Mixed Perturbation: Generating Directionally Diverse Perturbations for Adversarial Training

Changhun Hyun and Hyeyoung Park *

Posted Date: 1 October 2024

doi: 10.20944/preprints202410.0073.v1

Keywords: adversarial robustness; adversarial training; adversarial perturbations; evasion attack; multi-task learning

 doi:10.20944/preprints202410.0073.v1

Article

Mixed Perturbation: Generating Directionally Diverse Perturbations for Adversarial Training

Changhun Hyun [1] and Hyeyoung Park [2,*]

[1] School of Computer Science and Engineering, AI-driven Convergence Software Education Research Program, Kyungpook National University, Daegu 702-701, Korea
[2] School of Computer Science and Engineering, Kyungpook National University, Daegu 702-701, Korea
* Correspondence: hypark@knu.ac.kr

Abstract: The adversarial vulnerability of deep learning models is a critical issue that must be addressed to ensure the safe commercialization of AI technologies. Although numerous studies on adversarial defense methods are actively being conducted from various perspectives, most of them still provide limited robustness, and even the relatively trusted adversarial training is no exception. To develop more reliable defense methods, ongoing research exploring the properties and causes of adversarial vulnerabilities is essential. In this study, we focus on a hypothesis regarding the existence of adversarial examples: The adversarial examples represent low-probability "pockets" in the manifold. Assuming that the hypothesis holds true, we propose a method for generating perturbation: "mixed perturbation (MP)", which aims at discovering diverse pocket samples in a defensive perspective. The proposed method generates perturbations by leveraging information from both the main task and auxiliary tasks in multi-task learning scenarios, combining them through random weighted summation. The generated mixed perturbation intends to maintain the primary directionality of the main task perturbation to improve the model's main task recognition performance while introducing variability in the perturbation directions. We then utilize them for adversarial training to form more robust decision boundary. Through experiments and analyses conducted on five benchmark datasets, we validated the effectiveness of our proposed method.

Keywords: adversarial robustness; adversarial training; adversarial perturbations; evasion attack; multi-task learning

1. Introduction

In recent years, deep learning technology has rapidly advanced and is now widely utilized across various domains [1]. However, the security measures related to the stability and reliability of the AI technologies remain insufficient. The misuse and abuse cases using deepfakes [2], Generative AI [3], and other unethical applications [4] have become far from trivial. In early 2024, the European Union (EU) passed the AI Act [5], a regulatory law aimed at addressing issues related to AI misuse. Consequently, there is a growing interest and demand for stability and reliability in current AI systems.

In line with this situation, adversarial robustness of deep neural networks (DNNs) has received increasing attention. In 2013, Szegedy published research demonstrating that neural networks exhibit an intriguing property: adversarial vulnerability [6], which refers to the phenomenon where neural networks fail to defend against adversarial attacks designed to cause malfunction. Numerous studies in the literature have consistently emphasized the need for adequate countermeasures against this issue [7,8].

Adversarial attacks can be categorized based on their objectives, methodologies, and environments. Broadly, they can be divided into evasion attacks, poisoning attacks, model extraction attacks, and inversion attacks [9]. This study aims to enhance the robustness of

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 October 2024 doi:10.20944/preprints202410.0073.v1

DNNs against evasion attacks, which are the most widely researched type in the literature, and mainly addresses gradient-based attacks [6,10–13], which leverage gradient information from the input to generate adversarial examples.

Evasion attacks occur during the inference stage, where subtle and minor noise are added to the input data to create adversarial examples. Figure 1 illustrates the process of performing gradient-based evasion attack in a white-box attack environment, along with generated adversarial examples $\tilde{\boldsymbol{x}}$ and perturbations $\boldsymbol{\eta}$ according to different original input $\boldsymbol{x}$. While humans perceive these samples as belonging to the original class y, a DNN model $\boldsymbol{f}$ recognizes the input as a class other than the original class: $\boldsymbol{f}(\boldsymbol{\theta}, \tilde{\boldsymbol{x}}) \neq y$ (untargeted attack) or as a class t designated by the attacker: $\boldsymbol{f}(\boldsymbol{\theta}, \tilde{\boldsymbol{x}}) = t$ (targeted attack). This subtlety of evasion attacks can be particularly critical in fields such as medical diagnostic systems [14] and autonomous vehicles [15], where they may easily lead to significant loss of life and property.

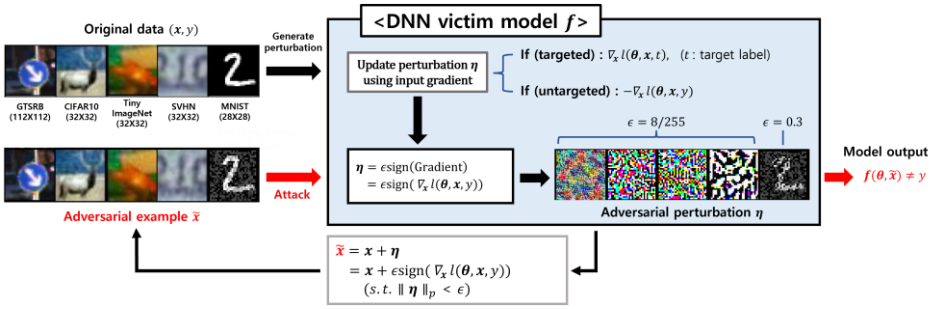


Figure 1. Illustration of the gradient-based adversarial attack process performed in a white-box attack environment, and examples of adversarial perturbations generated for five datasets.

As efforts to address such adversarial threats, adversarial defenses against evasion attacks have been studied from various perspectives, including gradient masking [16,17], defensive distillation [18,19], input transformation [20,21], adversarial detection [22–24], randomized smoothing [25], and adversarial training [6,10,12,26–28]. A summary of the methodologies and limitations of existing defenses is as follows:

- **Gradient masking** aims to prevent effective adversarial samples by adding non-differentiable layers or random noise to the model, blocking gradient-based attacks. However, it remains vulnerable to transfer-based black-box and adaptive attacks [29,30].
- **Defensive distillation** uses knowledge distillation to train a student model using soft labels obtained from a teacher model, reducing model's sensitivity to noise. Initially seen as strong, it was later shown to be insufficient against subtle and powerful attacks such as the C&W attack [13].
- **Input transformation** injects randomness or use transformation methods such as JPEG compression [31] and Gaussian blurring [32] to weaken adversarial attacks by reducing perturbations. However, these transformations can lead to information loss, degrading baseline model performance [33,34].
- **Adversarial detection** employs a separate detector to identify adversarial samples based on a threshold, utilizing statistical property-based consistency checks and ensemble models [35]. However, they struggle to defend against new or diverse attack types [36].
- **Randomized smoothing** is the application of differential privacy to adversarial robustness. By adding random noise to inputs and averaging predictions, it allows to reduce the model's sensitivity to small perturbations. Despite being considered as a reliable defense, it offers limited robustness against specific norm-bounded attacks, and efforts are ongoing to address this issue [37].
- **Adversarial training** utilizes adversarial samples in training to create a robust decision boundary. Its limitations include high computational costs, trade-off between clean and adversarial accuracies [38], and reduced robustness against unseen attacks, leading to ongoing research [27,28].

 doi:10.20944/preprints202410.0073.v1

Most of these defense strategies offer limited robustness [29,33,34,36–38]. The limited robustness can be attributed to the diversity of adversarial attack environments and methods; however, the fundamental reason lies in the lack of precise understanding of the causes of adversarial vulnerabilities. Although there are several hypotheses regarding the causes of adversarial vulnerabilities, there is no consensus yet [39]. Nevertheless, to develop reliable defense methods capable of defending any attack, it is essential to continuously study the nature of adversarial vulnerabilities.

An interesting approach to adversarial perturbations is the investigation of the direction of adversarial perturbations [40–42]. In Mao's research [40], the multi-task adversarial attack aims to generate perturbations directed at deceiving both tasks, utilizing auxiliary task information in the update process. The direction-aggregated attack [41] introduces a method that injects multiple random directions and averages them to provide diversity in the perturbation generation direction, thereby increasing the transferability of adversarial samples. The sibling attack [42] proposes a method that utilizes facial attribute tasks associated with facial images in facial recognition systems, updating perturbations alternately for each task to enhance the transferability of adversarial samples. These previous studies are related to this research in that they utilize auxiliary tasks or employ randomness for generating perturbations. However, they all have an offensive purpose to increase the attack success rate.

Inspired by the previous studies, we propose a method for generating perturbation: "mixed perturbation (MP)" which aims at discovering a more diverse set of pockets to form a more robust decision boundary. The contributions are summarized as follows:

- The proposed MP method leverages auxiliary tasks related to the main task within a multi-task learning model to generate perturbations, maintaining the primary directionality needed to improve the adversarial accuracy of the main task. At the same time, directional diversity is injected through a random weighted summation of the main task perturbation and those from the auxiliary tasks.
- In line with our previous work [43], the proposed method is applicable even when no auxiliary task is given for multi-task learning and provides better robustness in addition to improving the generalization performance of the multi-task model.
- We conduct the experiments on five benchmark datasets, demonstrating that the proposed method can improve adversarial robustness (adversarial accuracy) and recognition performance on original data (clean accuracy) in several attack methods and datasets. We also analyzed the relationship between the characteristics of data distribution and the diversity of perturbation directions, as well as their impact on the model's performance.

2. Materials and Methods

2.1. Generating Mixed Perturbation

Among the hypotheses regarding the existence of adversarial examples, we focus on the one from Szegedy's study: *the adversarial examples represent low-probability (high-dimensional) "pockets" in the manifold* [6]. In a defensive perspective, the concept of pockets around the decision boundary can now be associated with the adversarial example generation for adversarial training. In conventional adversarial training, the direction of the perturbation is determined based on the class labels of the main task, which is shown in Figure 2. The dashed square box represents the l_∞-norm bound constraint, and ϵ is the attack budget that limits the amount of change in pixel values. When performing untargeted attack, these pockets $\tilde{\boldsymbol{x}}$ can be found by adding perturbations $\boldsymbol{\eta}$ optimized in the direction that causes misclassification to a different class than the actual class y. The generation of adversarial example can be generally formulated as

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\eta}, \qquad \boldsymbol{\eta} = \epsilon \cdot \mathrm{sign}\big(\nabla_{\boldsymbol{x}} l(\boldsymbol{\theta}, \boldsymbol{x}, y)\big)., \qquad \|\boldsymbol{\eta}\|_p \leq \epsilon, \tag{1}$$

where l denotes the loss function, $\boldsymbol{\theta}$ is the model parameters, and $\|\boldsymbol{\eta}\|_p$ represents the l_p-norm constraint of the perturbation.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 October 2024 doi:10.20944/preprints202410.0073.v1

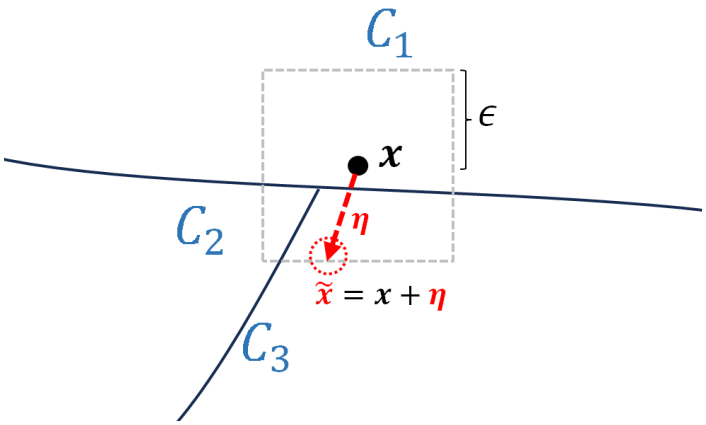


Figure 2. Generation of perturbation using main task information.

If the adversarial examples generated in this way are sufficient to form robust enough decision boundaries, then the adversarial robustness of the model should be achieved in

a substantial level. However, adversarially trained models often result in adversarial accuracies that are significantly lower than clean accuracies. This phenomenon can be interpreted on the one hand as a lack of model capacity [10,12] to learn strong enough robustness but on the other hand, the models with large capacity do not always guarantee higher robustness [45], and over-parameterization rather results in low robust generalization [44]. From a different interpretation, we assume the cause of low adversarial accuracy as a failure of conventional adversarial training method to find a diverse enough set of pockets to form sufficiently robust decision boundary.

In this regard, we seek a way to introduce diversity in the generation direction of adversarial perturbations with the objective of improving the adversarial accuracy of the main task. To enhance the adversarial accuracy of the main task from a defensive perspective in adversarial training, it is necessary to inject diversity with intention in the generation of adversarial perturbation, rather than relying on random directions. In a multi-task model, the auxiliary tasks are sets of class labels that are related to the main task and therefore contain information about the main task. In other words, the perturbations generated by the auxiliary task are likely to align in direction with those produced by the main task. Thus, perturbations generated using auxiliary tasks can be an appropriate means of providing diversity while maintaining the main directionality needed to fool the main task.

When there are a total of M tasks $(\mathbf{T}_0, \ldots, \mathbf{T}_m, \ldots, \mathbf{T}_M)$, where $\mathbf{T}_0$ represents the main task and $\mathbf{T}_\mathrm{m}$ is m th auxiliary task, a dataset for multi-task learning can be defined as $\mathrm{D} = \{(\boldsymbol{x}, y^0, \ldots, y^m, \ldots, y^M)\}$, where y^0 represents the target labels for the main task and y^m is those for the mth auxiliary task. The mixed perturbation $\boldsymbol{\eta}_{MP}$ proposed in this study integrates perturbations generated from the main task and auxiliary tasks, which can be formulated as follows:

$$\boldsymbol{\eta}_m = \underset{\boldsymbol{\eta}}{\operatorname{argmax}}\, l(\boldsymbol{\theta}_m, \boldsymbol{x} + \boldsymbol{\eta}, y^m),\ \text{s.t.}\ \|\boldsymbol{\eta}\|_p \leq \epsilon,$$
$$\boldsymbol{\eta}_{MP} = \sum_{m=0}^{M} \alpha_m \boldsymbol{\eta}_m, \qquad \sum_{m=0}^{M} \alpha_m = 1, \qquad (0.5 \leq \alpha_0 < 1.0). \tag{2}$$

Here, $\boldsymbol{\theta}_m$ represents the set of parameters for task $\mathbf{T}_m$ in a hard parameter sharing multi-task model, consisting of the shared layer weights $\boldsymbol{\theta}_s$ and the task-specific head weights $\boldsymbol{\theta}_{\mathbf{T}_m}$. The perturbation weight α_m for each task is randomly assigned for directional diversity, while ensuring that the total sum of all weights equals 1. In addition, as this study aims to improve adversarial accuracy of the main task, the minimum value of α_0 is set to 0.5 for ensuring the directionality of the main task perturbation.

To make the concept of MP clearer, we can consider a scenario where two tasks (main task and one auxiliary task) are given. As illustrated in Figure 3, C_{mk} represents the kth class label of the mth

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 October 2024 doi:10.20944/preprints202410.0073.v1

task, which originally belongs to the group of class labels $C_{mk} \in \{1, \dots, K_m\}$. When generating an adversarial perturbation to misclassify the clean data $\boldsymbol{x}$ which originally belongs to C_{01} as class C_{03}, the perturbation will be generated in the direction of $\boldsymbol{\eta}_0$, crossing the decision boundary between C_{01} and C_{03}. From the viewpoint of the auxiliary task, a perturbation will be generated in the direction of $\boldsymbol{\eta}_m$, causing $\boldsymbol{x}$ which originally belongs to C_{m1}, to be misclassified as C_{m2}. Therefore, if the auxiliary task is not identical to the main task, the resulting perturbation will have a directional distinction that cannot be captured solely from the main task perturbation.

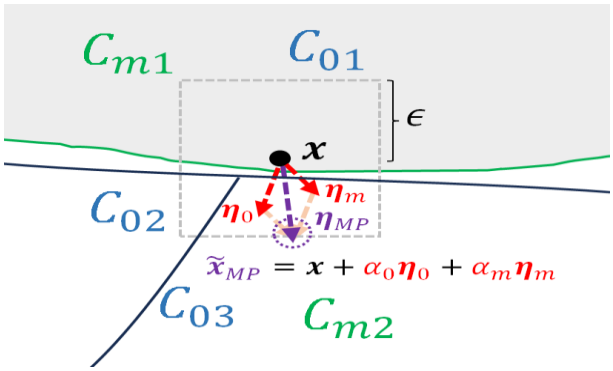


Figure 3. Illustration of the mechanism for generating mixed perturbation (MP).

Consequently, the adversarial examples generated by adding mixed perturbations to original data are expected to uncover pockets that are concealed in a broader range of directions across the manifold. In the following subsection, we apply this approach to a defensive scenario: adversarial training using mixed perturbation.

2.2. Adversarial Training of Victim models

2.2.1. Adversarial Training using Mixed Perturbation

We now try to leverage the adversarial examples obtained by the mixed perturbation as training samples for adversarial training and multi-task learning. Let us first review the objective function of adversarial training:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left[\max_{\|\boldsymbol{\eta}\|_p \le \epsilon} l(\boldsymbol{\theta}, \boldsymbol{x}+\boldsymbol{\eta}, y)\right], \tag{3}$$

and that of multi-task learning:

$$L = \sum_{n=1}^{N}\sum_{m=0}^{M} l(\boldsymbol{\theta}, \boldsymbol{x}_n, y_n^m)\,. \tag{4}$$

Where N is the total amount of data. By combining these two training algorithms, we can obtain the objective function for adversarial training with multi-task model, which is formulated as follows:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x},y^0\dots y^m)\sim\mathcal{D}}\left[\sum_{m=0}^{M} l(\boldsymbol{\theta}, \boldsymbol{x}+\boldsymbol{\eta}_0, y^m)\right]. \tag{5}$$

The objective function for adversarial training incorporating mixed perturbations in a multi-task model can be simply defined by substituting $\boldsymbol{\eta}$ from in equation (5) with $\boldsymbol{\eta}_{MP}$, while maintaining the equation (2) for mixed perturbation, which can be written as

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x},y^0\dots y^m)\sim\mathcal{D}}\left[\sum_{m=0}^{M} l(\boldsymbol{\theta}, \boldsymbol{x}+\boldsymbol{\eta}_{MP}, y^m)\right]. \tag{6}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 October 2024 doi:10.20944/preprints202410.0073.v1

Consequently, the objective function (6) aims to achieve the generalization effects of multi-task learning in addition to the improved robustness gained from adversarial training using mixed perturbations. The overall process of adversarial training using mixed perturbation is illustrated in Figure 4.

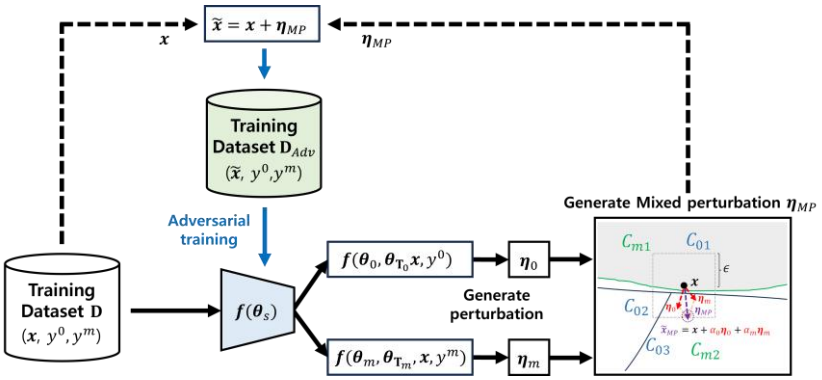


Figure 4. Illustration of the process of adversarial training using mixed perturbation.

2.2.2. Adversarial Training using Multi-task Attack

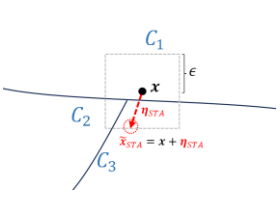
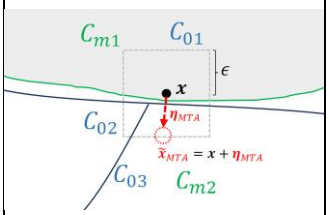
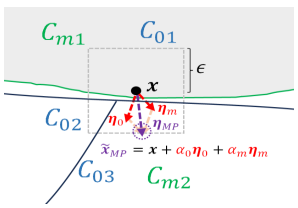
We compare the performance of our proposed method with existing research; we employ the multi-task attack (MTA) used in Mao's study [40], and utilized them for adversarial training, as it utilizes auxiliary tasks for generation of adversarial examples. In the context of a multi-task attack, the perturbations are designed to deceive the main task and all other tasks targeted by the attacker. The formulation for the multi-task perturbation $\boldsymbol{\eta}_{MTA}$ can be written as follows:

$$\boldsymbol{\eta}_{MTA} = \underset{\boldsymbol{\eta}}{\operatorname{argmax}} \sum_{m}^{M} l(\boldsymbol{\theta}, \boldsymbol{x} + \boldsymbol{\eta}, y^m), \text{ s.t. } \|\boldsymbol{\eta}\|_p \leq \epsilon. \tag{7}$$

The objective function for adversarial training using multi-task attack samples also can be simply defined by substituting the perturbation $\boldsymbol{\eta}$ with $\boldsymbol{\eta}_{MTA}$ in equation (5).

When contrasting the multi-task attack with the conventional single task attack (STA) [12] which relies solely on the main task for perturbation generation, it is observed that the multi-task attack samples generally exhibit a bit higher ability to mislead auxiliary tasks. However, they exhibit a reduced effectiveness in deceiving the main task [46]. Table 1 shows a comparison of the three different types of perturbations: STA, MTA, and MP. We compare the performance of the models trained with these three different types of perturbations in the Result section.

Table 1. Comparison of the three types of perturbations.

Type of Perturbation	STA [12]	MTA [40]	MP
Illustration			
Perturbation formula	$\boldsymbol{\eta}_{STA} = \underset{\boldsymbol{\eta}}{\operatorname{argmax}}\, l(\boldsymbol{\theta}_0, \boldsymbol{x} + \boldsymbol{\eta}, y^0)$	$\boldsymbol{\eta}_{MTA} = \underset{\boldsymbol{\eta}}{\operatorname{argmax}} \sum_{m}^{M} l(\boldsymbol{\theta}, \boldsymbol{x} + \boldsymbol{\eta}, y^m)$	$\boldsymbol{\eta}_m = \underset{\boldsymbol{\eta}}{\operatorname{argmax}}\, l(\boldsymbol{\theta}_m, \boldsymbol{x} + \boldsymbol{\eta}, y^m)$ $\boldsymbol{\eta}_{MP} = \sum_{m=0}^{M} \alpha_m \boldsymbol{\eta}_m, \quad \sum_{m=0}^{M} \alpha_m = 1$

 doi:10.20944/preprints202410.0073.v1

3. Results

Since this paper extends our previous work [43], we continued the experiments by using "advertorch" and utilized the same datasets, models, attack methods (including hyperparameters for attack samples generation) and other settings. Thus, we describe a brief summary regarding experimental settings in subsection 3.1.~3.2. The following subsection 3.3 explains metrics for evaluating and analyzing results. In subsection 3.4, we first compare the experimental results of adversarial training using the proposed mixed perturbation with the results from our previous work. Subsequently, subsection 3.5. compares the results using the three different perturbation generation methods: STA, MTA, and MP.

3.1. Datasets and Victim Models

In this study, we used five benchmark datasets: MNIST [47] and CIFAR-10 [48] are the most widely used datasets in computer vision, SVHN (Street View House Numbers) [49] is a real-world dataset for digit recognition problem, GTSRB (German Traffic Sign Recognition Benchmark) [50] is the one for evaluating the practical application, and Tiny ImageNet [51] is employed to evaluate applicability on larger datasets. For each dataset, we utilized two auxiliary tasks generated by methods proposed in our previous work [43]. The detailed components of each dataset are presented in Table 2.

Table 2 also presents the networks used for victim models used in experiments for each dataset. When training the victim models for all datasets, we equally used the SGD optimizer with a learning rate of 0.1, proceeded an average of 200 epochs training, and select the optimal checkpoint which shows the highest adversarial accuracy for the main task. Each dataset utilizes two auxiliary tasks, resulting in a total of three multi-task combinations, and each combination was adversarially trained using the three perturbation generation methods experimented in this study. The abbreviations of the three types of victim models are shown in Figure 5.

Table 2. Data composition and victim models of five benchmark datasets.

Dataset	Train	Test	Class	Channel	Size	Auxiliary task 1	Auxiliary task 2	Network
MNIST	60,000	10,000	10	Gray	28x28	odd, even	prime, composite	LeNet
SVHN	73,257	26,032	10	RGB	32x32	odd, even	prime, composite	ResNet-18
GTSRB	39,209	12,630	43	RGB	112x112	circle, polygon	character, symbol	AlexNet
CIFAR10	50,000	10,000	10	RGB	32x32	animal, vehicle	sky, ground, water	WRN-32
Tiny ImageNet	100,000	10,000	200	RGB	64x64	natural, artefact	animal, machine, others	ResNet-18

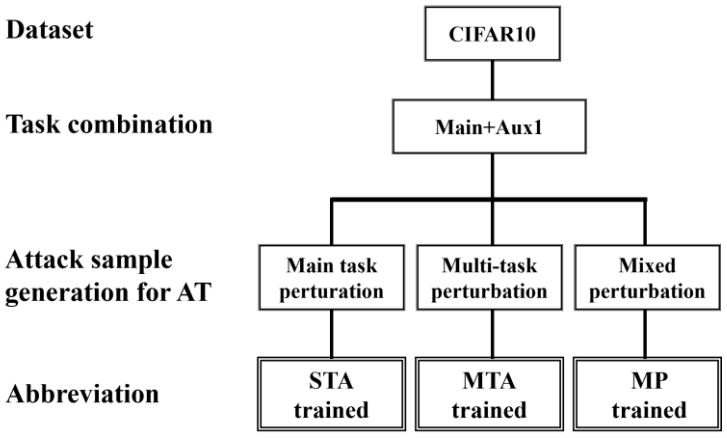


Figure 5. Three different types of victim models used in the experiments.

3.2. Attack Methods

In the experiments utilizing mixed perturbation, we assume a gray-box setting that is close to a white-box scenario, where the attacker has access to nearly all information about the victim model

 doi:10.20944/preprints202410.0073.v1

but lacks knowledge of the auxiliary task information. We selected projected gradient descent (PGD) [12] untargeted/targeted, and the Carlini & Wagner (CW) [13] untargeted attacks as the attack methods to evaluate the adversarial training model utilizing mixed perturbation. In the experiments comparing the three types of perturbations, we employed PGD untargeted and targeted attacks. The adversarial examples for adversarial training were generated using PGD untargeted attack, and the performance of the adversarially trained victim models against the same attack was also evaluated.

3.3. Metrics

The metrics for evaluating the performance of adversarially trained models consist of two key measures: accuracy on the unperturbed original data (clean accuracy) and accuracy on adversarial examples (adversarial accuracy). Note that the accuracies measured in this paper indicate recognition performance for the main task.

When applying the same defensive algorithm across multiple datasets, the degree of robustness may differ for each dataset (including factors such as whether there is improvement or degradation, variations in the extent of improvement, and variations caused by domain-specific characteristics [52]), and the types of defensible attacks may also vary. Therefore, in this study, we conducted an analysis based on the characteristics of the data. The value of data characteristics-based analysis regarding adversarial robustness has been demonstrated in the preceding study [53].

When analyzing the experimental results, we utilized the distributional properties of the datasets: inter-class variance. For calculating inter-class variance, we use a method that measures the average distances between data points across classes, since this is less sensitive to outliers compared to measuring distances from the mean vector, allowing for a more accurate representation of the distribution characteristics. The equation for inter-class variance can be written as

$$Inter-class\ variance = \frac{1}{K_0(K_0-1)} \sum_{k=1}^{K_0} \sum_{j=1(j\neq k)}^{K_0} mean\left(\mathbf{D}(\boldsymbol{X}_k, \boldsymbol{X}_j)\right). \quad (8)$$

Here, $\boldsymbol{X}_k$ and $\boldsymbol{X}_j$ denotes the set of data points belonging to each class, and $\mathbf{D}$ is the distance matrix that calculates the pairwise distances between data points in $\boldsymbol{X}_k$ and $\boldsymbol{X}_j$. Since these two data characteristics can be calculated before conducting experiments, we present the inter-class variances calculated for five benchmark datasets in Table 3.

Table 3. Average inter-class variance of five benchmark datasets.

Dataset	Inter-class variance
MNIST	10.3211
SVHN	14.4643
GTSRB	67.2574
CIFAR10	19.0552
Tiny ImageNet	19.2206

In addition, the angle between two vectors was measured to evaluate the directional diversity of mixed perturbation. We measured the angle between the original data $\boldsymbol{x}$ and the perturbations $\boldsymbol{\eta}_0$, $\boldsymbol{\eta}_m$, and $\boldsymbol{\eta}_0 + \boldsymbol{\eta}_m$, which are represented as $\measuredangle(\boldsymbol{x}, \boldsymbol{\eta}_0)$, $\measuredangle(\boldsymbol{x}, \boldsymbol{\eta}_m)$, and $\measuredangle(\boldsymbol{x}, \boldsymbol{\eta}_0 + \boldsymbol{\eta}_m)$. To evaluate how the directions of the perturbations generated by each task differ, we calculated the angle between the perturbation of the main task and the perturbation of the auxiliary tasks, which is denoted as $\measuredangle(\boldsymbol{\eta}_0, \boldsymbol{\eta}_m)$.

3.4. Experimental Results and Analysis

3.4.1. Experimental Results with Mixed Perturbation

The comparison of performance of the adversarially trained models using the proposed mixed perturbation is presented in Table 4. We first compare the performance of two adversarial training

 doi:10.20944/preprints202410.0073.v1

models: STA trained, and MP trained. The values highlighted in bold and underlined represent cases where the MP trained model outperforms the STA trained model. The two large columns on the left represent the datasets and task combinations for multi-task learning, respectively. The top row of the remaining four large columns indicates the type of input samples during the inference phase. For each type of input sample, the table allows for a comparison of the performance between models trained on conventional single task attack samples (STA trained) and those trained on attack samples generated with mixed perturbation (MP trained). For example, the value 99.09 at the top-left of Table 5 shows the clean accuracy of a MNIST model trained on the main task and the first auxiliary task, using adversarial samples generated by the conventional single-task attack.

The model trained using the proposed mixed perturbation shows enhanced robustness across the GTSRB, CIFAR10, and Tiny ImageNet datasets. Especially in case of GTSRB dataset, while the MP trained models exhibit lower robustness against PGD targeted and CW attacks compared to STA trained models, it demonstrates a notable improvement in adversarial accuracies against PGD untargeted attacks. When analyzing the three task combinations, the robustness against PGD untargeted attacks has increased by an average of 5.24%, peaking at an increase of 8.1%. Additionally, it's important to highlight that the clean accuracies have also improved, with an average increase of 3.66% and a maximum rise of 4.98%.

In case of CIFAR10, clean accuracies are decreased, which was somewhat compromised (trade-off), yet the adversarial accuracies against all types of attacks exhibit significant improvement. Notably, the robustness against PGD untargeted attacks rose by an average of 1.69%, while for PGD targeted attacks, the increase was around 1.91%. Additionally, the model exhibited an improvement of 0.61% against CW attacks.

Table 4. Comparison of classification accuracies of main task in STA and MP trained models.

Dataset	Task Combination	Clean samples		PGD targeted		PGD untargeted		CW untargeted	
		STA trained	MP trained	STA trained	MP trained	STA trained	MP trained	STA trained	MP trained
MNIST	Main+Aux1	99.09	99.05	98.03	**98.10**	94.46	94.35	82.96	**83.84**
	Main+Aux2	98.95	**99.08**	98.05	97.96	94.82	94.24	87.16	81.90
	Main+Aux1+Aux2	98.99	98.97	98.13	97.85	95.13	93.16	86.84	73.04
SVHN	Main+Aux1	90.42	90.31	69.99	**70.68**	54.26	52.32	56.14	53.64
	Main+Aux2	89.56	88.53	68.72	**69.22**	53.20	51.48	54.23	50.68
	Main+Aux1+Aux2	90.05	**91.05**	70.82	70.36	54.18	53.33	54.34	53.98
GTSRB	Main+Aux1	90.05	**92.90**	76.28	75.23	58.72	**63.48**	57.44	54.03
	Main+Aux2	88.87	**92.01**	75.12	73.90	59.20	**62.07**	56.65	54.69
	Main+Aux1+Aux2	89.83	**94.81**	76.18	73.64	59.94	**68.04**	56.48	51.53
CIFAR10	Main+Aux1	84.62	83.34	68.63	**70.44**	45.23	**48.12**	39.78	**41.56**
	Main+Aux2	84.79	84.18	67.83	**69.85**	45.77	**47.29**	40.68	**42.18**
	Main+Aux1+Aux2	84.94	82.57	67.85	**69.77**	46.68	**47.33**	42.13	40.68
Tiny ImageNet	Main+Aux1	17.71	**22.36**	18.21	**21.75**	6.16	5.80	7.83	**8.89**
	Main+Aux2	17.88	**21.43**	18.04	**20.33**	5.51	5.15	7.57	**8.10**
	Main+Aux1+Aux2	18.39	**22.52**	18.31	**21.20**	6.03	5.29	8.16	**8.36**

In the experiments on Tiny ImageNet, the results exhibit a notable increase in clean accuracies, averaging 4.11%. Furthermore, there was an average improvement in robustness of 2.91% against PGD targeted attacks and 0.60% against CW attacks. However, there was no increase in robustness against PGD untargeted attacks, which is unexpected since adversarial training is generally known as the ability to defend attack methods used in training phase. This anomaly requires further

 doi:10.20944/preprints202410.0073.v1

investigation, which will be addressed in our future research involving additional experiments and analyses aimed at enhancing the baseline performance of the Tiny ImageNet model.

For the MNIST and SVHN, the application of the proposed mixed perturbation did not yield a beneficial effect on the adversarial robustness of the models. Rather, there was an average decrease in robustness of 1.32% for SVHN and 2.76% for MNIST (excluding CW attacks), accompanied by a minor decrease in clean accuracy. Moreover, for the two-task combinations involving MNIST, models trained using mixed perturbation showed notable vulnerability to CW attacks. The underlying factors contributing to the lack of effectiveness of mixed perturbation on the MNIST and SVHN datasets will be explored in the following subsection.

3.4.2. Analysis on the Effect of Mixed Perturbation

We speculate that the rare instances where the MP trained models outperform the STA trained models in the SVHN and MNIST experiments may be attributed to the inter-class variance shown in Table 3. Here, we analyze the relation between distributional characteristic of the data and directional changes of the perturbation. The inter-class variance for MNIST and SVHN is relatively low among the five datasets, suggesting that the small distances between classes could make the model more sensitive to changes in the direction of perturbations. This means that a change in the direction of the perturbation due to the auxiliary task may change the directionality of the main task perturbation, which may lead to the resulting mixed perturbation being included within the decision region of an unintended class. Therefore, in such cases, a more refined approach is needed.

As one way to address this issue, we can consider assigning a larger main task perturbation weight $\boldsymbol{\alpha}_0$ to prevent unintended perturbations caused by the loss of primary directionality. Figures 6 and 7 represent graphs measuring the model's performance based on the variation of $\boldsymbol{\alpha}_0$ in models using the first auxiliary task of MNIST and the second auxiliary task of SVHN, respectively.

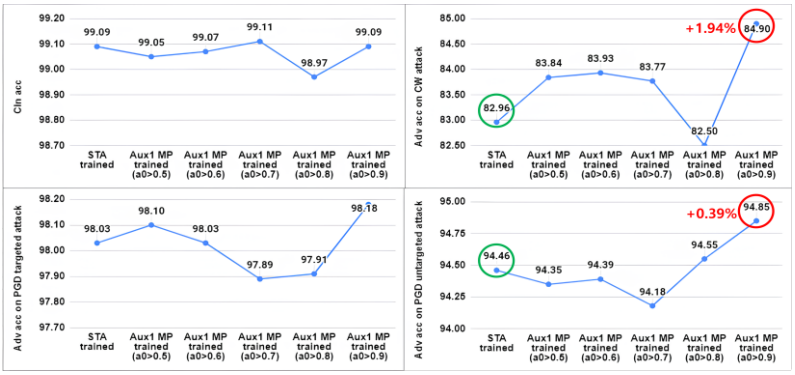


Figure 6. The change in accuracy of the MNIST MP trained model (using the first auxiliary task) with respect to variations in the value of $\boldsymbol{\alpha}_0$.

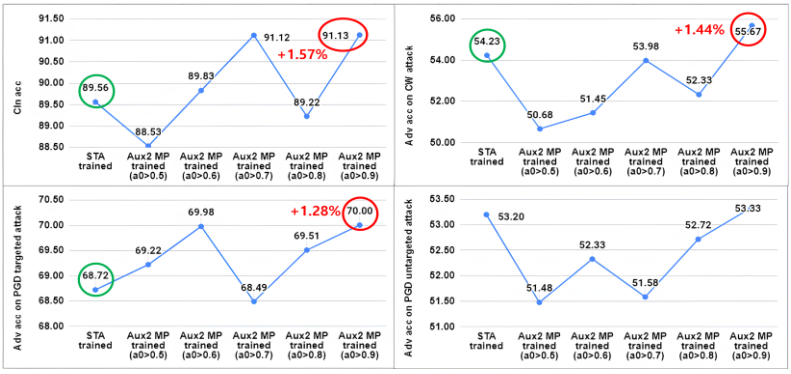


Figure 7. The change in accuracy of the SVHN MP trained model (using the second auxiliary task) with respect to variations in the value of $\boldsymbol{\alpha}_0$.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 October 2024 doi:10.20944/preprints202410.0073.v1

In each Figures 6 and 7, the green circle on the left represents the adversarial accuracy of the STA trained model, with the adjacent value corresponding to ($\boldsymbol{\alpha}_0 > 0.5$) is the adversarial accuracy shown in Table 4. The values indicated by the red circle on the right shows that a non-trivial performance improvement was achieved by assigning a larger $\boldsymbol{\alpha}_0$. Note that we conducted experiments on the remaining tasks of MNIST and SVHN; however, no positive effects like those observed in Figures 6 and 7 were achieved. These experiments revealed that the proposed mixed perturbation exhibits varying effectiveness depending on the auxiliary task used.

As another potential approach, we can consider setting the attack budget, represented by the epsilon ϵ, based on inter-class variance during the generation of perturbations, which will be one of our future works.

On the other hand, GTSRB gains the highest performance improvement due to mixed perturbation, which can also be inferred from its significantly high inter-class variance. This means that GTSRB can be expected to have a larger number of pockets between the classes. Consequently, the model may have been less sensitive to changes in perturbation directions caused by the mixed perturbation, allowing it to effectively locate valid pockets.

We analyze the impact of mixed perturbation on adversarial training by measuring the directions of perturbations. Table 5 presents the results of measuring the standard deviation of the angle between the perturbations, $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$, generated using the main task and the first auxiliary task for each dataset. We can see the changes in the standard deviation of the angle between $\boldsymbol{x}$ and the perturbation when $\boldsymbol{\eta}_1$ is added to $\boldsymbol{\eta}_0$, as well as the standard deviation of the angle between $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$.

Table 5. Standard deviation of the angles for each dataset.

Dataset	$\measuredangle(\boldsymbol{\eta}_0, \boldsymbol{\eta}_1)$	$\measuredangle(\boldsymbol{x}, \boldsymbol{\eta}_0)$	$\measuredangle(\boldsymbol{x}, \boldsymbol{\eta}_0 + \boldsymbol{\eta}_1)$	$\measuredangle(\boldsymbol{x}, \boldsymbol{\eta}_0 + \boldsymbol{\eta}_1) - \measuredangle(\boldsymbol{x}, \boldsymbol{\eta}_0)$
MNIST	8.88	3.67	4.04	0.37
SVHN	13.63	1.93	2.23	0.30
GTSRB	3.72	4.39	6.45	2.06
CIFAR10	5.15	3.01	4.08	1.07
Tiny ImageNet	5.07	2.88	3.81	0.93

We first assess the angle between original data point $\boldsymbol{x}$ and perturbations: $\measuredangle(\boldsymbol{x}, \boldsymbol{\eta}_0)$ and $\measuredangle(\boldsymbol{x}, \boldsymbol{\eta}_0 + \boldsymbol{\eta}_1)$**.** For all datasets, the addition of $\boldsymbol{\eta}_1$ to $\boldsymbol{\eta}_0$ led to an increase in standard deviation of the perturbation directions. Specifically, GTSRB increased by 2.06, CIFAR10 by 1.07, Tiny ImageNet by 0.93, SVHN by 0.30, and MNIST by 0.37. However, the increases in standard deviation for SVHN and MNIST were notably lower than those of the other datasets. This indicates that the intended diversification of perturbation directions through mixed perturbation was not effectively performed for these two datasets.

When analyzing the angle $\measuredangle(\boldsymbol{\eta}_0, \boldsymbol{\eta}_1)$, we observe that the standard deviations for MNIST and SVHN are considerably higher compared to those for GTSRB, CIFAR10, and Tiny ImageNet. This implies that the addition of $\boldsymbol{\eta}_1$ to $\boldsymbol{\eta}_0$ introduces more variability in direction. Although this could potentially inject more diversity in perturbation directions, it also increases the likelihood of unnecessary directional changes. In other words, while the objective of the proposed mixed perturbation is to improve adversarial accuracy for the main task by capturing unidentified pockets, the directional change need not to be substantial enough to alter the main directionality of $\boldsymbol{\eta}_0$.

3.4.3. Comparison of Three Perturbation Generation Methods

The proposed mixed perturbation generates distinct perturbations for the main task and auxiliary tasks, which are subsequently combined by random weighted summation to introduce more directional diversity into the adversarial examples. On the other hand, when the main and auxiliary tasks are provided, multi-task attacks [40,54] can also be considered as a method for generating adversarial examples for adversarial training. However, when operating under the same

 doi:10.20944/preprints202410.0073.v1

attack budget, adversarial examples generated by multi-task attacks tend to have weaker strength in deceiving the main task compared to single task attacks. When utilizing multi-task attack samples for adversarial training, the objective is to improve adversarial robustness for all tasks. Consequently, we expected that the proposed mixed perturbation aligns better with the objective of this study, as it focuses on improving the adversarial accuracy of the main task.

We compared the performance of three different adversarially trained models against PGD targeted and untargeted attacks in Table 6. The values corresponding to the STA trained and MP trained are identical to those shown in Table 4. The values highlighted in bold and underlined represent the cases showing the highest adversarial accuracies among three adversarially trained models. For PGD targeted attacks, the adversarial training method that exhibits the highest adversarial accuracies varied across datasets. Since we primarily use adversarial examples generated through PGD untargeted attack for adversarial training, it is challenging to observe a consistent tendency in which approach performs best against PGD targeted attack.

Table 6. Comparison of adversarial accuracies of adversarially trained models trained using three different perturbations.

Dataset	Task Combination	PGD targeted			PGD untargeted		
		STA trained	MTA trained	MP trained	STA trained	MTA trained	MP trained
MNIST	Main+Aux1	98.03	97.85	**98.10**	**94.46**	93.77	94.35
	Main+Aux2	**98.05**	97.83	97.96	**94.82**	93.90	94.24
	Main+Aux1+Aux2	**98.13**	97.91	97.85	**95.13**	93.55	93.16
SVHN	Main+Aux1	69.99	70.33	**70.68**	**54.26**	51.78	52.32
	Main+Aux2	68.72	**69.31**	69.22	**53.20**	51.16	51.48
	Main+Aux1+Aux2	70.82	**71.20**	70.36	**54.18**	51.50	53.33
GTSRB	Main+Aux1	**76.28**	76.00	75.23	58.72	59.84	**63.48**
	Main+Aux2	75.12	**75.27**	73.90	59.20	59.91	**62.07**
	Main+Aux1+Aux2	**76.18**	74.60	73.64	59.94	59.44	**68.04**
CIFAR10	Main+Aux1	67.83	68.69	**70.44**	45.23	45.82	**48.12**
	Main+Aux2	67.85	68.89	**69.85**	45.77	45.67	**47.29**
	Main+Aux1+Aux2	68.64	68.94	**69.77**	46.68	46.05	**47.33**
Tiny ImageNet	Main+Aux1	18.21	**23.64**	21.75	6.16	**6.55**	5.80
	Main+Aux2	18.04	**22.21**	20.33	5.51	**7.45**	5.15
	Main+Aux1+Aux2	18.30	18.25	**21.20**	6.03	**6.30**	5.29

Against PGD untargeted attacks, the SVHN and MNIST datasets show no robustness gain in MTA and MP trained models compared to STA trained models, suggesting that training with directionally diversified adversarial examples was ineffective. However, for the GTSRB and CIFAR10 datasets, training with the proposed mixed perturbation improved the adversarial accuracies more effectively than MTA or STA training. Especially in case of CIFAR10, all task combinations of MP trained models show improved adversarial accuracies against both attacks. In the case of Tiny ImageNet, the MTA-trained model was more robust against both attack methods. While the proposed method showed partial effectiveness, no clear overall tendency was identified.

4. Discussion

As a result, the overall experimental results varied depending on the dataset, the task combination of the multi-task model, and the type of attack. Nonetheless, in several cases, significant improvements in adversarial robustness were observed, and the cases where both clean accuracy and adversarial accuracy were improved are particularly noteworthy.

 doi:10.20944/preprints202410.0073.v1

Although we utilized three different analysis metrics to interpret the experimental results, but additional analysis is required for further investigation of the properties of adversarial robustness. The characteristics of the data varies widely [53], and a more comprehensive analysis from various perspectives is necessary.

The proposed adversarial training method using mixed perturbations has the limitation of increased computational time, as illustrated in Table 7, due to the need for generating two or more perturbations within each batch loop. Notably, the time required for perturbation generation varies based on the type of auxiliary task and dataset used. This highlights the importance of exploring ways to reduce computational overhead. Additionally, since mixed perturbation does not consistently outperform existing adversarial training techniques across all datasets and attack types, it is necessary to refine and enhance the methodology.

Table 7. Comparison of the time taken for one epoch of training based on the number of auxiliary tasks required for perturbation generation (second).

Dataset	Model	Main task	One Aux task	Two Aux tasks
MNIST	LeNet	47.89	74.55	76.12
SVHN	ResNet18	1352.04	2054.92	2248.94
GTSRB	AlexNet	502.66	533.75	659.65
CIFAR10	WRN-32	5032.58	7818.13	11393.31
Tiny ImageNet	ResNet18	1290.05	2148.45	2795.09

5. Conclusions

In this paper, we proposed a perturbation generation method to identify diverse pockets exist in the data manifold for better adversarial training. The proposed mixed perturbation leverages auxiliary tasks which are correlated with the main task, enables the generation of perturbations with directionality for defensive objective. To increase the diversity of the perturbations generated, we perform random weighted summation under the condition that the directionality of the main task perturbation is preserved.

Through the experiments on five datasets, we confirmed that our proposed method can improve the adversarial robustness of the main task. In addition, we analyzed the experimental results based on the characteristics of the data and demonstrated the effectiveness of the proposed mixed perturbation.

Our proposed MP method in adversarial training is expected to be particularly effective in datasets with high inter-class variance. Furthermore, when combined with methods that train the model in a way that makes the distinction between classes more explicit, it is expected to gain significant improvement in adversarial robustness.

Our future work will aim to verify and analyze the root causes and properties of adversarial vulnerabilities, with the goal of applying it to more practical applications such as medical analysis systems.

Author Contributions: Conceptualization, Changhun Hyun and Hyeyoung Park; Data curation, Changhun Hyun; Formal analysis, Changhun Hyun and Hyeyoung Park; Funding acquisition, Hyeyoung Park; Investigation, Changhun Hyun; Methodology, Changhun Hyun and Hyeyoung Park; Project administration, Changhun Hyun; Resources, Hyeyoung Park; Software, Changhun Hyun; Supervision, Hyeyoung Park; Validation, Changhun Hyun; Visualization, Changhun Hyun; Writing – original draft, Changhun Hyun; Writing – review & editing, Changhun Hyun and Hyeyoung Park.

Funding: The research was funded by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2021-II212068, Artificial Intelligence Innovation Hub).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study is provided with manually generated auxiliary tasks, which are available at https://github.com/ChanghunHyun/Self-defined_MTL.

Acknowledgments: This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2021-II212068, Artificial Intelligence Innovation Hub).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sarker, I. H. Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, **2021**, *2(6)*, 420.
2. Gambín, Á. F., Yazidi, A., Vasilakos, A., Haugerud, H., & Djenouri, Y. Deepfakes: Current and future trends. *Artificial Intelligence Review*, **2024**, *57(3)*, 64.
3. Marchal, N., Xu, R., Elasmar, R., Gabriel, I., Goldberg, B., & Isaac, W. Generative AI misuse: A taxonomy of tactics and insights from real-world data. *arXiv* **2024**, arXiv:2406.13843.
4. Anderljung, M., & Hazell, J. Protecting society from AI misuse: When are restrictions on capabilities warranted? *arXiv* **2023**, arXiv:2303.09377.
5. Madiega, T. Artificial intelligence act. European Parliament: European Parliamentary Research Service, 2021.
6. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. Intriguing properties of neural networks. In 2nd International Conference on Learning Representations, 2014, January.
7. Nguyen, A., Yosinski, J., & Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 427-436.
8. Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. Unsolved problems in ML safety. *arXiv* **2021**, arXiv:2109.13916.
9. Oprea, A., & Vassilev, A. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations (No. NIST Artificial Intelligence (AI) 100-2 E2023 (Withdrawn)). National Institute of Standards and Technology, 2023.
10. Goodfellow, I. J., Shlens, J., & Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572, 2014.
11. Kurakin, A., Goodfellow, I. J., & Bengio, S. Adversarial examples in the physical world. In Artificial Intelligence Safety and Security, 2018, pp. 99-112. Chapman and Hall/CRC.
12. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations, 2018, February.
13. Carlini, N., & Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the IEEE Symposium on Security and Privacy (SP), 2017, pp. 39-57.
14. Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. Adversarial attacks on medical machine learning. *Science*, **2019**, *363(6433)*, 1287-1289.
15. Angelos, F., Panagiotis, T., Rowan, M., Nicholas, R., Sergey, L., & Yarin, G. Can autonomous vehicles identify, recover from, and adapt to distribution shifts? In Proceedings of the IEEE International Conference on Machine Learning, 2020, pp. 3145-3153, November.
16. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. Practical black-box attacks against machine learning. In Proceedings of the ACM Asia Conference on Computer and Communications Security (ASIACCS), 2017, pp. 506-519, April.
17. Yue, K., Jin, R., Wong, C. W., Baron, D., & Dai, H. Gradient obfuscation gives a false sense of security in federated learning. In 32nd USENIX Security Symposium (USENIX Security 23), 2023, pp. 6381-6398.
18. Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In 2016 IEEE Symposium on Security and Privacy (SP), 2016, pp. 582-597. IEEE.
19. Kuang, H., Liu, H., Wu, Y., Satoh, S. I., & Ji, R. Improving adversarial robustness via information bottleneck distillation. Advances in Neural Information Processing Systems, 2024, 36.
20. Nesti, F., Biondi, A., & Buttazzo, G. Detecting adversarial examples by input transformations, defense perturbations, and voting. *IEEE Transactions on Neural Networks and Learning Systems*, **2021**, *34(3)*, pp. 1329-1341.

21. A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer, "Deflecting adversarial attacks with pixel deflection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8571-8580.
22. Chen, J., Raghuram, J., Choi, J., Wu, X., Liang, Y., & Jha, S. Revisiting adversarial robustness of classifiers with a reject option. In Proceedings of the AAAI Workshop on Adversarial Machine Learning Beyond, 2021, December.
23. Crecchi, F., Melis, M., Sotgiu, A., Bacciu, D., & Biggio, B. FADER: Fast adversarial example rejection. *Neurocomputing*, **2022**, *470*, pp. 257-268.
24. Aldahdooh, A., Hamidouche, W., Fezza, S. A., & Déforges, O. Adversarial example detection for DNN models: A review and experimental comparison. *Artificial Intelligence*, **2022**, Review, pp. 1-60.
25. Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., & Jana, S. Certified robustness to adversarial examples with differential privacy. In Proceedings of the IEEE Symposium on Security and Privacy (SP), 2019, pp. 656-672, May.
26. Wang, H., Zhang, A., Zheng, S., Shi, X., Li, M., & Wang, Z. Removing batch normalization boosts adversarial training. In Proceedings of the International Conference on Machine Learning (ICML), 2022, pp. 23433-23445, June.
27. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., & Jordan, M. Theoretically principled trade-off between robustness and accuracy. In International Conference on Machine Learning, 2019, pp. 7472-7482, May. PMLR.
28. Wong, E., Rice, L., & Kolter, J. Z. Fast is better than free: Revisiting adversarial training. *arXiv* **2020**, arXiv:2001.03994.
29. Athalye, A., Carlini, N., & Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Proceedings of the International Conference on Machine Learning (ICML), 2018, pp. 274-283, July.
30. Tramer, F., Carlini, N., Brendel, W., & Madry, A. On adaptive attacks to adversarial example defenses. In Advances in Neural Information Processing Systems (NeurIPS), 2020, 33, pp. 1633-1645.
31. Liu, Z., Liu, Q., Liu, T., Xu, N., Lin, X., Wang, Y., & Wen, W. Feature distillation: DNN-oriented JPEG compression against adversarial examples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 860-868, June.
32. Xu, W. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv* **2017**, arXiv:1704.01155.
33. Nesti, F., Biondi, A., & Buttazzo, G. Detecting adversarial examples by input transformations, defense perturbations, and voting. *IEEE Transactions on Neural Networks and Learning Systems*, **2021**, *34(3)*, 1329-1341.
34. Chen, Y., Zhang, M., Li, J., Kuang, X., Zhang, X., & Zhang, H. Dynamic and diverse transformations for defending against adversarial examples. In 2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2022, pp. 976-983, December.
35. Klingner, M., Kumar, V. R., Yogamani, S., Bär, A., & Fingscheidt, T. Detecting adversarial perturbations in multi-task perception. In 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2022, pp. 13050-13057, October.
36. Carlini, N., & Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec), 2017, pp. 3-14, November.
37. Pfrommer, S., Anderson, B. G., & Sojoudi, S. Projected randomized smoothing for certified adversarial robustness. *arXiv* **2023**, arXiv:2309.13794.
38. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. Robustness may be at odds with accuracy. In Proceedings of the International Conference on Learning Representations (ICLR), 2019, pp. 1-24.
39. H. Sicong et al., "Interpreting adversarial examples in deep learning: A review," *ACM Computing Surveys*, **2023**.
40. Mao, C., et al. Multitask learning strengthens adversarial robustness. In Proceedings of the European Conference on Computer Vision (ECCV), 2020, pp. 158-174.
41. Huang, T., Menkovski, V., Pei, Y., Wang, Y., & Pechenizkiy, M. Direction-aggregated attack for transferable adversarial examples. *ACM Journal of Emerging Technologies in Computing Systems (JETC)*, **2022**, *18(3)*, 1-22.

 doi:10.20944/preprints202410.0073.v1

42. Li, Z., Yin, B., Yao, T., Guo, J., Ding, S., Chen, S., & Liu, C. Sibling-attack: Rethinking transferable adversarial attacks against face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 24626-24637.
43. Hyun, C., & Park, H. Multi-task learning with self-defined tasks for adversarial robustness of deep networks. *IEEE Access*, **2024**, *12*, pp. 83248–83259.
44. Hassani, H., & Javanmard, A. The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression. *The Annals of Statistics*, **2024**, *52(2)*, 441-465.
45. Wu, B., Chen, J., Cai, D., He, X., & Gu, Q. Do wider neural networks really help adversarial robustness? Advances in Neural Information Processing Systems, 2021, 34, 7054-7067.
46. Lee, S. W., Lee, R., Seo, M. S., Park, J. C., Noh, H. C., Ju, J. G., ... & Choi, D. G. Multi-task learning with task-specific feature filtering in low-data condition. *Electronics*, **2021**, *10(21)*, 2691.
47. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **1998**, *86(11)*, 2278-2324.
48. Krizhevsky, A., & Hinton, G. Learning multiple layers of features from tiny images. 2009.
49. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., & Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011, vol. 2011, no. 5, pp. 7, December.
50. Stallkamp, J., Schlipsing, M., Salmen, J., & Igel, C. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, **2012**, *32*, 323-332.
51. Le, Y., & Yang, X. Tiny ImageNet Visual Recognition Challenge. CS 231N, 2015, vol. 7, no. 7, pp. 3.
52. Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., & Lu, F. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, **2021**, *110*, 107332.
53. Xiong, P., Tegegn, M., Sarin, J. S., Pal, S., & Rubin, J. It Is All About Data: A Survey on the Effects of Data on Adversarial Robustness. *ACM Computing Surveys*, **2024**, 56(7), 1-41.
54. Ghamizi, S., Cordy, M., Papadakis, M., & Le Traon, Y. Adversarial robustness in multi-task learning: Promises and illusions. In Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(1), pp. 697-705, June.
55. Deng, J., Guo, J., Xue, N., & Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, (pp. 4690-4699).