposium on Security and Privacy*, pages 19–35. IEEE, 2018.

[20] Yonghan Jung, Jin Tian, and Elias Bareinboim. Double machine learning density estimation for local treatment effects with instruments. *Advances in Neural Information Processing Systems*, 34:21821–21833, 2021.

[21] Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating identifiable causal effects through double machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12113–12122, 2021.

[22] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[23] Edward H Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

[24] Junho Kim, Byung-Kwan Lee, and Yong Man Ro. Distilling robust and non-robust features in adversarial examples by information bottleneck. *Advances in Neural Information Processing Systems*, 34:17148–17159, 2021.

[25] Junho Kim, Byung-Kwan Lee, and Yong Man Ro. Demystifying causal features on adversarial examples and causal inoculation for robust network by adversarial instrumental variable regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12302–12312, 2023.

[26] Sylvia Klosin. Automatic double machine learning for continuous treatment effects. *arXiv preprint arXiv:2104.10334*, 2021.

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

[28] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2017.

[29] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7, 2015.

[30] Byung-Kwan Lee, Junho Kim, and Yong Man Ro. Masking adversarial damage: Finding adversarial saliency for robust

and sparse network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15126–15136, 2022.

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[32] Lester Mackey, Vasilis Syrgkanis, and Ilias Zadik. Orthogonal machine learning: Power and limitations. In *International Conference on Machine Learning*, pages 3375–3383, 2018.

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

[34] Jerzy Neyman. C ($\alpha$) tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 1–21, 1979.

[35] Jerzy Neyman and Elizabeth L Scott. Asymptotically optimal tests of composite hypotheses for randomized experiments with noncontrolled predictor variables. *Journal of the American Statistical Association*, 60(311):699–721, 1965.

[36] Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, and Shuicheng Yan. Robustness and accuracy could be reconcilable by (proper) definition. In *International Conference on Machine Learning*, pages 17258–17277. PMLR, 2022.

[37] Judea Pearl. *Causality*. Cambridge university press, 2009.

[38] Yao Qin, Xuezhi Wang, Alex Beutel, and Ed Chi. Improving calibration through the relationship with adversarial robustness. *Advances in Neural Information Processing Systems*, 34:14358–14369, 2021.

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

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

[41] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104, 2020.

[42] Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

[43] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[44] Y. E. Sagduyu, Y. Shi, and T. Erpek. Iot network security from the perspective of adversarial deep learning. In *International Conference on Sensing, Communication, and Networking*, pages 1–9, 2019.

[45] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018.

[46] Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? In *International Conference on Learning Representations*, 2022.

[47] Carl-Johann Simon-Gabriel, Yann Ollivier, Leon Bottou, Bernhard Schölkopf, and David Lopez-Paz. First-order adversarial vulnerability of neural networks and input dimension. In *International Conference on Machine Learning*, pages 5809–5817. PMLR, 2019.

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[49] Leslie N Smith. Cyclical learning rates for training neural networks. In *IEEE Winter Conference on Applications of Computer Vision*, pages 464–472. IEEE, 2017.

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

[51] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021.

[52] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[54] Xianmin Wang, Jing Li, Xiaohui Kuang, Yu an Tan, and Jin Li. The security of machine learning in an adversarial setting: A survey. *Journal of Parallel and Distributed Computing*, 130:12 – 23, 2019.

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

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

[57] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020.

[58] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020.

[59] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations*, 2018.

[60] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable fea-

tures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[61] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, 2016.

[62] 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, *International Conference on Machine Learning*, volume 97, pages 7472–7482, 09–15 Jun 2019.

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

[64] Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *International Conference on Learning Representations*, 2019.

[65] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021.

[66] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020.