A. Proofs

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

$$\mathbb{E}\left[\tilde{z}_1^{\text{adv}}\right] = y, \quad \mathbb{E}\left[\tilde{z}_i^{\text{adv}}\right] = (\eta - \lambda)y. \tag{8}$$

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, 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

$$\begin{aligned} &\tilde{\ell}(\tilde{\boldsymbol{z}}, \tilde{y}; \gamma \boldsymbol{w}) \\ &= -\tilde{t} \ln \sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}) - (1 - \tilde{t}) \ln (1 - \sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}})), \end{aligned} \tag{13}$$

where $\tilde{t} = \frac{1}{2}(\tilde{y} + 1)$. 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

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

Hence,

$$\mathbb{E}\left[\tilde{z}_i^{\text{adv}}\right] = \eta y - \lambda y, \tag{16}$$

where $i \in \{2, \cdots, d+1\}$ and $t = \frac{1}{2}(y + 1)$. □

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

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

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

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

□

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

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

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 $w_1 > 0$, with high probability, the signs of $\tilde{z}_1^{\text{adv}}$ and the auxiliary loss gradient with respect to $\tilde{z}_1^{\text{adv}}$ are*

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

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,

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

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 non-robust feature regularization, 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, with high probability, the gradient of $\tilde{\ell}$ with respect to $\tilde{z}_i : i \in \{2, \cdots, d+1\}$ can be rewritten as

$$\begin{aligned}\frac{\partial \tilde{\ell}}{\partial \tilde{z}_i} &= \frac{\gamma}{d}\left(\sigma(\gamma \boldsymbol{w}^\top \tilde{\boldsymbol{z}}) - \frac{1}{2}\right) \\ &= \frac{\gamma}{d}\left(\frac{1}{2}(1+\zeta\gamma y) - \frac{1}{2}\right) = \frac{\zeta y}{2d}.\end{aligned} \tag{27}$$

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

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

A.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, and $\hat{\gamma} = \text{sign}(\gamma)$. 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*

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

Theorem 5. *Let $\tilde{\ell}(;\gamma\boldsymbol{w})$ be the loss function of the auxiliary task and $\hat{\gamma} = \text{sign}(\gamma)$. Then, with high probability, the signs of $\tilde{z}_i^{\text{adv}} : i \in \{2, \cdots, d+1\}$ and the auxiliary loss gradient with respect to $\tilde{z}_i^{\text{adv}}$ are*

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

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 $w_1 > 0$, with high probability, the signs of $\tilde{z}_1^{\text{adv}}$ and the auxiliary loss gradient with respect to $\tilde{z}_1^{\text{adv}}$ are*

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

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.

B. 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; 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.

C. Robust dataset analysis

Ilyas et al. [21] 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 learning through the 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 6. 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.

D. Comparison with other related methods

Semi-supervised learning. Carmon et al. [4] and Stanforth et al. [38] 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. [4] collected in-distribution data of the CIFAR-10 dataset from 80 Million Tinyimages dataset [39] and used the unlabeled data with pseudo labels. Carmon et al. [38] 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 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 [4]. 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. [4] 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 trans-

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

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. 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

fer because of the considerable domain discrepancy between the primary and auxiliary datasets.

Pre-training. Hendrycks et al. [17] 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 [30] 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 [17] 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 bene-

Table 7. Comparison (accuracy %) of the effectiveness of BiaMAT with the semi-supervised [4] and pre-training [17] methods on the CIFAR datasets.

Primary dataset	Method	Auxiliary dataset	Clean	AA
CIFAR-10	Hendrycks et al. [17]	CIFAR-100	80.21	42.36
		ImageNet	87.11	55.30
	Carmon et al. [4]	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
	Gowal et al. [16]	Generated data [20]	85.07	57.62
	TRADES+BiaMAT (ours)	CIFAR-100	87.02	55.48
		Places365	87.18	55.24
		ImageNet	88.03	56.64
CIFAR-100	Hendrycks et al. [17]	ImageNet	59.23	28.79
	Carmon et al. [4]	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
	Gowal et al. [16]	Generated data [20]	60.66	29.94
	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

fit from the application of state-of-the-art adversarial training methods [43, 4]. 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) [24] 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. However, OAT does not derive useful information in terms of robust feature learning from auxiliary datasets. 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

Table 9. Results on CIFAR-10 when ImageNet-100k is auxiliary

Method	Clean	AA
OAT [24]	86.28	51.54
BiaMAT	**88.23**	**57.01**

Table 10. Performance improvements on CIFAR-10 (WRN16-8)

Clean			AA		
BiaMAT	[16]	BiaMAT+[16]	BiaMAT	[16]	BiaMAT+[16]
84.51	82.68	83.71	51.48	52.74	53.21

large amount of useful information in terms of non-robust feature regularization rather than robust feature learning, the improvements resulting from the applications of OAT and BiaMAT can be similar.

BiaMAT has two advantages over OAT and RST: (i) OAT and RST assume that the given auxiliary dataset is out-distribution (OOD) and in-distribution (ID), respectively. Hence, if a dataset contains both OOD and ID samples, they need an additional filtering process. On contrary, BiaMAT is an end-to-end method that does not need any filtering; (ii) If the assumptions on auxiliary datasets do not hold, OAT and RST will perform badly. *E.g.*, OAT using ImageNet-100k (100k ImageNet samples closest to CIFAR-10) as an auxiliary dataset deteriorates the robustness on CIFAR-10. Tab. 9 indicates that in that case the BiaMAT model outperforms the OAT model by a large margin.

Generated data. Recently, Gowal et al. [16] leveraged generative models [20] to artificially increase the training dataset size. They showed that state-of-the-art robust accuracy can be achieved by using the increased training dataset. To be specific, they demonstrated that their proposed method yields the desired effect under the following conditions: (i) The pre-trained non-robust classifier (pseudo-label generator) must be accurate on all realistic inputs. (ii) The generative model accurately approximate the true data distribution. From these conditions, we can infer the limitations of their method. That is, the effectiveness of their method is highly dependent on the quality of the generative and classification models that are solely trained on the original training dataset; in fact, Tab. 7 demonstrates that the use of synthetic data leads to a significant robustness improvement on CIFAR-10 ($+3.69\%$), whereas a much smaller robustness improvement on CIFAR-100 ($+1.12\%$) than that induce by BiaMAT ($+3.05\%$). In addition, Tab. 7 shows that while [16] significantly improves robustness against AA, it has no effect on Clean. Based on these, we investigate whether the combination of [16] and BiaMAT, which considerably increases Clean, has a synergistic effect. Tab. 10 indicates that BiaMAT can further improve [16].

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

E. Implementation details

In all our experiments, we employed commonly used data augmentation techniques such as random cropping and flipping. On the CIFAR datasets, we used WRN28-10 [42] and WRN34-10 for AT and TRADES, respectively. On ImgNet100, we used WRN16-10.

Datasets. The CIFAR-10 dataset [22] contains 50K training and 10K test images over ten classes. The CIFAR-100 dataset [22] includes 50K training and 10K test images over one hundred classes. Each image in CIFAR-10 and CIFAR-100 consists of 32×32 pixels. The ImageNet dataset [12] has 1,281,167 training and 100,000 test images over 1,000 classes. Chrabaszcz et al. [8] created downsampled versions of ImageNet. These datasets (ImageNet32x32 and ImageNet64x64) [8] contain the identical number of images and their classes as the original ImageNet dataset. The images therein are downsampled versions having pixel sizes of 32×32 and 64×64, 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 tiny images dataset [39], 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.

Table 1. For the experiments in Table 1, we executed 100 training epochs on CIFAR-10. 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 ℓ_∞-bound were set to 2e-4 and $\frac{8}{255}$, respectively.

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 ℓ_∞-bound were set to 2e-4 and $\frac{8}{255}$, respectively. Based on a recent study [31], for the models associated with TRADES, we executed 110 training epochs (including 5 warm-up epochs) on the CIFAR datasets and ImgNet100. 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 ℓ_∞-bound were set to 5e-4 and 0.031, respectively.

The hyperparameter α and π for each model presented in Table 1 is summarized in Table 12. From Table 12, 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 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 α 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 π, 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 [17].

E.1. Ablation study on the hyperparameter π

Here, we provide the results of ablation study on π in Table 13. From the results of the AT+BiaMAT model, the effectiveness of BiaMAT is smooth near the optimal π 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 π 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 β [43]. Therefore, in future work, we will develop advanced algorithms that adaptively control the threshold in BiaMAT for learning stability.

F. 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 primary dataset classes can contribute more to robust feature learning. From this motivation, we design our algorithm to use the expectation of random labels for the less-related samples. In particular, we adopt an automatic confidence-based sample selection strategy, widely used in existing novelty detection literature [19]. 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.*, ω 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 14 shows the average higher-than-threshold ratio (*i.e*, the ratio of samples contribute to learn

Table 12. The hyperparameter α and π for each model in Table 2

.

Primary dataset	Method	Auxiliary dataset	α	π
CIFAR-10	AT+BiaMAT	SVHN CIFAR-100 Places365 ImageNet	1.0	0.55
	TRADES+BiaMAT	CIFAR-100 Places365 ImageNet	0.5	0.5
CIFAR100	AT+BiaMAT	Places365 ImageNet	1.0	0.5
	TRADES+BiaMAT	Places365 ImageNet	1.0	0.3
ImgNet100	AT+BiaMAT	Places365 ImgNet900	1.0	0.5

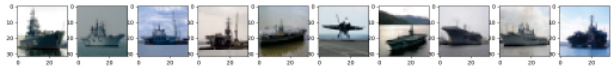
Table 13. The results of ablation study on π. Primary dataset: CIFAR-10; Auxiliary dataset: ImageNet.

Method	π	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

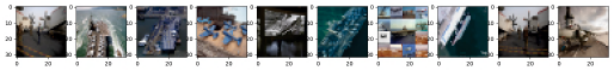
robust features) of ImageNet training images by the model. We show the average higher-than-threshold ratio for each CIFAR-10 superclasses. We match classes of two datasets by using the ImageNet synset following CINIC-10 [11][1].

In Table 14, 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 learning than less related samples ("Others" in Table 14). 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

[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



(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

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 [19].

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), ("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

Table 14. 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 **??**).

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

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 features 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 learning 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.

From these observations, we conclude that by learning robust features from extra images but related to the primary dataset, a model can learn more diverse and fine-grained features, resulting in better robustness generalizability.