# BIASED MULTI-DOMAIN ADVERSARIAL TRAINING

**Anonymous authors**
Paper under double-blind review

## 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 using freely available 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 an additional analysis based on shuffle testing. Our experimental results indicate that BiaMAT can effectively utilize the robust and non-robust features present in various auxiliary datasets. 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.

## 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 (Dhillon et al., 2018; Xie et al., 2019) 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. (2018), the sample complexity of adversarially robust generalization is substantially higher than that of standard generalization. To address this issue, several recent studies (Carmon et al., 2019; Stanforth et al., 2019) 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 adversarially robust generalization ability of a classifier on a primary dataset based on the use of auxiliary datasets. The proposed method yields the desired effect based on the following assumption:

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

Robust features (Ilyas et al., 2019) exhibit human-perceptible patterns, and if two datasets are sufficiently similar from a human perspective, it can be inferred that they share robust features. On the other hand, non-robust features are imperceptible to the human visual system, and thus, we cannot determine whether Assumption 1 is correct. Fortunately, several recent studies have provided empirical evidence in support of the existence of a common non-robust feature space among diverse image datasets (Naseer et al., 2019; Lee et al., 2021). Therefore, unlike existing state-of-the-art methods (Carmon et al., 2019; Stanforth et al., 2019), which employ in-distribution data, under BiaMAT, the distributions of the auxiliary datasets and the corresponding primary dataset can differ. For example, by applying BiaMAT, we can leverage CIFAR-100 (Krizhevsky et al., 2009), Places365 (Zhou et al., 2017), or ImageNet (Chrabaszcz et al., 2017; Deng et al., 2009) as an auxiliary dataset for adversarial training on CIFAR-10 (Krizhevsky et al., 2009).

The proposed method achieves an inductive transfer between adversarial training on the primary dataset (referred to as the "primary task") and auxiliary datasets (referred to as "auxiliary tasks"). In other words, BiaMAT learns primary and auxiliary tasks in parallel within the framework of multi-domain learning (Nam & Han, 2016), 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 (Caruana, 1997). In addition, based on the dichotomy between robust and non-robust features, 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 *consistency transfer* and *robust feature transfer* and assess the contribution of each through shuffle testing (Caruana, 1997). Our experimental results on CIFAR-10, CIFAR-100, 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. In summary, our paper makes the following contributions:

- We propose the use of BiaMAT to improve the adversarial robustness of classifiers on primary datasets by leveraging auxiliary datasets that are more accessible than extra in-distribution datasets.
- We analyze the proposed method by applying the shuffle test to show that the effects of BiaMAT arise from a combination of consistency transfer and robust feature transfer.
- We introduce a confidence-based selection strategy, which enables the proposed method to leverage diverse image datasets without resulting in negative transfer.
- By applying BiaMAT with various auxiliary datasets, we show that different datasets can be related in terms of adversarial robustness, even though they seem unrelated from a human perspective.

## 2 BIASED MULTI-DOMAIN ADVERSARIAL TRAINING

We first describe consistency learning and robust feature learning in Section 2.1. A naive method to reduce the sample complexity of adversarially robust generalization by using auxiliary datasets is presented in Section 2.2. In Section 2.3, we theoretically analyze how auxiliary datasets can induce training data amplification in multi-domain learning by using the dichotomy between robust and non-robust features. Finally, in Section 2.4, a confidence-based selection strategy is proposed to address the negative transfer problem encountered in the naive method when a large domain discrepancy exists between the primary and auxiliary datasets.

### 2.1 PRELIMINARY: ROBUST AND NON-ROBUST FEATURES

Tsipras et al. (2018) 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:

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

where $x_1$ is a robust feature that strongly correlates to the label ($p \geq 0.5$), and the remaining features $x_2, \ldots, x_{d+1}$ are non-robust features that weakly correlate to the label ($0 < \eta < 1$). 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:

$$f(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}_{\text{unif}}^{\top} \boldsymbol{x}), \text{ where } \boldsymbol{w}_{\text{unif}} = \left[0, \frac{1}{d}, \ldots, \frac{1}{d}\right]. \tag{2}$$

Most importantly, they indicated the importance of adversarial training via the following lemma:

**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 *consistency learning* and *robust feature learning*, respectively. We discuss additional related work in Appendix C.

### 2.2 A NAIVE BIASED MULTI-DOMAIN ADVERSARIAL TRAINING METHOD

Multi-domain learning (Dredze et al., 2010) is a strategy for improving the performance of tasks that solve the same problem across multiple (but related) domains by sharing information across these domains. In standard settings, domains typically share semantic features but have different image distributions. For example, in the Office dataset (Saenko et al., 2010), 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 (Naseer et al., 2019). In particular, the use of domain-agnostic adversarial examples (Naseer et al., 2019) and robust training methods that leverage different datasets (Chan et al., 2020; Lee et al., 2021) demonstrates that common adversarial spaces can exist across considerably 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.

To examine this, we consider a multi-domain learning problem on $T$ datasets $D_t = \{(\boldsymbol{x}_i^t, y_i^t)\}_{i=1}^{n_t} \subset \mathcal{X} \times \mathcal{Y}_t$, where $t \in \{1, \dots, T\}$. The hypothesis of the problem is denoted by $H = \{h_t : \mathcal{X} \to \mathcal{Y}_t\}_{t=1}^T$, and $h_t = f_t \circ g$. Here, $f_t$ is the prediction function that outputs class probabilities for dataset $D_t$, and $g$ is the shared feature embedding function. Figure 1 provides an overview of this problem. The parameterized functions $\{f_t\}_{t=1}^T$ and $g$ are trained in parallel on $\{D_t\}_{t=1}^T$. The loss function for $D_t$ is defined as $\ell_t = \mathbb{E}_{(\boldsymbol{x},y)\sim D_t}\left[\ell_{\text{adv}}(\boldsymbol{x}, y; h_t, S)\right]$, where $\ell_{\text{adv}}$ is the adversarial loss, and $S$ represents the set of perturbations an adversary can apply. Any existing adversarial losses (Madry et al., 2017; Zhang et al., 2019) can be employed for $\ell_{\text{adv}}$. We focus on the $\ell_\infty$-robustness, the most common robustness scenario considered in the field of heuristic defenses (Madry et al., 2017; Zhang et al., 2019; Lee et al., 2020). Our goal is to attain a small adversarial loss on the primary dataset. Thus, assuming that $D_1$ is the primary dataset, the proposed method minimizes the following loss:

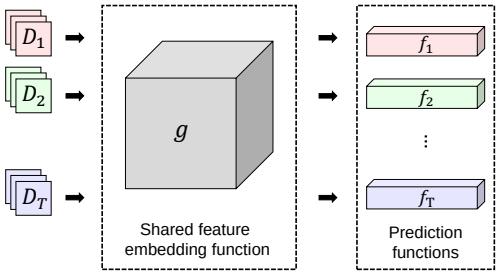



Figure 1: Conceptual network architecture of a multi-domain learning model. Domains share a feature embedding function, but each has its own prediction function.

$$\mathcal{L} = \ell_1 + \frac{\alpha}{T-1}\sum_{t=2}^{T}\ell_t, \quad \text{where } \alpha \in [0, 1]. \tag{3}$$

$\alpha$ is a hyperparameter that biases the multi-domain learning toward the primary dataset. The detailed procedure of the naive BiaMAT is described in Appendix A. Although the proposed naive BiaMAT can improve the adversarial robustness of classifiers, it can also cause negative transfer (or hurt performance) depending on the auxiliary datasets used (Table 1). To address this issue, we introduce a confidence-based selection strategy in Section 2.4.

### 2.3 THEORETICAL MOTIVATION

In this section, we analyze the proposed method from the perspective of consistency learning and robust feature learning, which are the two effects of adversarial training. In particular, based on Section 2.1, we define a simple Gaussian model to demonstrate how the proposed method induces training data amplification using an auxiliary dataset that satisfies Assumption 1.

**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:

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

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". The correlation between the two tasks can

also be represented by a covariance matrix; however, since this does not change our theoretical results, we use a correlation coefficient for simplicity. In Equation 4, $z_1$ is a robust feature that is strongly correlated with the label, whereas the other features $z_2, \cdots, z_{d+1}$ are non-robust features that are weakly correlated with the label ($0 < \eta < 1$). If the two datasets are highly correlated in terms of robust and non-robust features, from Equation 2, it is evident that the following linear classification models can achieve a high standard accuracy on the primary and auxiliary datasets, respectively, although they have a low degree of adversarial robustness:

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

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 consistency learning from 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 application of shuffle testing enables us to dissociate the compound effect of the proposed method into the effects of consistency transfer and robust feature transfer.

**Consistency transfer** First, we construct the adversarial feature vector $\tilde{\boldsymbol{z}}^{\text{adv}} = g(\tilde{\boldsymbol{x}} + \boldsymbol{\delta}) : \boldsymbol{\delta} \in S$ against our classification model, 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 (Madry et al., 2017).

**Lemma 2.** *The expectation of the adversarial feature vector against the auxiliary task is*

$$\mathbb{E}\left[\tilde{z}_1^{\text{adv}}\right] = y, \quad \mathbb{E}\left[\tilde{z}_i^{\text{adv}}\right] = (\eta - \lambda) y, \quad \text{where } i \in \{2, \cdots, d+1\} \text{ and } \eta < \lambda < 1. \tag{6}$$

Proof is in Appendix B. Our classification model is trained on $\tilde{\boldsymbol{z}}^{\text{adv}}$ by applying the stochastic gradient descent to the cross-entropy loss. 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. 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*

$$\mathbb{E}\left[\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i^{\text{adv}}}\right] = \frac{\gamma}{d}\mathbb{E}\left[\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}^{\text{adv}}) - \gamma t - \frac{1-\gamma}{2}\right] = \frac{1}{d}\mathbb{E}\left[\sigma(\boldsymbol{w}^\top \boldsymbol{z}^{\text{adv}}) - t\right] = \mathbb{E}\left[\frac{\partial \ell}{\partial z_i^{\text{adv}}}\right], \tag{7}$$

where $t = \frac{1}{2}(y+1)$. The theoretical results in the cases of $|\gamma| < 1$ (weak correlation) are discussed in Appendix B. From Lemma 2 and Theorem 1, for $i \in \{2, \cdots, d+1\}$, it can be seen that $\text{sign}\left(\mathbb{E}\left[\tilde{z}_i^{\text{adv}}\right]\right) = \text{sign}\left(\mathbb{E}\left[\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i^{\text{adv}}}\right]\right) = -y$. That is, the application of a gradient descent guides the shared feature embedding function to pay less attention to non-robust features in the images. 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 consistency learning. Therefore, this can be considered data amplification for consistency learning, and we define this effect of the proposed method as *consistency transfer*.

**Robust feature transfer** 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 consistency learning. We define this effect of the proposed method as *robust feature transfer*. To empirically assess whether the proposed method can improve the robust generalization via robust feature transfer, we conduct a shuffle test in which: (i) the class labels are shuffled among all samples in an auxiliary dataset, that is, the true labels in an auxiliary dataset are replaced by random labels; (ii) a classifier is then trained by a naive approach (Section 2.2) using the shuffled auxiliary dataset, and the trained model is evaluated on the primary task. The original shuffle test was used by Caruana (1997) to show that the advantages of multitask learning depend on the training signals for the

auxiliary tasks; here, we use the shuffle test to separately observe the two effects (consistency and robust feature transfers) induced by BiaMAT.

To investigate the effect of adversarial training on the shuffled auxiliary dataset, we replace the true labels in the auxiliary data defined in Equation 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:

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

For the case in which the auxiliary task is adversarially trained on $(\tilde{\boldsymbol{z}}, q)$ pairs by applying the stochastic gradient descent to 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 $|\gamma| = 1$, 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*

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

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 consistency transfer. 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 transfer, we assign a positive number to the weight (defined in Equation 5) 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 $w_1 > 0$, the signs of $\tilde{z}_1^{\text{adv}}$ and the auxiliary loss gradient with respect to $\tilde{z}_1^{\text{adv}}$ are*

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

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 from 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 $y^{\text{random}} = [\frac{1}{c}, \ldots, \frac{1}{c}]$, where $c$ is the number of the classes in the primary dataset, to all the auxiliary data for the shuffle test. A theoretical analysis of our Gaussian model demonstrating that the use of the expectation of random labels can achieve consistency transfer is provided in Appendix B, and the shuffle test results are described and discussed in Section 3.2.

### 2.4 A CONFIDENCE-BASED SELECTION STRATEGY

We can consider the case in which the use of a shuffled auxiliary dataset results in better adversarial robustness than that produced by the use of the normal (unshuffled) version of the dataset. In adversarial settings, a common non-robust feature space can exist between considerably different datasets (Naseer et al., 2019; Lee et al., 2021). That is, the high applicability of the proposed method arises from consistency transfer. By contrast, recent studies (Tsipras et al., 2018; Ilyas et al., 2019; Santurkar et al., 2019) have shown that robust features exhibit human-perceptible patterns. Hence, an auxiliary dataset that has a weak relationship with a primary dataset from a human perspective can assist the primary task by enhancing consistency learning 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". As an example, Ilyas et al. (2019) showed that, by

constructing a non-robust dataset based on randomly and uniformly selected target classes, it is possible to train a classifier that outperforms one that is trained with a non-robust dataset containing false (and deterministically selected) robust features. In this respect, random labels can enhance the effectiveness of the proposed method by preventing irrelevant robust feature transfer.

On this basis, we introduce a confidence-based selection strategy in BiaMAT. The confidence-based method first (i) trains a classifier from scratch on the primary dataset; (ii) after a few epochs (warm-up epochs), 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-than-threshold auxiliary data in each training batch based on the confidence in the primary classes; (iv) uses the low confidence data in a shuffle-testing manner ($y^{\text{random}}$) and the remaining auxiliary data as in the naive method ($y^t$). In summary, the confidence-based method enables neural networks to flexibly leverage diverse datasets for adversarial training, without requiring the class distribution match between the primary and auxiliary datasets. The pseudo-code for the overall procedure of BiaMAT is presented in Algorithm 1.

## 3 Experimental results and discussion

### 3.1 Experimental setup

**Datasets** We complement our analysis with experiments conducted on CIFAR-10, CIFAR-100, and ImageNet. ImageNet is resized (Chrabaszcz et al., 2017) to dimensions of $64 \times 64$ and then randomly divided into datasets that contain 100 and 900 classes, which are termed ImgNet100 and ImgNet900, respectively. SVHN (Netzer et al., 2011), 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 downsampled to a dimension of $32 \times 32$, and ImageNet32x32 is leveraged.

**Adversarial attack methods** Fast gradient sign method (FGSM) (Goodfellow et al., 2014) is an one-step attack using the sign of the gradient. Madry et al. (2017) proposed an iterative application of the FGSM method (PGD). Carlini & Wagner (CW) (Carlini & Wagner, 2017) attack is a targeted attack that maximize the logit of a target class and minimize that of ground-truth. Autoattack (AA) (Croce & Hein, 2020) is an ensemble attack that consists of two PGD extensions, one white-box attack (Croce & Hein, 2019), and one black-box attack (Andriushchenko et al., 2019).

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

**Require:** Primary dataset $D_1$, auxiliary datasets $\{D_t\}_{t=2}^T$, model parameter $\boldsymbol{\theta}$, batch size $n$, training iterations $K$, warmup iterations $K_w$, learning rate $\tau$, hyperparameters $\alpha \in \mathbb{R}^+$ and $\pi \in \mathbb{R}^+$

**for** $k = 1$ **to** $K_w$ **do**
  /* Warm-up without auxiliary datasets */
  $\mathcal{L} \leftarrow \frac{2}{n} \sum_{\{\boldsymbol{x}_i^1, \boldsymbol{y}_i^1\}_{i=1}^{n/2} \sim D_1} \ell_{\text{adv}}(\boldsymbol{x}_i^1, y_i^1; h_1, S)$
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \tau \cdot \nabla_\theta \mathcal{L}$
**end for**
confidence threshold $\omega \leftarrow \pi \frac{2}{n} \sum_{i=1}^{n/2} \max h_1(\boldsymbol{x}_i^1)$
**for** $k = K_w + 1$ **to** $K$ **do**
  $\ell_1 \leftarrow \frac{2}{n} \sum_{\{\boldsymbol{x}_i^1, \boldsymbol{y}_i^1\}_{i=1}^{n/2} \sim D_1} \ell_{\text{adv}}(\boldsymbol{x}_i^1, y_i^1; h_1, S)$
  **for** $D_t$ **in** auxiliary datasets $\{D_t\}_{t=2}^T$ **do**
    $\ell_{\text{high}}, \ell_{\text{low}} \leftarrow 0, 0$
    /* The confidence-based selection strategy for auxiliary datasets */
    **for** $\boldsymbol{x}_i^t, y_i^t$ **in** $\{\boldsymbol{x}_i^t, \boldsymbol{y}_i^t\}_{i=1}^{n/2(T-1)} \sim D_t$ **do**
      **if** $\max h_1(\boldsymbol{x}_i^t) < \omega$ **then**
        $\ell_{\text{low}} \mathrel{+}= \frac{2(T-1)}{n} \ell_{\text{adv}}(\boldsymbol{x}_i^t, y^{\text{random}}; h_1, S)$
      **else**
        $\ell_{\text{high}} \mathrel{+}= \frac{2(T-1)}{n} \ell_{\text{adv}}(\boldsymbol{x}_i^t, y_i^t; h_t, S)$
      **end if**
    **end for**
    $\ell_t \leftarrow \ell_{\text{low}} + \ell_{\text{high}}$
  **end for**
  $\mathcal{L} \leftarrow \ell_1 + \frac{\alpha}{T-1} \sum_{t=2}^T \ell_t$
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \tau \cdot \nabla_\theta \mathcal{L}$
**end for**
**Output:** adversarially robust classifier $h_1 = f_1 \circ g$

**Implementation details** In our experiments, we adopt the adversarial training methods proposed by Madry et al. (2017) and Zhang et al. (2019) as the baseline methods, denoted by AT and TRADES, respectively. On the CIFAR datasets, we use WRN28-10 (Zagoruyko & Komodakis, 2016) 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

Table 1: Shuffle test results and accuracy (%) comparison of the models trained by AT and the naive BiaMAT using auxiliary datasets on CIFAR-10. More results on the effectiveness of the naive BiaMAT can be found in Appendix G. Red numbers denote worse robustness than the vanilla AT.

| Method | Auxiliary dataset | Shuffled | Clean | AA |
|---|---|---|---|---|
| AT | - | - | 87.37 | 48.53 |
| AT+BiaMAT (naive) | SVHN | ✗ | 87.23 | 47.44 |
| | | ✓ | 87.05 | 48.53 |
| | CIFAR-100 | ✗ | 87.65 | 48.48 |
| | | ✓ | 87.12 | 49.89 |
| | Places365 | ✗ | 87.15 | 48.88 |
| | | ✓ | 88.58 | 49.24 |
| | ImageNet | ✗ | 89.01 | 50.33 |
| | | ✓ | 87.87 | 49.81 |

or 110 epochs with a batch size of 256 (128 primary and 128 auxiliary data samples, respectively). To evaluate adversarial robustness, we apply several attacks, including PGD, CW, and AA, with an $\ell_\infty$-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 (Wong et al., 2020) 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 $\pi$, are summarized and discussed in Appendix I.

### 3.2 Naive version of BiaMAT and shuffle test results

In Section 2.3, we describe the shuffle test as a tool to understand the effects of the proposed method. Table 1 lists the results of executing the shuffle test on CIFAR-10 using several auxiliary datasets. Here, the naive BiaMAT without shuffling is equivalent to Algorithm 1 where $K_w = 0$ and $\omega = 0$, whereas the shuffled version is set to $K_w = 0$ and $\omega = 1$. From Table 1, we find that the application of the naive BiaMAT with shuffling leads to better adversarial robustness than that induced by the non-shuffled counterpart (except for ImageNet), even though image-label mappings in the auxiliary datasets are disrupted. These results demonstrate that the robust feature transfer induced by the naive method using all the data in each of the SVHN, CIFAR-100, and Places365 datasets is detrimental to the primary task on CIFAR-10. Specifically, the use of SVHN exhibits the most severe negative transfer among all the auxiliary datasets tested, and this is consistent with the fact that SVHN differs most from CIFAR-10 from the robust feature perspective (see Appendix H for more details). Furthermore, SVHN is the only dataset that does not derive improved robustness compared with the baseline method (AT) in the shuffle test, indicating that it may not share enough non-robust features with CIFAR-10 to achieve consistency transfer. By contrast, Table 1 shows that the application of the naive method using ImageNet leads to significant improvements in adversarial robustness. In particular, the fact that blocking robust feature transfer using random labels leads to less performance improvements indicates that beneficial inductive transfer in terms of robust feature learning can be achieved from the auxiliary task on ImageNet. That is, the shuffle test results demonstrate that ImageNet shares a large number of robust features as well as non-robust features with CIFAR-10. More details and further analysis on the results listed in Table 1 can be found in Appendix G.

### 3.3 Adversarial robustness under various attacks

Table 2 summarizes the improvements in the adversarial robustness of the models obtained from the application of BiaMAT (the results on ImgNet100 can be found in Appendix E). The proposed method can freely use various auxiliary datasets as it avoids negative transfer through the application of the confidence-based selection strategy; in fact, a comparison of Tables 1 and 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

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

| 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** |
| CIFAR100 | 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** |

ratio $\frac{n_{\text{high}}}{n_{\text{aux}}}$, where $n_{\text{aux}}$ and $n_{\text{high}}$ 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 transfer as well as consistency transfer for an auxiliary dataset. Figure 2 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 transfer, and each of the SVHN, CIFAR100, and Places365 datasets are mostly used in the shuffle-testing manner, which is consistent with the results listed in Table 1. Additional analysis is in Appendix K.

We conduct several experiments to further investigate the proposed method. (Appendix D) 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 between the primary and auxiliary datasets is more important to BiaMAT than the number of auxiliary datasets. (Appendix J) To demonstrate that BiaMAT can achieve robust feature transfer, we construct robust datasets (Ilyas et al., 2019) from the AT and AT+BiaMAT models and normally train models from scratch on each robust dataset; the results show that the robust dataset developed using the AT+BiaMAT model results in more accurate and robust models than those trained on the robust dataset of the AT model.

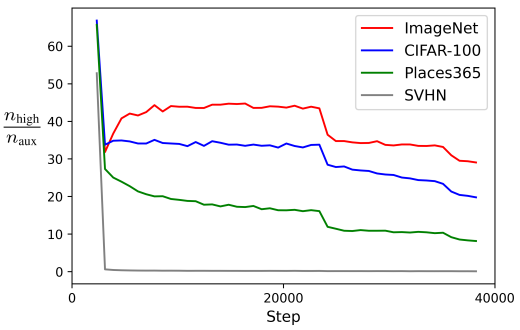



Figure 2: Ratio $\frac{n_{\text{high}}}{n_{\text{aux}}}$ for each auxiliary dataset with respect to the primary task on CIFAR-10.

### 3.4 Comparison with other related methods

Carmon et al. (2019) 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

Table 3: Comparison (accuracy %) of the effectiveness of BiaMAT with the semi-supervised (Carmon et al., 2019) and pre-training (Hendrycks et al., 2019a) 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. (2019a) | CIFAR-100 | 80.21 | 45.68 | 44.52 | 42.36 |
| | ImageNet | 87.11 | 57.16 | 55.43 | 55.30 |
| Carmon et al. (2019) | 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 |

distribution of additional data. For instance, Carmon et al. (2019) collected the in-distribution data of the CIFAR-10 dataset from 80 Million Tinyimages dataset (Torralba et al., 2008) 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 Carmon et al. (2019). In particular, we sort the ImageNet data 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 Table 3, the Carmon et al. (2019) 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. (2019a) 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 (Hendrycks et al., 2019a) models and then adversarially fine-tune them on CIFAR-10. The results in Table 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. More results and further discussions on Table 3 can be found in Appendix F.

## 4 Conclusions and future directions

In this study, we develop BiaMAT, a method that uses freely available auxiliary datasets to reduce the large gap between training and test errors in adversarial training. We demonstrate from the results of a shuffle test that the effectiveness of the proposed method can be attributed to two factors: consistency transfer and robust feature transfer. In particular, we show that while existing methods are vulnerable to negative transfer due to the distributional discrepancy between auxiliary and primary data, the proposed method 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 the proposed method 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 out-of-distribution detection strategies (Liang et al., 2018), are implemented.

## 5 REPRODUCIBILITY STATEMENT

The code and pre-trained models of our study are available at: `https://github.com/BiaMAT/BiaMAT_under_review`. The proofs of the theoretical results in our work can be found in Appendix B. The implementation details, including the training times of the models, the hyperparameters ($\alpha$ and $\pi$), datasets, and architectures, are provided in Section 3.1 and Appendix I.

## REFERENCES

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.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

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*, pp. 11190–11201, 2019.

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

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.

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.

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

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

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.

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.

Guneet S. Dhillon, Kamyar Azizzadenesheli, Jeremy D. Bernstein, Jean Kossaifi, Aran Khanna, Zachary C. Lipton, and Animashree Anandkumar. Stochastic activation pruning for robust adversarial defense. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=H1uR4GZRZ`.

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

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

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

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019b. URL `https://openreview.net/forum?id=HyxCxhRcY7`.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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.

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

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.

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. URL https://openreview.net/forum?id=eIHYL6fpbkA.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1VGkIxRZ.

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.

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*, pp. 4293–4302, 2016.

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*, pp. 12885–12895, 2019.

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

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 (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 512–523. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/0607f4c705595b911a4f3e7a127b44e0-Paper.pdf.

Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Xb8xvrtB8Ce.

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

Shibani Santurkar, Andrew Ilyas, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Image synthesis with a single (robust) classifier. In *Advances in Neural Information Processing Systems*, pp. 1262–1273, 2019.

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

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

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

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.

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

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

Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 501–509, 2019.

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

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 (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/zhang19p.html. https://github.com/yaodongyu/TRADES.

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.

## A THE OVERALL PROCEDURE USED BY THE NAIVE VERSION OF BIAMAT

**Algorithm 2** The naive version of biased multi-domain adversarial training

**Require:** Primary dataset $D_1$, auxiliary datasets $\{D_t\}_{t=2}^{T}$, hyperparameter $\alpha$, model parameter $\boldsymbol{\theta}$, batch size $n$, training iterations $K$, learning rate $\tau$

**for** $k = 1$ **to** $K$ **do**
  primary mini-batch $\{\boldsymbol{x}_i^1, \boldsymbol{y}_i^1\}_{i=1}^{n/2} \sim D_1$
  compute loss:
  **for** $t = 1$ **to** $T$ **do**
    auxiliary mini-batch $\{\boldsymbol{x}_i^t, \boldsymbol{y}_i^t\}_{i=1}^{n/2(T-1)} \sim D_t$
    $\ell_t \leftarrow \frac{2(T-1)}{n} \sum_{i=1}^{n/2(T-1)} \ell_{\text{adv}}(\boldsymbol{x}_i^t, y_i^t; h_t, S)$
  **end for**
  $\mathcal{L} \leftarrow \ell_1 + \frac{\alpha}{T-1} \sum_{t=2}^{T} \ell_t$
  model update:
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \tau \cdot \nabla_\theta \mathcal{L}$
**end for**
**Output:** adversarially robust classifier $h_1 = f_1 \circ g$

## B PROOFS

**Lemma 2.** *The expectation of the adversarial feature vector against the auxiliary task is*

$$\mathbb{E}\left[\tilde{z}_1^{\text{adv}}\right] = y, \quad \mathbb{E}\left[\tilde{z}_i^{\text{adv}}\right] = (\eta - \lambda)y, \quad \text{where } i \in \{2, \cdots, d+1\} \text{ and } \eta < \lambda < 1. \tag{8}$$

*Proof.* Our model comprises a non-linear feature embedding function $g : \mathcal{X} \to \mathcal{Z}$ and a linear classifier $f_{\gamma \boldsymbol{w}} : \mathcal{Z} \to \mathcal{Y}$. In addition, the theoretical model is based on two principles that reflect the behaviors of neural networks against adversarial examples: (i) the signs of the non-robust features $\tilde{z}_i : i \in \{2, \cdots, d+1\}$ are switched by an adversary with high probability; (ii) the sign of the

robust feature $z_1$ is not easily switched by an adversary. The objective of an adversary is to find an adversarial perturbation $\boldsymbol{\delta}^\star = \arg\max_{\boldsymbol{\delta}\in S} \tilde{\ell}(g(\tilde{x}+\boldsymbol{\delta}), \tilde{y}; \gamma\boldsymbol{w})$. Because $f_{\gamma\boldsymbol{w}}$ is linear, we can easily determine the optimal adversarial direction in the feature space $\mathcal{Z}$ using $\nabla_g \tilde{\ell}(g(\tilde{x}+\boldsymbol{\delta}), \tilde{y}; \gamma\boldsymbol{w})$. Since the scale of the adversarial perturbation in the feature space is a problem of maximizing the convex function $\tilde{\ell}(g(\tilde{x}+\boldsymbol{\delta}), \tilde{y}; \gamma\boldsymbol{w})$, as the scale of the perturbations increases, the situation is better from the adversarial point of view. However, the these principles limit the scale range. By (i), $\lambda_i > \eta = |\mathbb{E}[\tilde{z}_i]|$, where $i \in \{2, \cdots, d+1\}$; by (ii), $\lambda_1 < 1 = |\mathbb{E}[\tilde{z}_1]|$. Therefore, without loss of generality, the adversarial feature vector $\tilde{\boldsymbol{z}}^{\text{adv}}$ can be approximated by $\tilde{\boldsymbol{z}} + \lambda \cdot \text{sign}(\nabla_{\tilde{\boldsymbol{z}}} \tilde{\ell}(\tilde{\boldsymbol{z}}, \tilde{y}; \gamma\boldsymbol{w}))$ (we set $\eta < \lambda = \lambda_1 = \cdots = \lambda_{d+1} < 1$ for simplicity).

The loss function of the auxiliary task is formulated as

$$\tilde{\ell}(\tilde{\boldsymbol{z}}, \tilde{y}; \gamma\boldsymbol{w}) = -\tilde{t} \ln \sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}) - (1-\tilde{t}) \ln(1 - \sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}})), \quad \text{where } \tilde{t} = \frac{1}{2}(\tilde{y}+1). \tag{13}$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[\tilde{z}_1^{\text{adv}}\right] &= \mathbb{E}\left[\tilde{z}_1 + \lambda \cdot \text{sign}\left(\frac{\partial \tilde{\ell}}{\partial \tilde{z}_1}\right)\right] = y + \mathbb{E}\left[\lambda \cdot \text{sign}\left(\gamma w_1 (\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}) - \tilde{t})\right)\right] = y, \\ \mathbb{E}\left[\tilde{z}_i^{\text{adv}}\right] &= \mathbb{E}\left[\tilde{z}_i + \lambda \cdot \text{sign}\left(\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i}\right)\right] = \eta y + \mathbb{E}\left[\lambda \cdot \text{sign}(\gamma w_i (\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}) - \tilde{t}))\right]. \end{aligned} \tag{14}$$

We have

$$\text{sign}(\gamma w_i (\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}) - \tilde{t})) = \text{sign}(w_i) \cdot \text{sign}(\gamma \sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}) - \gamma \tilde{t}) = -y. \tag{15}$$

Hence,

$$\mathbb{E}\left[\tilde{z}_i^{\text{adv}}\right] = \eta y - \lambda y, \quad \text{where } i \in \{2, \cdots, d+1\} \text{ and } t = \frac{1}{2}(y+1). \tag{16}$$

□

**Theorem 1.** *Let $\ell(; \boldsymbol{w})$ and $\tilde{\ell}(; \gamma\boldsymbol{w})$ be the loss functions of the primary and auxiliary tasks, respectively. 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*

$$\mathbb{E}\left[\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i^{\text{adv}}}\right] = \frac{\gamma}{d}\mathbb{E}\left[\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}^{\text{adv}}) - \gamma t - \frac{1-\gamma}{2}\right] = \frac{1}{d}\mathbb{E}\left[\sigma(\boldsymbol{w}^\top \boldsymbol{z}^{\text{adv}}) - t\right] = \mathbb{E}\left[\frac{\partial \ell}{\partial z_i^{\text{adv}}}\right], \tag{9}$$

*Proof.* The expectation of the gradient of $\tilde{\ell}$ with respect to $\tilde{z}_i^{\text{adv}} : i \in \{2, \cdots, d+1\}$ is

$$\mathbb{E}\left[\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i^{\text{adv}}}\right] = \mathbb{E}\left[\frac{\gamma}{d}(\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}^{\text{adv}}) - \tilde{t})\right]. \tag{18}$$

Based on Equation 15, we obtain

$$\mathbb{E}\left[\frac{\gamma}{d}(\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}^{\text{adv}}) - \tilde{t})\right] = \frac{\gamma}{d}\mathbb{E}\left[\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}^{\text{adv}}) - \gamma t - \frac{1-\gamma}{2}\right] = \frac{1}{d}\mathbb{E}\left[\sigma(\boldsymbol{w}^\top \boldsymbol{z}^{\text{adv}}) - t\right]. \tag{19}$$

□

**Theorem 2.** *Let $\tilde{\ell}(; \gamma\boldsymbol{w})$ be the loss function of the auxiliary task. Then, if $|\gamma| = 1$, 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*

$$\text{sign}(\tilde{z}_i^{\text{adv}}) = -\gamma q = \text{sign}\left(\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i^{\text{adv}}}\right) \quad \text{with high probability.} \tag{11}$$

*Proof.* The gradient of $\tilde{\ell}$ with respect to $\tilde{z}_i : i \in \{2, \cdots, d+1\}$ is

$$\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i} = \frac{\gamma}{d}(\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}) - r), \quad \text{where } r = \frac{1}{2}(q+1). \tag{20}$$

Therefore, the adversarial feature $\tilde{z}_i^{\text{adv}}$ can be calculated as $\tilde{z}_i^{\text{adv}} = \tilde{z}_i - \lambda\gamma q$. Because $\mathbb{E}[\tilde{z}_i] = \eta y$ and $\eta < \lambda$, the sign of $\tilde{z}_i^{\text{adv}}$ is equal to $-\gamma q$ with high probability. In addition, the gradient of $\tilde{\ell}$ with respect to $\tilde{z}_i^{\text{adv}}$ is given as

$$\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i^{\text{adv}}} = \frac{\gamma}{d}(\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}^{\text{adv}}) - r). \tag{21}$$

Considering the adversarial vulnerability of our classification model, we can rewrite $\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}^{\text{adv}})$ as $\frac{1}{2}(1 - \zeta q)$, where $\zeta \in (0, 1)$. Then,

$$\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i^{\text{adv}}} = \frac{\gamma}{d}\left(\frac{1}{2} - \frac{\zeta q}{2} - \frac{q}{2} - \frac{1}{2}\right) = \frac{-\gamma q}{2d}(1 + \zeta). \tag{22}$$

Hence, the sign of $\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i^{\text{adv}}}$ is equal to $-\gamma q$ with high probability. □

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

$$\text{sign}(\tilde{z}_1^{\text{adv}}) = y, \quad \text{sign}\left(\frac{\partial \tilde{\ell}}{\partial \tilde{z}_1^{\text{adv}}}\right) = -\gamma q \quad \text{with high probability.} \tag{12}$$

*Proof.* The gradient of $\tilde{\ell}$ with respect to $\tilde{z}_1$ is

$$\frac{\partial \tilde{\ell}}{\partial \tilde{z}_1} = \gamma w_1(\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}) - r), \quad \text{where } r = \frac{1}{2}(q + 1). \tag{23}$$

Assuming that the classification model is still vulnerable to adversarial examples, the adversarial feature $\tilde{z}_1^{\text{adv}}$ is given as $\tilde{z}_1^{\text{adv}} = \tilde{z}_1 - \lambda\gamma q$. Because $\mathbb{E}[\tilde{z}_1] = y$ and $\lambda < 1$, the sign of $\tilde{z}_i^{\text{adv}}$ is equal to $y$ with high probability. In addition, the gradient of $\tilde{\ell}$ with respect to $\tilde{z}_1^{\text{adv}}$ is

$$\frac{\partial \tilde{\ell}}{\partial \tilde{z}_1^{\text{adv}}} = \gamma w_1(\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}^{\text{adv}}) - r). \tag{24}$$

Considering the adversarial vulnerability of our classification model, $\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}^{\text{adv}})$ can be rewritten as $\frac{1}{2}(1 - \zeta q)$, where $\zeta \in (0, 1)$. Then,

$$\frac{\partial \tilde{\ell}}{\partial \tilde{z}_1^{\text{adv}}} = \gamma w_1\left(\frac{1}{2} - \frac{\zeta q}{2} - \frac{q}{2} - \frac{1}{2}\right) = \frac{-\gamma q w_1}{2}(1 + \zeta). \tag{25}$$

Hence, the sign of $\frac{\partial \tilde{\ell}}{\partial \tilde{z}_1^{\text{adv}}}$ is equal to $-\gamma q$ with high probability. □

If we use $\mathbb{E}[q] = 0$ instead of sampled random labels $q$ for consistency learning, the gradient of $\tilde{\ell}$ with respect to $\tilde{z}_i : i \in \{2, \cdots, d + 1\}$ is

$$\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i} = \frac{\gamma}{d}\left(\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}) - \frac{1}{2}\right). \tag{26}$$

Based on the high standard accuracy of our classification model, the gradient of $\tilde{\ell}$ with respect to $\tilde{z}_i : i \in \{2, \cdots, d + 1\}$ can be rewritten as

$$\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i} = \frac{\gamma}{d}\left(\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}) - \frac{1}{2}\right) = \frac{\gamma}{d}\left(\frac{1}{2}(1 + \zeta\gamma y) - \frac{1}{2}\right) = \frac{\zeta y}{2d} \quad \text{with high probability.} \tag{27}$$

Therefore, the adversarial feature $\tilde{z}_i^{\text{adv}}$ can be calculated as $\tilde{z}_i^{\text{adv}} = \tilde{z}_i + \lambda y$. Because $\mathbb{E}[\tilde{z}_i] = \eta y$ and $\eta < \lambda$, the sign of $\tilde{z}_i^{\text{adv}}$ is equal to $y$ with high probability. In addition, the gradient of $\tilde{\ell}$ with respect to $\tilde{z}_i^{\text{adv}}$ is given as

$$\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i^{\text{adv}}} = \frac{\gamma}{d}\left(\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}^{\text{adv}}) - \frac{1}{2}\right). \tag{28}$$

Because $\tilde{z}_i^{\text{adv}} = \tilde{z}_i + \lambda y$, $\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}^{\text{adv}})$ can be approximated by $\frac{1}{2}(1+\gamma y)$. Then,

$$\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i^{\text{adv}}} = \frac{y}{2d}. \tag{29}$$

Hence, the signs of $\tilde{z}_i^{\text{adv}}$ and the auxiliary loss gradient with respect to $\tilde{z}_i^{\text{adv}}$ are

$$\text{sign}(\tilde{z}_i^{\text{adv}}) = y = \text{sign}\left(\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i^{\text{adv}}}\right) \quad \text{with high probability.} \tag{30}$$

### B.1 When $|\gamma| < 1$

When $|\gamma| < 1$ (weak correlation), our theorems can be replaced as follows:

**Theorem 4.** *Let $\ell(;\boldsymbol{w})$ and $\tilde{\ell}(;\gamma\boldsymbol{w})$ be the loss functions of the primary and auxiliary tasks, respectively. Then, the sign of the expectation of the gradient of $\tilde{\ell}$ with respect to $\tilde{z}_i^{\text{adv}} : i \in \{2, \cdots, d+1\}$ is*

$$\begin{aligned} &\text{sign}\left(\mathbb{E}\left[\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i^{\text{adv}}}\right]\right) = \text{sign}\left(\mathbb{E}\left[\frac{\gamma\hat{\gamma}}{d}\sigma(|\gamma|\boldsymbol{w}^\top \tilde{\boldsymbol{z}}^{\text{adv}}) - t\right]\right) = -y \\ &= \text{sign}\left(\mathbb{E}\left[\frac{1}{d}\sigma(\boldsymbol{w}^\top \boldsymbol{z}^{\text{adv}}) - t\right]\right) = \text{sign}\left(\mathbb{E}\left[\frac{\partial \ell}{\partial z_i^{\text{adv}}}\right]\right), \quad \text{where } \hat{\gamma} = \text{sign}(\gamma) \end{aligned} \tag{31}$$

**Theorem 5.** *Let $\tilde{\ell}(;\gamma\boldsymbol{w})$ be the loss function of the auxiliary task and $\hat{\gamma} = \text{sign}(\gamma)$. Then, 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*

$$\text{sign}(\tilde{z}_i^{\text{adv}}) = -\hat{\gamma} q = \text{sign}\left(\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i^{\text{adv}}}\right) \quad \text{with high probability.} \tag{32}$$

**Theorem 6.** *Let $\tilde{\ell}(;\gamma\boldsymbol{w})$ be the loss function of the auxiliary task and $\hat{\gamma} = \text{sign}(\gamma)$. Then, if $|\gamma| = 1$ and $w_1 > 0$, the signs of $\tilde{z}_1^{\text{adv}}$ and the auxiliary loss gradient with respect to $\tilde{z}_1^{\text{adv}}$ are*

$$\text{sign}(\tilde{z}_1^{\text{adv}}) = y, \quad \text{sign}\left(\frac{\partial \tilde{\ell}}{\partial \tilde{z}_1^{\text{adv}}}\right) = -\hat{\gamma} q \quad \text{with high probability.} \tag{33}$$

The theorems in the cases of $|\gamma| < 1$ show that the scale of the correlation coefficient does not change our main idea. Moreover, the training signals generated from the auxiliary task are weakened as $|\gamma|$ approaches 0 (shown in Equation 31). Note that we consider only a common robust and non-robust feature space between the primary and auxiliary data in our theoretical model. Therefore, negative transfer, induced by learning exclusive features of auxiliary tasks, cannot be described in our model.

## C Related work

**Adversarial training** Adversarial training (Madry et al., 2017; Szegedy et al., 2013) strengthens adversarial robustness by substituting adversarial examples for training samples. Given a dataset $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \in \mathbb{R}^{d+1}$ is an example in the input space and $y_i$ is its associated label, the objective of adversarial training is to make the classifier robust by minimizing the adversarial loss:

$$\min_\theta \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim D}\left[\max_{\delta \in S} \mathcal{L}(\boldsymbol{x}+\boldsymbol{\delta}, y; \theta)\right], \tag{34}$$

where $\mathcal{L}(;\theta)$ is the loss function, and $S$ represents the set of perturbations an adversary can apply, which is generally the set of $\ell_p$-bounded perturbations. As the goal of the adversary is to lower the accuracy of the classifier, the 0-1 loss should be used as the loss function $\mathcal{L}$. However, a surrogate loss will replace the 0–1 loss when there are accompanying computational problems. In previous studies, Madry et al. (2017) utilized the cross-entropy loss as the objective function, and Zhang et al. (2019) suggested the optimization of the following loss:

$$\min_h \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim D}\left[\phi(h(\boldsymbol{x})y) + \max_{\delta \in S} \phi(h(\boldsymbol{x})h(\boldsymbol{x}+\boldsymbol{\delta}))\right], \tag{35}$$

Table 4: Performance improvements (accuracy %) on CIFAR-10 following application of the proposed method using various datasets. The best result is indicated in bold.

| Method | Auxiliary dataset | Clean | $PGD^{100}$ | $CW^{100}$ | AA |
|---|---|---|---|---|---|
| AT | - | 87.37 | 50.87 | 50.93 | 48.53 |
| AT+BiaMAT | SVHN | 87.34 | 51.90 | 51.40 | 48.61 |
| | CIFAR-100 | 87.22 | 55.93 | 52.09 | 50.08 |
| | SVHN, CIFAR-100 | 87.61 | 54.58 | 52.03 | 49.88 |
| | Places365 | 87.76 | 57.00 | 51.70 | 49.48 |
| | ImageNet | **88.75** | **57.63** | **53.04** | **50.78** |
| | Places365,ImageNet | 87.88 | 56.22 | 51.86 | 49.58 |

where $h$ is a hypothesis, and $\phi$ is the classification-calibrated loss. The first term in Equation 35 increases the natural accuracy, and the second term minimizes the *difference* between the output of the natural example and the adversarial example.

**Robust and non-robust features** Ilyas et al. (2019) indicated that the adversarial vulnerabilities of neural networks could be attributed to the existence of non-robust features. Robust features are useful under adversarial perturbations, whereas non-robust features are useful only for standard classification and not adversarial robustness. They noted that the non-robust features were inherent to the dataset and were well-generalized yet brittle. Their hypothesis suggested that any two independent models trained on a given dataset were likely to learn similar non-robust features; thus, perturbations disturbing such features would also apply to both models.

Naseer et al. (2019) demonstrated the existence of a common adversarial space among different datasets by producing domain-agnostic adversarial perturbations; they showed that the adversarial examples generated using an adversarial function trained on Paintings, Cartoons, or Medical data could decrease the performance of the classifier trained on ImageNet with a high success rate. These findings indicate that a common non-robust feature space can exist between considerably different datasets.

## D THE EFFECTS OF THE USE OF MORE AUXILIARY DATASETS

We investigate the effects of the use of more auxiliary datasets under the proposed method and provide the experimental results in Table 4. The results demonstrate that the use of more auxiliary datasets does not always lead to further improvements in adversarial robustness. The results on CIFAR-10 indicate that the use of both SVHN and CIFAR-100 results in a lower degree of robustness than that achieved by using CIFAR-100 alone. Likewise, leveraging a combination of ImageNet and Places365 leads to more vulnerable models than that utilizing only ImageNet. In other words, the relationship between the primary and auxiliary datasets is more important to the proposed method than the number of auxiliary datasets.

In fact, this result is a general phenomenon that can be easily observed even in non-adversarial setting. To show this, we conducted an additional test in which: (1) the CIFAR-10 training set was classified into datasets that contain 25000, 12500, and 12500 samples, namely cifar-A, cifar-B, and cifar-C, respectively. We added uniform noise to the cifar-C dataset to sparsify the information included in the cifar-C dataset; (2) a classifier (ResNet18) was then trained on cifar-A using cifar-B and cifar-C as extra datasets with a batch size of 128 and evaluated on the test set. The results in Table 5 indicate that although cifar-B and cifar-C each result in performance improvement as an additional data set, the use of both cifar-B and cifar-C results in a test accuracy lower than that achieved by using cifar-B alone. We hypothesize that that this is because the density of information in the training dataset is more important than the total amount of information included in the training dataset in terms of the minibatch gradient descent. In other words, when DNNs are trained with a small batch size, the quality of each minibatch gradient is more important than the total amount of information in the dataset. To confirm this, we additionally run the abovementioned experiments with larger batch sizes;

Table 5: Comparison (accuracy %) of the effectiveness of data augmentation (cifar-B and cifar-C) on cifar-A.

| Batch size | Dataset | Test error (mean±std over 5 runs) |
|---|---|---|
| 128 | cifar-A | 9.58±0.21 |
| | cifar-A + cifar-B | **7.32**±0.14 |
| | cifar-A + cifar-C | 9.15±0.26 |
| | cifar-A + cifar-B +cifar-C | 7.45±0.21 |
| 256 | cifar-A | 10.48±0.21 |
| | cifar-A + cifar-B | **8.06**±0.18 |
| | cifar-A + cifar-C | 9.78±0.25 |
| | cifar-A + cifar-B +cifar-C | 8.12±0.20 |
| 384 | cifar-A | 11.08±0.35 |
| | cifar-A + cifar-B | 8.58±0.22 |
| | cifar-A + cifar-C | 10.70±0.25 |
| | cifar-A + cifar-B +cifar-C | **8.29**±0.21 |
| 512 | cifar-A | 11.49±0.20 |
| | cifar-A + cifar-B | 9.22±0.12 |
| | cifar-A + cifar-C | 11.21±0.27 |
| | cifar-A + cifar-B +cifar-C | **8.94**±0.20 |
| 1024 | cifar-A | 13.22±0.25 |
| | cifar-A + cifar-B | 10.55±0.21 |
| | cifar-A + cifar-C | 12.85±0.33 |
| | cifar-A + cifar-B +cifar-C | **10.23**±0.17 |

Table 6: Performance improvements (accuracy %) on ImgNet100 following application of the proposed method using Places365 and ImgNet900. The best results are indicated in bold.

| Primary dataset | Method | Auxiliary dataset | Clean | $PGD^{100}$ | $CW^{100}$ | AA |
|---|---|---|---|---|---|---|
| ImgNet100 | AT | - | 66.60 | 35.46 | 31.90 | 29.54 |
| | AT+BiaMAT | Places365 | **70.04** | **40.52** | 33.24 | 30.64 |
| | | ImgNet900 | 68.00 | 40.18 | **35.00** | **32.88** |

in fact, Table 5 reveal that the use of both cifar-B and cifar-C results in a higher test accuracy than that achieved by using cifar-B alone in large batch settings.

## E THE EXPERIMENTAL RESULTS ON IMGNET100

Table 6 shows that the proposed method allows improvements to both standard and robust generalizations on ImgNet100 (we use WRN16-10 on ImgNet100).

## F COMPARISON WITH OTHER RELATED METHODS

**Semi-supervised** Carmon et al. (2019) and Stanforth et al. (2019) proposed a semi-supervised learning technique by augmenting the training dataset with unlabeled in-distribution data. The main difference between them and BiaMAT is the distribution of additional data leveraged. For instance, Carmon et al. (2019) collected in-distribution data of the CIFAR-10 dataset from 80 Million Tinyimages dataset (Torralba et al., 2008) and used the unlabeled data with pseudo labels. Stanforth et al. (2019) categorized CIFAR-10 into labeled and unlabeled data. Their theoretical analysis also assumed that the unlabeled data were in-distribution, and when out-of-distribution data were used instead, a large performance drop can be observed. Therefore, while no assumptions are required for the classes of the primary and auxiliary datasets in our scenario, the semi-supervised methods

Table 7: Comparison (accuracy %) of the effectiveness of BiaMAT with the semi-supervised (Carmon et al., 2019) and pre-training (Hendrycks et al., 2019a) methods on the CIFAR datasets.

| Primary dataset | Method | Auxiliary dataset | Clean | AA |
|---|---|---|---|---|
| CIFAR-10 | Hendrycks et al. (2019a) | CIFAR-100 | 80.21 | 42.36 |
| | | ImageNet | 87.11 | 55.30 |
| | Carmon et al. (2019) | CIFAR-100 | 82.61 | 50.81 |
| | | Places365 | 83.95 | 52.81 |
| | | ImageNet | 85.42 | 53.79 |
| | | ImageNet-500k | 86.02 | 55.63 |
| | | ImageNet-250k | 86.51 | 56.27 |
| | | ImageNet-100k | 86.87 | 56.56 |
| | TRADES+BiaMAT (ours) | CIFAR-100 | 87.02 | 55.48 |
| | | Places365 | 87.18 | 55.24 |
| | | ImageNet | 88.03 | 56.64 |
| CIFAR-100 | Hendrycks et al. (2019a) | ImageNet | 59.23 | 28.79 |
| | Carmon et al. (2019) | Places365 | 56.74 | 26.22 |
| | | ImageNet | 63.45 | 27.71 |
| | | ImageNet-500k | 64.90 | 28.64 |
| | | ImageNet-250k | 66.18 | 29.49 |
| | | ImageNet-100k | 65.40 | 30.61 |
| | TRADES+BiaMAT (ours) | Places365 | 64.58 | 29.24 |
| | | ImageNet | 65.82 | 31.87 |

Table 8: Comparison (accuracy %) of the effectiveness of pre-training-based method using pre-trained ImageNet model on CIFAR-10 according to fine-tuning method.

| Fine-tuning | Clean | PGD20 | PGD100 |
|---|---|---|---|
| AT | 87.11 | 57.29 | 56.99 |
| TRADES | 83.97 | 57.17 | 57.07 |

are ineffective when the primary and auxiliary datasets do not share the same class distribution. To demonstrate this, we assign pseudo labels to the auxiliary data using a classifier trained on each primary dataset and configure each training batch to contain the same amount of primary data and pseudo-labeled data as in Carmon et al. (2019). In particular, we sort the ImageNet data based on the confidence in the primary dataset classes and select the top $(N \times 10)$k (or top $(N \times 1)$k) samples for each class in CIFAR-10 (or CIFAR-100); this is denoted by ImageNet-$(N \times 100)$k. In Table 7, the Carmon et al. (2019) method exhibits lower compatibility than the proposed method. In particular, 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.

**Pre-training** Hendrycks et al. (2019a) demonstrated that ImageNet pre-training can significantly improve adversarial robustness on CIFAR-10. Although adversarial training on ImageNet is expensive, fine-tuning on the primary dataset does not require an extensive number of computations once the pre-trained model has been acquired. However, once this has been done, it is difficult to obtain benefit from the application of cutting-edge methods in the fine-tuning phase because the hypothesis converges in the same basin in the loss landscape (Neyshabur et al., 2020) when trained from pre-trained weights. For example, as shown in Table 2, TRADES generally achieves higher adversarial robustness than AT. However, fine-tuning a pre-trained ImageNet model (Hendrycks et al., 2019a) through AT and TRADES, respectively, produces two models that exhibit similar levels of adversarial accuracy on CIFAR-10 (see Table 8). By contrast, the proposed method can directly benefit from the application of state-of-the-art adversarial training methods (Zhang et al., 2019; Carmon et al.,

Table 9: Accuracy (%) comparison of the models trained by AT and the naive version of BiaMAT using various datasets.

| Primary dataset | Method | Auxiliary dataset | Clean | PGD100 | CW100 | AA |
|---|---|---|---|---|---|---|
| CIFAR-10 | AT | - | 87.37 | 50.87 | 50.93 | 48.53 |
| | AT+BiaMAT (naive) | SVHN | 87.23 | 49.80 | 50.15 | 47.44 |
| | | CIFAR-100 | 87.65 | 50.40 | 50.79 | 48.48 |
| | | SVHN, CIFAR-100 | 87.64 | 50.79 | 51.39 | 48.90 |
| | | Places365 | 87.15 | 51.39 | 51.46 | 48.88 |
| | | ImageNet | **89.01** | **52.67** | **53.15** | **50.33** |
| | | Places365, ImageNet | 88.36 | 52.17 | 52.43 | 49.81 |
| CIFAR100 | AT | - | 62.59 | 26.80 | 26.07 | 24.13 |
| | AT+BiaMAT (naive) | Places365 | 62.78 | 27.46 | 26.76 | 24.89 |
| | | ImageNet | **65.87** | **29.55** | **28.73** | **26.26** |

2019). BiaMAT does not require complex operations and can also leverage a variety of datasets, whereas the pre-training method is effective only when a dataset that has a distribution similar to that of the primary dataset and a sufficiently large number of samples is used. To demonstrate this difference empirically, we adversarially pre-train the CIFAR-100 and ImageNet models and then adversarially fine-tune them on CIFAR-10. The results in Table 7 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 are not suitable for application of the CIFAR-10 task. Conversely, BiaMAT avoids such negative transfer through the application of a confidence-based selection strategy. That is, these results emphasize the high compatibility of the proposed method with a variety of datasets.

**Out-of-distribution data augmented training** Out-of-distribution data augmented training (OAT) (Lee et al., 2021) was proposed as a means of supplementing the training data required for adversarial training. Under the assumption that non-robust features are shared among different datasets, the authors theoretically demonstrated that using out-of-distribution data with a uniform distribution label can reduce the contribution of non-robust features and empirically demonstrated that their method promotes the adversarial robustness of a model. OAT is similar to our proposed method in that it improves adversarial robustness by using additional data with a distribution that differs from that of the primary data. In fact, OAT is identical to the shuffle-testing described in Section 2.3. In other words, OAT does not derive useful information in terms of robust feature learning from auxiliary datasets; that is, it does not achieve robust feature transfer. This is because OAT can only eliminate the contribution of features from the auxiliary dataset. Therefore, BiaMAT outperforms OAT when the auxiliary dataset has a close relationship with the primary dataset in terms of robust features. By contrast, if the auxiliary dataset contains a large amount of useful information in terms of consistency learning rather than robust feature learning, the improvements resulting from the applications of OAT and BiaMAT can be similar.

## G More results on the effectiveness of the naive version of BiaMAT

Table 9 summarizes the effects of the naive version of BiaMAT on the robust generalization performance on the CIFAR datasets. Additionally, we investigate the importance of the number of classes and that of samples per class in an auxiliary dataset in terms of negative transfer, when the correlation between the primary and auxiliary datasets is high. For this, we define a subsampling ratio $\frac{n_{\text{sub}}}{n_{\text{all}}}$, where $n_{\text{all}}$ and $n_{\text{sub}}$ denote the numbers of data samples in an auxiliary dataset and in a subset of the auxiliary dataset, respectively; Figure 2 shows the changes in the effectiveness of the naive method (under AA) on the robust generalization performance on CIFAR-10 using ImageNet with the subsampling ratio

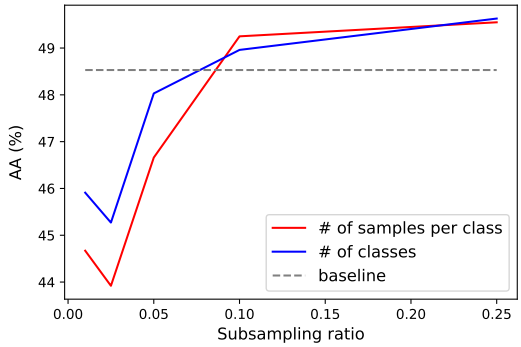



Figure 3: The changes in the effectiveness of the naive method on the robust generalization performance on CIFAR-10 (baseline: AT) using ImageNet with the subsampling ratio $\frac{n_{\text{sub}}}{n_{\text{all}}} \in \{0.01, 0.025, 0.05, 0.1, 0.25\}$.

Table 10: FID to CIFAR-10.

| Dataset | FID |
|---|---|
| SVHN | 9.55 |
| CIFAR-100 | 3.02 |
| Places365 | 4.62 |
| ImageNet | 3.21 |

$\frac{n_{\text{sub}}}{n_{\text{all}}} \in \{0.01, 0.025, 0.05, 0.1, 0.25\}$. As shown in the figure, after a certain level of auxiliary dataset size is satisfied, the influences of the number of classes and that of samples per class in the auxiliary dataset on the effectiveness of multi-domain learning are similar. However, when the auxiliary dataset size is very small ($\frac{n_{\text{sub}}}{n_{\text{all}}} \leq 0.05$), a small number of samples per class causes more negative transfer when compared to a small number of classes.

## H SVHN DIFFERS MOST FROM CIFAR-10 FROM THE ROBUST FEATURE PERSPECTIVE

We can approximate the difference between different datasets from the robust feature perspective by using a robust classifier. Specifically, we measure the Frechet inception distance (FID) (Heusel et al., 2017) between the CIFAR-10 and auxiliary datasets using the hidden representation of an adversarially trained CIFAR-10 classifier (AT) and summarize the results in Table 10. As presented, SVHN differs most from CIFAR-10 from a robust classifier perspective.

Although CIFAR-100 is the closest dataset to CIFAR-10 as shown in Table 10, CIFAR-100 is not the auxiliary dataset that resulted in the largest performance improvement through the naive BiaMAT (Table 1). This is because CIFAR-100 has a much smaller number of classes or samples per class than Places365 and ImageNet. Please refer to Appendix G for further discussions.

## I IMPLEMENTATION DETAILS

**Datasets** CIFAR-10 (Krizhevsky et al., 2009) consists of 50,000 training images and 10,000 test images in 10 classes. CIFAR-100 (Krizhevsky et al., 2009) consists of 50,000 training images and 10,000 test images in 100 classes. Both CIFAR-10 and CIFAR-100 images have sizes of $32 \times 32$ pixels. ImageNet Deng et al. (2009) consists of 1,281,167 training images and 100,000 test images in 1,000 classes. Chrabaszcz et al. (2017) provided downsampled variants of the ImageNet dataset. The ImageNet32x32 and ImageNet64x64 datasets (Chrabaszcz et al., 2017) have the same number of classes and images as ImageNet, but the images are downsampled to sizes of $32 \times 32$ and $64 \times 64$ pixels, respectively. SVHN is obtained from a very large set of images from urban areas in various countries using Google Street View. The CIFAR datasets are labeled subsets of the 80 million

Table 11: The training times of the models in our experiments.

| Primary dataset | Method | Training time (h) |
|---|---|---|
| CIFAR | AT | 34 |
| | AT+BiaMAT (naive) | 56 |
| | AT+BiaMAT | 56.5 |
| | TRADES | 52 |
| | TRADES+BiaMAT | 103 |
| ImgNet100 | AT | 119 |
| | AT+BiaMAT | 196 |

Table 12: The hyperparameter $\alpha$ for each model in Table 9

| Primary dataset | Method | Auxiliary dataset | $\alpha$ |
|---|---|---|---|
| CIFAR-10 | AT+BiaMAT (naive) | SVHN | 0.5 |
| | | CIFAR-100 | 0.5 |
| | | SVHN, CIFAR-100 | 0.5 |
| | | Places365 | 0.5 |
| | | ImageNet | 1.0 |
| | | Places365, ImageNet | 1.0 |
| CIFAR100 | AT+BiaMAT (naive) | Places365 | 0.5 |
| | | ImageNet | 1.0 |

tiny images dataset (Torralba et al., 2008), and the 80 million tiny images dataset contains images downloaded from seven independent image search engines: Altavista, Ask, Flickr, Cydral, Google, Picsearch, and Webshots. The Places365 images are queried from several online image search engines (Google Images, Bing Images, and Flickr) using a set of WordNet synonyms. The ImageNet images are collected from online image search engines and organized by the semantic hierarchy of WordNet.

**Training time** The training times of the models are summarized in Tables 11. We used a single Tesla V100 GPU with CUDA10.2 and CuDNN7.6.5. Because of the increased training dataset size (and batch size) in the proposed method, the training time was almost twice that of the baseline method. Furthermore, a comparison of AT+BiaMAT(naive) and AT+BiaMAT revealed that the proposed confidence-based selection strategy requires negligible time.

**Tables 1 and 9** For the experiments in Table 1 and 9, we executed 100 training epochs on the CIFAR datasets. The initial learning rate was set to 0.1, and the learning rate decay was applied at 60% and 90% of the total training epochs with a decay factor of 0.1. Weight decay factor and $\ell_\infty$-bound were set to 2e-4 and $\frac{8}{255}$, respectively. To observe the best performance that each auxiliary dataset can produce through the naive version of BiaMAT, we used different $\alpha$ values for each auxiliary dataset. The hyperparameter $\alpha$ for each model presented in Table 9 is summarized in Table 12.

**Table 2** For the models associated with AT, we executed 100 training epochs (including 5 warm-up epochs) on CIFAR-10, CIFAR-100, and ImgNet100. The initial learning rate was set to 0.1, and the learning rate decay was applied at 60% and 90% of the total training epochs with a decay factor of 0.1. Weight decay factor and $\ell_\infty$-bound were set to 2e-4 and $\frac{8}{255}$, respectively. Based on a recent study (Pang et al., 2021), for the models associated with TRADES, we executed 110 training epochs (including 5 warm-up epochs) on CIFAR-10 and CIFAR-100. The initial learning rate was set to 0.1, and the learning rate decay was applied at the 100th epoch and 105th epoch with a decay factor of 0.1. Weight decay factor and $\ell_\infty$-bound were set to 5e-4 and 0.031, respectively.

The hyperparameter $\alpha$ and $\pi$ for each model presented in Table 1 is summarized in Table 13. From Table 13, it can be observed that when the proposed method is applied with AT, it produces good results around $\alpha = 1.0$ and $\pi = 0.5$ regardless of the primary dataset used. However, when the proposed method is applied with TRADES, the optimal set of hyperparameters are dependent on the

Table 13: The hyperparameter $\alpha$ and $\pi$ for each model in Table 2

.

| Primary dataset | Method | Auxiliary dataset | $\alpha$ | $\pi$ |
|---|---|---|---|---|
| CIFAR-10 | AT+BiaMAT | SVHN<br>CIFAR-100<br>Places365<br>ImageNet | 1.0 | 0.55 |
| | TRADES+BiaMAT | CIFAR-100<br>Places365<br>ImageNet | 0.5 | 0.5 |
| CIFAR100 | AT+BiaMAT | Places365<br>ImageNet | 1.0 | 0.5 |
| | TRADES+BiaMAT | Places365<br>ImageNet | 1.0 | 0.3 |
| ImgNet100 | AT+BiaMAT | Places365<br>ImgNet900 | 1.0 | 0.5 |

characteristics of the primary task, such as the scale of training loss and its learning difficulty. For example, the primary task on CIFAR-10 achieves a lower training loss than that on CIFAR-100, and thus, a smaller $\alpha$ value is required when the primary dataset is CIFAR-10 than that required when the primary dataset is CIFAR-100. In addition, when the proposed method is applied to improve the sample complexity of a high-difficulty task, the confidence-based selection strategy becomes sensitive to the hyperparameter $\pi$, because the threshold used by the strategy is determined based on the confidences of the sampled primary data. Therefore, as a future research direction, we aim to develop an algorithm that can stably detect the data samples causing negative transfer.

When CIFAR-10 is the primary dataset, we use the same adversarial loss function for the primary and auxiliary tasks under BiaMAT. However, this setting can be problematic when the TRADES+BiaMAT model is trained on CIFAR-100. TRADES uses the prediction of natural examples instead of labels to maximize the adversarial loss. In this respect, when an insufficient training time is applied to a challenging dataset, such as CIFAR-100 and ImageNet, low-quality training signals can arise owing to the inaccurate predictions. Therefore, in our experiment, the cross-entropy loss with labels is used for auxiliary tasks when the primary dataset is CIFAR-100. The application of the cross-entropy loss function allows the TRADES+BiaMAT models to achieve a high level of adversarial robustness on CIFAR-100, as shown in Table 2.

**Pre-training** In the pre-training phase, the model was adversarially trained on the auxiliary dataset according to the implementation details described in Section 3.1. The fine-tuning phase commenced from the best checkpoint of the pre-training phase. We adversarially fine-tuned the entire layers of the pre-trained model on the primary dataset. The learning rate was set according to the global step over the pre-training and fine-tuning phase. For example, if the best checkpoint was acquired at the 65th epoch in the pre-training phase, the learning rate of the fine-tuning phase commenced at 0.01 and decreased to 0.001 after 25 epochs. When SVHN and CIFAR-100 were used as the auxiliary datasets, the abovementioned type of learning rate schedule rendered better robustness than that achieved by fine-tuning the model with a fixed learning rate (Hendrycks et al., 2019a).

### I.1 Ablation study on the hyperparameter $\pi$

Here, we provide the results of ablation study on $\pi$ in Table 14. From the results of the AT+BiaMAT model, the effectiveness of BiaMAT is smooth near the optimal $\pi$ when it is applied with AT. In the results of TRADES+BiaMAT, however, it can be seen that the effectiveness of the proposed method is relatively sensitive to $\pi$ when it is applied with TRADES. We speculate that this is because of the relatively complex loss function of TRADES, which introduces another regularization hyperparameter $\beta$ (Zhang et al., 2019). Therefore, in future work, we will develop advanced algorithms that adaptively control the threshold in BiaMAT for learning stability.

Table 14: The results of ablation study on $\pi$. Primary dataset: CIFAR-10; Auxiliary dataset: ImageNet.

| Method | $\pi$ | AA |
|---|---|---|
| AT+BiaMAT | 0.45 | 49.85 |
| | 0.50 | 50.35 |
| | 0.55 | 50.78 |
| | 0.60 | 50.32 |
| | 0.65 | 50.35 |
| | 0.70 | 50.69 |
| TRADES+BiaMAT | 0.45 | 56.42 |
| | 0.50 | 56.64 |
| | 0.55 | 56.21 |
| | 0.60 | 54.70 |
| | 0.65 | 54.95 |
| | 0.70 | 54.04 |

Table 15: Accuracy (%) comparison of the models (WRN34-10) trained on each robust dataset generated from the AT and AT+BiaMAT models.

| Source model | Clean | FGSM (mean±std over 5 runs) |
|---|---|---|
| AT | 87.49±0.20 | 30.79±1.16 |
| AT+BiaMAT | **88.19**±0.16 | **31.82**±1.06 |

## J Robust dataset analysis

Ilyas et al. (2019) generated a robust dataset containing only robust features (relevant to an adversarially trained model) to demonstrate their existence in images. In particular, they optimized:

$$\min_{x_r} \|g(x_r) - g(x)\|_2$$

, where $x$ is the target image and $g$ is the feature embedding function. They initialized $x_r$ as a different randomly chosen image from the training set. Thus, the robust dataset consists of optimized $x_r$–target label $y$ pairs.

To confirm robust feature transfer through application of the proposed method, we construct robust datasets from the AT and AT+BiaMAT models. We then normally train models from scratch on each robust dataset using the cross-entropy loss and list the results in Table 15. As shown, the robust dataset developed using the model trained with the proposed method results in more accurate and robust models than those trained on the robust dataset of the baseline model. The proposed method thus enables neural networks to learn better robust features via inductive transfer between adversarial training on the primary and auxiliary datasets (i.e., robust feature transfer).

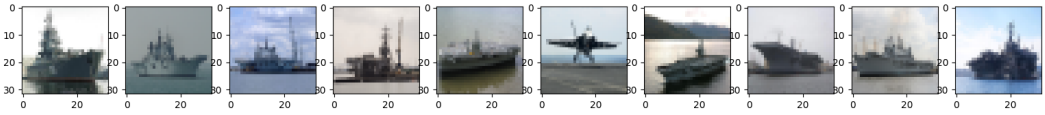
## K Additional analysis of the confidence-based selection strategy

Since robust features exhibit human-perceptible patterns, we conjecture that auxiliary data samples more related to the original dataset classes can contribute more to robust feature transfer. From this motivation, we design our algorithm to shuffle labels of the less-related samples. In particular, we adopt an automatic confidence-based sample selection strategy, widely used in existing novelty detection literature (Hendrycks et al., 2019b). To understand how the proposed confidence-based selection strategy works in practice, we analyze the ratio of samples having higher confidences than the confidence threshold (*i.e.*, $\omega$ in Algorithm 1). If a sample contributes more to learn robust features, it tends to have a higher confidence score than less contributed samples.

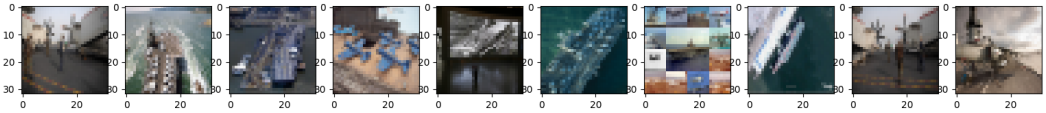
We use the AT+BiaMAT model in Table 2, trained on the CIFAR-10 dataset with the ImageNet auxiliary dataset. The model shows 88.75% clean accuracy and 50.78% robust accuracy on AA. Table

Table 16: Average higher-than-threshold ratio of the ImageNet training images by the AT+BiaMAT-trained CIFAR-10 classifier. The fine-grained ImageNet classes are mapped to CIFAR-10 superclasses by the WordNet hierarchy. "All" denotes the entire training ImageNet images. "Deer" and "Horse" classes has zero error because there is only one ImageNet class matched to each of them (Table 17).

| CIFAR-10 Superclass | Average higher-than-threshold ratio | Standard error |
|---|---|---|
| Airplane | 0.849 | 0.096 |
| Automobile | 0.706 | 0.163 |
| Bird | 0.554 | 0.143 |
| Cat | 0.501 | 0.136 |
| Deer | 0.720 | - |
| Dog | 0.592 | 0.103 |
| Frog | 0.653 | 0.070 |
| Horse | 0.819 | - |
| Ship | 0.677 | 0.215 |
| Truck | 0.763 | 0.129 |
| Others (dismatched) | 0.290 | 0.196 |
| All | 0.335 | 0.219 |



(a) The top-10 highest confident samples from "aircraft carrier" class.



(b) The top-10 lowest confident samples from "aircraft carrier" class.

Figure 4: The top-10 highest and lowest confident ImageNet training samples ("aircraft carrier" class) by the BiaMAT trained classifier on CIFAR-10.

16 shows the average higher-than-threshold ratio (*i.e*, the ratio of samples contribute to learn robust features) of ImageNet training images by the model. We show the average higher-than-threshold ratio for each CIFAR-10 superclasses, where the mapping is shown in Table 17. We match classes of two datasets by using the ImageNet synset following CINIC-10 (Darlow et al., 2018)[1].

In Table 16, we observe that the related classes show higher selection ratio (larger than 50%) than the dismatched classes (29%) and the entire average (33.5%). In other words, the auxiliary samples with CIFAR-10 superclasses contribute more to robust feature transfer than less related samples ("Others" in Table 16). We also illustrate the samples from the class "aircraft carrier", showing 87.0% higher-than-threshold ratio in Figure 4. In the figure, the highest confident samples plausibly match to the CIFAR-10 superclasses, such as "Ship" and "Airplane". On the other hand, the lowest confident samples, therefore their labels are shuffled during the training, seem to be less related to the CIFAR-10 superclasses and the original CIFAR-10 training images. The low confident samples can take a role of "out-of-distributed" dataset that can improve the confidence-based selection strategy as shown in Hendrycks et al. (2019b).

Finally, we take a look into the "Others" classes as well. While the CIFAR-10 related classes show high higher-than-threshold ratios, we also witness that some classes not highly related to the CIFAR-10 superclasses, but weakly related to them also show high higher-than-threshold ratios. For example, ("grey whale", 0.750), ("promontory", 0.749), ("breakwater", 0.734), ("dock", 0.730),

[1]We follow the official synset mapping used by CINIC-10 `https://github.com/BayesWatch/cinic-10/blob/master/synsets-to-cifar-10-classes.txt`

Table 17: The mapping between CIFAR-10 superclasses and ImageNet classes for Table 16.

| CIFAR-10 superclass | ImageNet classes |
|---|---|
| Airplane | airliner, amphibian |
| Automobile | beach wagon, convertible, sports car, ambulance, jeep, limousine, racer, cab, Model T |
| Bird | kite, white stork, ostrich, bustard, American egret, albatross, oystercatcher, red-breasted merganser, dowitcher, bee eater, redshank, red-backed sandpiper, goldfinch, black stork, crane, ruddy turnstone, bald eagle, partridge, magpie, black grouse, vulture, sulphur-crested cockatoo, junco, chickadee, American coot, spoonbill, quail, little blue heron, goose, indigo bunting, bulbul, pelican, brambling, limpkin, coucal, robin, ptarmigan, house finch, European gallinule, ruffed grouse, bittern, water ouzel, drake, peacock, jay, prairie chicken, jacamar, black swan, hummingbird, African grey, hornbill, hen, great grey owl, cock, king penguin, knot, toucan, lorikeet, flamingo, macaw |
| Cat | Persian cat, tabby, Egyptian cat, Siamese cat, Angora, lynx, cheetah, tiger cat, lion, cougar, leopard, jaguar, snow leopard, tiger |
| Deer | bison |
| Dog | Japanese spaniel, Maltese dog, Pekinese, Shih-Tzu, Samoyed, Saint Bernard, Pomeranian, white wolf, Brabancon griffon, Great Pyrenees, Newfoundland, miniature poodle, toy terrier, toy poodle, chow, kit fox, Arctic fox, Mexican hairless, coyote, red wolf, red fox, standard poodle, hyena, dhole, Eskimo dog, Great Dane, Rhodesian ridgeback, keeshond, Pembroke, Chihuahua, bull mastiff, dingo, Cardigan, timber wolf, boxer, basenji, grey fox, pug, African hunting dog, Leonberg, dalmatian |
| Frog | tailed frog, tree frog, bullfrog |
| Horse | sorrel |
| Ship | yawl, speedboat, fireboat, lifeboat, canoe, gondola |
| Truck | pickup, police van, trailer truck, minivan, moving van, tow truck, fire engine, garbage truck, tractor |

("geyser", 0.728), and ("sandbar", 0.717) are not directly included in the CIFAR-10 superclasses, but share the similar environmental backgrounds (e.g., "grey whale" and "ship" are usually on the ocean background). The multi-domain learning strategy by BiaMAT let the model learn an auxiliary information by discriminating between such weakly related auxiliary classes and the CIFAR-10 superclasses. Our BiaMAT can learn better robust features by the additional tasks to discriminate weak auxiliary classes from the target classes.

To sum up, our confidence-based selection strategy let the model learn better robust feature transfer from plausible extra images, while less plausible images improve the performance of the confidence-based selection strategy. At the same time, the multi-domain learning strategy by BiaMAT makes the model learn discriminative features between the samples highly correlated with target classes and the sample weakly correlated with targets (e.g., "grey whale"), thus BiaMAT shows a good robust feature transfer capability. Therefore, BiaMAT can learn diverse and fine-grained features using extra images related to the target classes without suffering from the negative transfer, resulting in showing better robustness generalizability.