ADVERSARIAL ROBUSTNESS AND GENERALIZATION

by

Duygu Serbes

B.S., Mechanical Engineering, Boğaziçi University, 2019

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering
Boğaziçi University

2023

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my advisor, Assist. Prof. İnci Meliha Baytaş, for her endless support, guidance, and belief in me throughout this journey. Her unwavering commitment to my success, along with her invaluable feedback and encouragement, has been the driving force behind my accomplishments.

I would also like to extend my sincere thanks to the members of the jury for accepting to be a part of my journey and for providing valuable suggestions.

To my friends and family, thank you for being my constant source of love, encouragement, and support. Your unwavering belief in me has given me the strength and determination to push through the toughest of times. I would like to express my gratitude to each and every person who has crossed my path, whether in a small or significant way. I am truly grateful for all the journey in shaping my understanding and perspective of the world.

This work is supported by Boğaziçi University Research Fund under grant number 17004. The numerical calculations reported in this thesis were fully/partially performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources) and TETAM, Telecommunications and Informatics Technologies Research Center at Bogazici University.

ABSTRACT

ADVERSARIAL ROBUSTNESS AND GENERALIZATION

In light of recent discoveries regarding adversarial attacks, the necessity for robust models in deep learning has become increasingly critical. Adversarial training is considered one of the most effective approaches to defending against such attacks. However, a key challenge of it is the trade-off between adversarial robustness and generalization. The generalizability of robust models in adversarial training is affected by the diversity of perturbations, as they can overfit if the model only learns a limited attack pattern. Although stronger attacks can enhance robustness, their use may cause performance drops when classifying natural images. This thesis investigates the factors that affect the success of adversarial training and proposes solutions to mitigate some of these factors by utilizing new attack augmentation and generation methods. In that regard, we propose an adversarial training method that enhances adversarial directions by augmenting them from a one-step attack. The proposed framework is inspired by the feature scattering adversarial training and generates a principal adversarial direction based on the distance of the inter-sample relationships in a perturbed mini-batch. The principal direction is augmented by sampling new adversarial directions in a 45-degree region from it. The proposed method does not necessitate additional backpropagation steps than feature scattering. Experimental results on popular benchmark datasets indicate that the method consistently improves adversarial robustness without sacrificing natural accuracy. Furthermore, in this thesis, we propose integrating generalization-boosting techniques, namely mixup and shift-invariance, into the adversarial training framework. The proposed techniques aim to improve the data representations and robustness of models through convex data augmentation and by making the models invariant to small shifts. The effectiveness of our proposals is evaluated under white-box attacks on benchmark datasets.

ÖZET

HASMANE DAYANIKLILIK VE GENELLEŞME

Derin öğrenme ağlarının hasmane ataklara karşı hassas ve kırılgan olduğuna dair bulgular derin öğrenmede bu ataklara karşı güçlü modellere ihtiyacı ortaya koymuştur. Hasmane eğitim, bu tür saldırılara karşı savunmada en etkili yaklaşımlardan biri olarak kabul edilir. Ancak, dayanıklılık ve genelleme arasında bir denge sağlamak, hasmane eğitimin temel bir zorluğudur. Ayrıca hasmane eğitiminde genelleme yeteneği, atakların belirli örüntülere ezberlemesine sebep olabileceğinden kullanılan atakların çeşitliliğinden etkilenir. Öte yandan daha güçlü saldırılar sağlamlığı arttırsa da, temiz görüntülerin sınıflandırılmasında performans düşüşüne neden olabilir. Bu tezde hasmane eğitiminin başarısını etkileyen faktörler incelenerek bu faktörleri destekleyecek çözümler önerilir. Bu bağlamda atak yönünü birden fazla yönde çoğaltan ve fazladan geri yayılım operasyonu gerektirmeyen bir hasmane eğitim yöntemi önermekteyiz. Bu hasmane eğitim yaklaşımı, öz nitelik saçılma olarak bilinen hasmane eğitimden esinlenilerek, atak yönü eğitim kümesindeki atağa uğramış örnekler arasındaki mesafeye dayanarak belirlenir. Ana hasmane yönden maksimum 45 derece olacak şekilde yeni hasmane yönleri elde edilir. Çeşitli popüler veri kümelerinde yapılan deneysel sonuçlar, bu yöntemin doğal doğruluğu kaybetmeden hasmane dayanıklılık doğruluğunu tutarlı bir şekilde artırdığını göstermektedir. Ayrıca, bu tezde litarütürde genelleşme konusunda başarı sağlamış yöntemler hasmane eğitim yönteminin geliştirilmesi amacıyla incelenmiştir. Bu bağlamda, mixup ve kaydırmaya dayanıklılık yaklaşımlarının önerilen hasmane eğitim yöntemi ile birleştirilmesi önerilmektedir. Önerilerimiz, hasmane eğitimde veri temsilinin kalitesini arttırmayı ve modelleri küçük kaydırmalara karşı duyarsız hale getirerek genelleme performansının yanı sıra hasmane dayanıklılığını da arttırmayı amaçlamaktadır. Önerilerimizin etkinliği beyaz kutu saldırılarına karşı değerlendirilmiştir.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . iii
ABSTRACT . . . iv
ÖZET . . . v
LIST OF FIGURES . . . viii
LIST OF TABLES . . . x
LIST OF SYMBOLS . . . xii
LIST OF ACRONYMS/ABBREVIATIONS . . . xiv
1. INTRODUCTION . . . 1
1.1. Motivation . . . 1
1.2. Contributions and Outline of the Thesis . . . 4
2. BACKGROUND AND RELATED WORK . . . 7
2.1. Adversarial Attacks . . . 7
2.1.1. Fast Gradient Signed Method (FGSM) . . . 8
2.1.2. Projected Gradient Descent (PGD) . . . 9
2.1.3. Carlini and Wagner Attack (CW) . . . 11
2.1.4. AutoAttack . . . 11
2.2. Adversarial Robustness . . . 12
2.2.1. Adversarial Training with Gradient-based Attacks . . . 13
2.2.2. Adversarial Training with Generative-based Attacks . . . 15
2.2.3. Adversarial Training with Mixup . . . 16
2.2.4. Other Adversarial Training Methods . . . 18
2.3. Shift-Invariance Approaches . . . 19
3. METHOD . . . 22
3.1. Robustness with PDA-AT . . . 22
3.1.1. Adversarial Attack Augmentation . . . 26
3.2. Robustness with Shift-Invariance Approaches . . . 29
3.2.1. Low-Pass Filtering in Shift-Invariance . . . 29
3.2.2. Adaptive Polyphase Sampling in Shift-Invariance . . . 31

3.2.3. Robustness with Mixup-based Adversarial Training 33
3.2.3.1. Mixup + Feature Scatter 33
3.2.3.2. Feature Scatter + Mixup 33
3.2.3.3. Mixup in OT Distance 34
3.2.3.4. Mixup in OT Distance and CE Loss 35
4. EXPERIMENTS 36
4.1. Datasets 36
4.2. Deep Neural Networks Architectures 38
4.3. Training Scheme 38
4.4. Results of PDA-AT 39
4.4.1. Baseline Studies 40
4.4.2. Attack Diversity 41
4.4.3. Results against White-Box Attacks 42
4.4.4. Results against Black-Box Attacks 47
4.4.5. Variations of Perturbation Augmentations 49
4.4.6. Visualization of the Loss Landscape 50
4.5. Results of Shift-Invariance Approaches 51
4.5.1. Results of LPF Models 52
4.5.2. Results of APS Models 52
4.6. Results of Mixup-based Approaches 54
5. CONCLUSION 56
REFERENCES 59

LIST OF FIGURES

Figure 2.1. FGSM attack with $\epsilon = 0.02$ on Imagenet images [6] for the model pre-trained ResNet34 [10, 41]. The codebase provided by [40] is used for the generations of visualizations. The top-5 predictions of the model are listed in the plots. 9

Figure 2.2. Allowable perturbation space within the ℓ_2 ball with radius ϵ center at $\mathbf{x}$. 10

Figure 2.3. Toy example of max-pooling operation to see the effect on shift-invariance performance of the CNNs. 21

Figure 3.1. Proposed PDA-AT method. 23

Figure 3.2. Perturbation augmentation in 2-Dimensional space. 27

Figure 3.3. PDA-AT Framework. Robustness for T epochs, D dataset, θ network parameters, ϵ a perturbation budget, τ learning rate, n mini-batch size, α_{max} maximum deviation angle. 28

Figure 3.4. Traditional Feature Scatter [34] Framework. Adversarial Training for T epochs, D dataset, θ network parameters, ϵ a perturbation budget, τ learning rate, n mini-batch size. 31

Figure 3.5. Mixuped Feature Scatter [34] Framework. Adversarial Training for T epochs, D dataset, θ network parameters, ϵ a perturbation budget, τ learning rate, n mini-batch size. 34

Figure 4.1. Diversity of adversarial attack directions in training. The element-wise sum of the elements of the adversarial direction tensor is shown as a change in standard deviations in the top plot. Standard deviations of cosine similarities between random reference direction and adversarially generated direction is illustrated in bottom plot. 41

Figure 4.2. Loss landscapes of proposed and baselines. The visualizations are created based on the codebase provided by [73]. 51

LIST OF TABLES

Table 4.1. Robustness comparison of proposed and baseline models against white-box attacks on CIFAR-10 dataset. The asterisk symbol (*) denotes results from scratch training, while dagger (†) symbols indicate results sourced from the study of Baytas and Deb [47]. All other values of baseline models are the best results reported by the primary studies, and those marked with a hyphen (-) are currently unavailable. 43

Table 4.2. Robustness comparison of proposed and baseline models against white-box attacks on CIFAR-100 dataset. The asterisk symbol (*) denotes results from scratch training, while dagger (†) symbols indicate results sourced from the study of Baytas and Deb [47]. All other values of baseline models are the best results reported by the primary studies, and those marked with a hyphen (-) are currently unavailable. 44

Table 4.3. Robustness comparison of proposed and baseline models against white-box attacks on SVHN dataset. The asterisk symbol (*) denotes results from scratch training, while dagger (†) symbols indicate results sourced from the study of Baytas and Deb [47]. All other values of baseline models are the best results reported by the primary studies, and those marked with a hyphen (-) are currently unavailable. 46

Table 4.4. Accuracy comparison of models for Tiny-Imagenet dataset. All models are trained from scratch. 47

Table 4.5. Accuracy comparison of models for GTSRB dataset. All models are trained from scratch. 47

Table 4.6. Results against black-box attack transferred from natural model on CIFAR-10. 48

Table 4.7. Results against black-box attack transferred from natural model on CIFAR-100. 48

Table 4.8. Results against black-box attack transferred from AT [21] model on CIFAR-10. 49

Table 4.9. Adversarial robustness performance of various perturbation augmentation permutations. 50

Table 4.10. Adversarial accuracy comparison of adversarially trained shift-invariant models by LPS. 53

Table 4.11. Adversarial accuracy comparison of adversarially trained shift-invariant models with APS and anti-aliasing. 54

Table 4.12. Adversarial accuracy comparison of proposed models enhanced with mixup methodology on CIFAR-10 dataset. 55

LIST OF SYMBOLS

C	Cost function of transporting
D	Data distribution
c	Constant
$\mathcal{D}$	Optimal transport distance
f	DNN model prediction function
$\tilde{\mathcal{F}}$	Upsampling map function
Δh	Shifting amount in height
Δw	Shifting amount in width
$\widetilde{J}$	Updated loss function
$\mathcal{L}$	Loss function of DNN model
ℓ	Norm function
n	Mini-batch size
t	Number of attack iterations
S	Allowable input domain
x	Input of DNN model
x'	Adversarial example
$\tilde{x}$	Mixuped input
y	Ground-truth label
$\tilde{y}$	Mixuped output
α	Random angle parameter
δ	Adversarial perturbation
ϵ	Perturbation budget of adversarial attack
η	Step size of adversarial attack
λ	Proportion of mixup data
∇_x	Gradient with respect to x
ϕ	Proportion of the adversarial term
$\prod$	Projection operation

τ	Learning rate
θ	Parameters of DNN

LIST OF ACRONYMS/ABBREVIATIONS

2D	Two Dimensional
AT	Adversarial Training
AA	AutoAttack
AWP	Adversarial Weight Perturbation
APS	Adaptive Polyphase Sampling
CE	Cross-Entropy
CIFAR	Canadian Institute For Advanced Research
CNN	Convolution Neural Networks
CW	Carlini-Wagner
DL	Deep Learning
DNN	Deep Neural Networks
FAB	Fast Adaptive Boundary
FGSM	Fast Gradient Signed Method
FS	Fature Scatter
GAMA	Guided Adversarial Margin Attack
GAN	Genarative Adversarial Network
GPU	Graphical Processing Units
GTSRB	German Traffic Sign Recognition Benchmark
JS	Jensen-Shannon
KL	Kullback-Leibler
LPF	Low-Pass Filtering
OT	Optimal Transport
PDA-AT	Perturbation Direction Augmentation for Adversarial Training with Diverse Attacks
PGD	Projected Gradient Descent
RGB	Red Green Blue
TETAM	Telecommunications and Informatics Technologies Research Center at Bogazici University

TUBITAK	Scientific and Technological Research Council of Turkey
TRUBA	Turkish National e-Science e-Infrastructure
SGD	Stochastic Gradient Descent
SVHN	Street View House Number
ULAKBIM	High Performance and Grid Computing Center

1. INTRODUCTION

Deep learning (DL) has rapidly emerged as a potent tool in artificial intelligence, achieving remarkable results in various complex problems such as image classification [1], speech recognition [2], natural language processing [3], self-driving cars [4] and healthcare [5]. With recent developments, deep neural networks have surpassed human performance in some applications. As a result, numerous industrial companies have been able to use DL approaches to enhance their products and resolve issues that traditional software engineering was unable to handle. The success of deep learning can be attributed to the development of large-scale datasets [6–8], powerful computational resources [9], advanced neural network architectures [3], [10,11] and optimization techniques [12, 13]. Despite the significant advancements achieved in the field, there remains a great deal of work to be accomplished in order to grasp and utilize the potential of deep neural networks (DNNs) fully.

1.1. Motivation

While there have been numerous studies exploring the potential of deep learning methods to tackle complicated tasks, Szegedy et al. [14] made the initial discovery of an intriguing characteristic of neural networks against adversarial examples in 2014. As a result of various studies [14–17], it is found that DNNs are vulnerable to adversarial examples, which are malicious inputs designed to cause a deep neural network to misclassify an input and produce an incorrect output. The empirical evidence indicates that the introduction of small perturbations that are perceptually indistinguishable can result in severe mispredictions by the model.

Upon this discovery, concerns have been raised regarding the reliability and robustness of deep neural networks in the face of adversarial attacks, particularly in safety-critical applications such as medical robotics [18], self-driving cars [19], and finance [20]. Therefore, researchers have shown a significant deal of interest in building

robust models against adversarial attacks. Also, sophisticated attacks emphasize the urgent need for further research and development in enhancing adversarial robustness.

To alleviate the vulnerability of the DNNs against adversarial examples, various adversarial defense techniques have been developed. Among several effective defense approaches, adversarial training (AT) [21–23] became one of the mainstream and proved to be successful. AT aims to mitigate the issue by training the model on adversarial counterparts of natural images, i.e., the samples corrupted with imperceptible perturbations to the humans, yet capable of fooling the DNNs.

Generally, AT is formulated as a robust optimization problem that takes the form of a non-convex non-concave minimax problem [21, 24]. The inner maximization process is used to create attacks, while model parameters are updated based on the outer minimization using the adversarial loss induced by the generated images with the following formula:

$$\min_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim D} \left[\max_{\boldsymbol{\delta}\in S} \mathcal{L}\left(f\left(\mathbf{x}+\boldsymbol{\delta},\boldsymbol{\theta}\right),y\right) \right] \tag{1.1}$$

where $f(\cdot,\boldsymbol{\theta})$ denotes a neural network parameterized by $\boldsymbol{\theta}$, D represents a joint data distribution of the benign image $\mathbf{x}$ with corresponding label y. δ is the adversarial perturbation, S is the set of possible perturbations within an $\boldsymbol{\delta}$-ball, such as the ℓ_∞ norm ball, and the loss function is represented with $\mathcal{L}$, which is generally selected as cross-entropy for classification tasks.

Adversarial attacks to be used in AT can be created using a single-step attack [15] or multi-step attack approach [21]. To achieve state-of-the-art robustness, Madry et al. [21] proposed training of adversarial examples generated with multi-step projected gradient descent (PGD) attack. By the studies conducted, the effectiveness of AT for models depends on the strength and diversity of the adversarial examples [21,25]. With the knowledge of this fact, recent research has been focused on devising new methods to generate more diverse and stronger adversarial attacks with the aim of producing more robust models that can better withstand adversarial manipulation.

Despite being widely recognized as one of the most effective methods for defending against adversarial attacks, adversarial training continues to present a number of challenges and limitations that require careful consideration and resolution. One such challenge is the issue of label leaking [26], which arises when the model exploits the inherent regularities in the adversarial example construction process. As a result, the model may be able to directly determine the correct class label by decoding the additive perturbation that causes trained models to perform better on adversarial examples than on clean examples, without relying on the underlying content of the image. In addition, gradient masking can create a deceptive illusion of robustness by generating obfuscated gradients. Obfuscated gradients refer to a situation where the model cannot be updated with useful gradients, leading to unsuccessful iterative optimization-based methods [27].

Furthermore, recent studies have brought to light that the one-step attack approach of AT is susceptible to catastrophic overfitting as a potential explanation for the lack of diversity in attacks [23,28]. Catastrophic overfitting leads to 0% robust accuracy against multi-step attacks, resulting in a complete erosion of the model's robustness. In addition to catastrophic overfitting, robust overfitting causes a degradation of the robustness of unseen data under the same adversary used during training [29].

Another drawback of AT is the time complexity and resource dependency. In addition to the gradient computation required for network parameter update, inner maximization needs multiple gradient computations when a multi-step attack approach is used to create adversarial images. Therefore, AT can take a longer time [29,30] (e.g., 3 to 30 times) to reach a robust model than natural training, which is trained with natural images. It makes AT more challenging to scale models for large datasets such as ImageNet.

Although standard AT techniques can enhance the robustness of DNNs, they are usually accompanied by a decline in standard accuracy. Therefore, the use of ATs comes with a trade-off between robustness and generalization. Numerous studies

have shown that DNNs trained with ATs can suffer from a significant decrease in generalization performance on unseen natural images, especially in the presence of complex and strong attacks [21] and inherent feature behavior [31] of robust and non-robust models. This decrease in generalization performance can also be attributed to the inherent feature behavior of robust and non-robust models [31]. To address the trade-off between robustness and generalization in ATs, researchers have proposed several strategies, such as generative adversarial training [32], mixup [33], and feature scattering-based approaches [34].

The purpose of this thesis is to alleviate the drawbacks of AT and present potential solutions through the exploration of novel attack augmentation and generation techniques as well as the investigation of well-known generalization methods, namely mixup, and shift-invariance. Additionally, we propose an AT approach that aims to enhance the diversity of adversarial perturbations without requiring additional back-propagation steps while maintaining generalization performance. This is particularly important in safety-critical applications where misclassification can result in severe consequences.

1.2. Contributions and Outline of the Thesis

The purpose of this thesis is to address the limitations of AT in enhancing the robustness of DNNs against adversarial attacks. To fulfill this goal, we have developed several approaches that were meticulously analyzed and investigated. Initially, we have explored the integration of successful generalization methods such as shift-invariance [35, 36] and mixup [33] into AT scheme. AT with mixup is proposed and analyzed to look for expanding input representation through the augmentation of convex interpolation in the training process. Moreover, we experimentally analyze the effect of employing shift-invariant models in AT scheme, intending to leverage the generalization success of shift-invariant models for AT through various experiments. The efficacy of the required architectural modifications into AT is assessed and validated through experimental evaluations.

Furthermore, we have proposed Perturbation Direction Augmentation for Adversarial Training with Diverse Attacks (PDA-AT) method, which is an end-to-end adversarial training scheme that achieves state-of-the-art performance in terms of robustness and generalization. PDA-AT enhances adversarial perturbation diversity by the generation of the attack direction based on the increase in the feature matching distance in the perturbed mini-batch without requiring any additional backpropagation operation. Also, in order to have more attack diversity, it proposed attack augmentation aimed at combating overfitting issues. The key contributions of this thesis are as follows:

- We propose an AT scheme that utilizes the mixup [33] method and data augmentation based on the convex interpolation of input and targets, with the aim of improving the generalization performance of AT. The proposed scheme is evaluated and analyzed under white-box attack settings on the CIFAR-10 dataset.
- We introduce an approach to enhance the generalization of AT by incorporating shift-invariance techniques such as low-pass filtering (LPF) [35] and adaptive polyphase sampling (APS) [36]. The proposed methodology aims to take advantage of the shift-invariance methods, which are able to achieve boosting generalization by making models robust to small shifts and transitions. The proposed training scheme is evaluated and analyzed under white-box attack settings on the CIFAR-10 dataset.
- In contrast to the traditional feature scattering AT technique, our proposed method, PDA-AT, generates perturbations that enhance the optimal transport distance among various perturbed images within the mini-batch rather than the distance between the natural and randomly perturbed correspondences. Based on empirical evidence, we demonstrate that our approach can effectively obtain more diverse adversarial directions for each iteration.
- Our proposed method, PDA-AT, augments adversarial perturbations by generating new directions within a 45-degree region from the principal adversarial direction, which is generated by the gradient of the optimal transport distance between the mini-batch perturbed samples. The augmentation is achieved by

utilizing cosine similarity. This approach aims to improve the network's ability to handle a wider range of adversarial attacks.

- PDA-AT method fulfills the improvement of adversarial robustness in certain circumstances. It is noteworthy that the difference in accuracy between natural and adversarial samples obtained by the Fast Gradient Sign Method (FGSM) is greatly diminished.
- The proposed methodology aims to alleviate the trade-off between robustness and generalization. Through augmentation of the perturbations, the PDA-AT approach provides increased robustness while maintaining the accuracy of the original test samples.
- The experimental evaluation on various benchmark datasets, namely CIFAR-10, CIFAR-100, SVHN, Tiny-Imagenet, and The German Traffic Sign Recognition Benchmark (GTSRB), reveals the notable performance enhancement achieved by the proposed PDA-AT approach with proposed attack direction generation and augmented perturbations, compared to existing state-of-the-art AT techniques reported in the literature.

2. BACKGROUND AND RELATED WORK

Deep neural networks, specifically convolutional neural networks (CNNs) achieve outstanding performance in a multitude of computer vision tasks. Following the initial realization of the adversarial example phenomenon by Szegedy et al. [14], the unprecedented vulnerability of CNNs has garnered considerable attention in recent times. The presence of adversarial examples poses a grave safety risk for the widespread deployment and utilization of deep learning models. Therefore, several methods have been proposed and developed to mitigate the effects of adversarial attacks.

2.1. Adversarial Attacks

Adversarial attacks can be categorized into three groups depending on the level of knowledge the attacker has regarding the targeted model. White-box attacks refer to scenarios where the attacker has complete knowledge of the targeted model, including network architecture, parameters, gradients, training method, and training data [37, 38]. These attacks are deemed the most effective as the attacker has complete access to model training and they are able to generate adversarial examples to manipulate any weakness of DNNs.

On the other hand, gray box attacks involve the attacker having a limited understanding of the targeted model. For instance, they may know the architecture of the model and training data but have no right to access the exact parameters of it [38]. Whereas, the attacker has no knowledge regarding the targeted model under black-box settings [38]. However, they can interact with the model through input/output queries and leverage previous adversarial examples' prediction scores to generate new samples.

To summarize, white-box attacks are the most powerful but require complete knowledge of the targeted model. In contrast, black-box attacks are the most challenging to be successful as an attacker but also the most realistic attack. Gray box

attacks are in between, necessitating some knowledge of the model but not full access to its inner workings. To emphasize the effectiveness of adversarial attacks on neural networks, the most popular white-box attack methods will be reviewed and explained throughout this section.

2.1.1. Fast Gradient Signed Method (FGSM)

To provide effective adversarial training, Goodfellow et al. [15] firstly proposed a method, Fast Gradient Signed Method (FGSM), to compute an adversarial perturbation for a given image in the l_∞ neighbor by exploiting gradient information of the model. FGSM performs only one step update along the direction of the gradient to increase the adversarial loss, $\mathcal{L}(\mathbf{x}, y, \boldsymbol{\theta})$, in the steepest direction [39] as follows:

$$\mathbf{x}^{FGSM} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, y, \boldsymbol{\theta})) \tag{2.1}$$

where $\mathbf{x}$ denotes the natural input to the model, y is the targets associated with $\mathbf{x}$, $\boldsymbol{\theta}$ is the parameters of neural network and ϵ is the adversarial budget to control the maximum allowable magnitude of the perturbation.

Figure 2.1 presents a comparative analysis of natural and adversarial images alongside the perturbations with $\epsilon = 0.02$. The code base provided by [40] is used for the generation of the showcases. The first column displays natural images, while the second column depicts their corresponding adversarial counterparts. Despite the presence of perturbations, the adversarial images are nearly imperceptible to the naked eye, as they appear visually identical to their natural counterparts. The third column of the figure highlights the perturbations applied to the natural images. These perturbations are highly subtle and imperceptible, yet they are sufficient to fool a state-of-the-art deep neural network model. Specifically, the perturbations were utilized to generate adversarial images that can mislead a pre-trained ResNet34 model, provided by PyTorch [10, 41] and trained on the ImageNet dataset [6]. In the fourth row of the figure, the top five predictions of the pre-trained ResNet34 model are listed for each adversar-

ial image. It is noteworthy that the model misclassified all images, despite achieving correct classification with above 90% confidence on the natural images. These results demonstrate the susceptibility of deep neural network models to adversarial attacks and underscore the need for more robust and reliable models in practical applications.

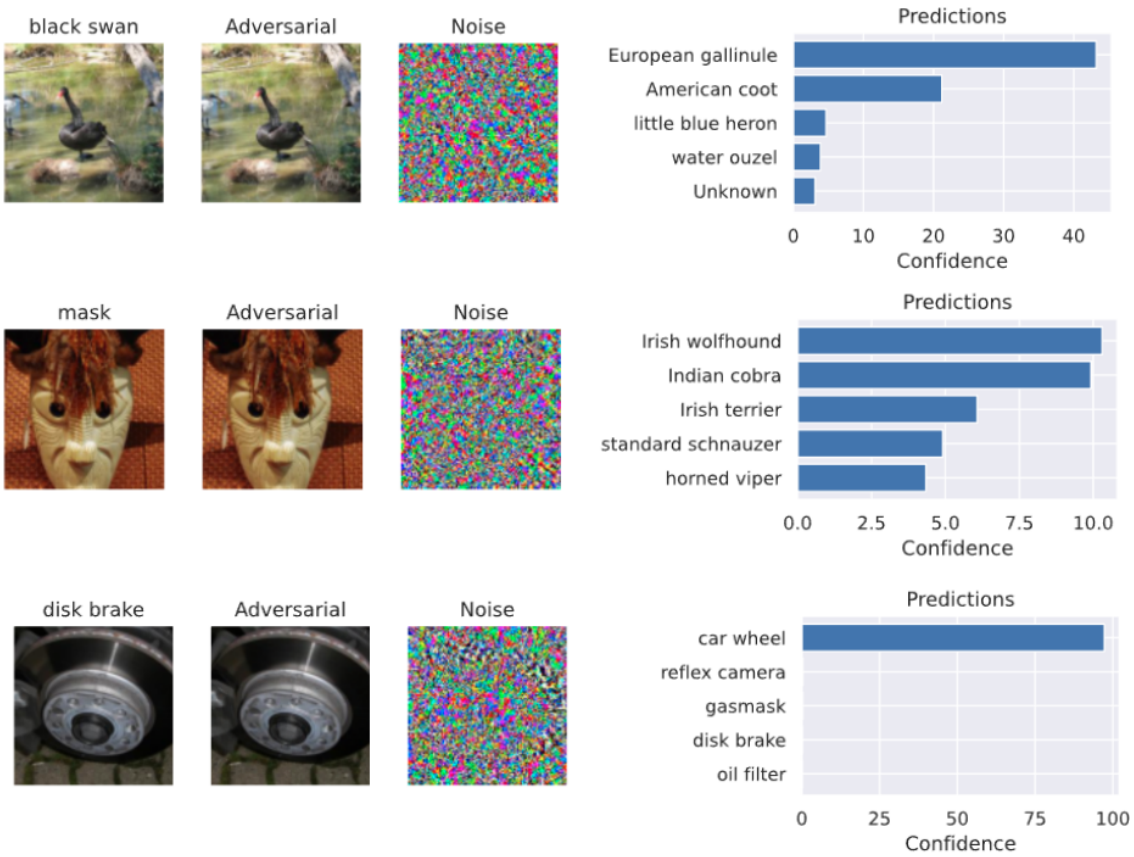


Figure 2.1. FGSM attack with $\epsilon = 0.02$ on Imagenet images [6] for the model pre-trained ResNet34 [10, 41]. The codebase provided by [40] is used for the generations of visualizations. The top-5 predictions of the model are listed in the plots.

2.1.2. Projected Gradient Descent (PGD)

The Projected Gradient Descent (PGD) [21] attack is widely believed to be one of the most powerful attack methods, generating adversarial examples by an iterative

gradient-based adversarial attack. PGD [21] extends FGSM to multiple steps to get stronger attacks. The PGD attack can be formulated mathematically as follows

$$\mathbf{x}^{t+1} = \prod_{\mathbf{x}+S} \left(\mathbf{x}^t + \eta \cdot \operatorname{sign} \left(\nabla_{\mathbf{x}^t} \mathcal{L} \left(\mathbf{x}, y, \boldsymbol{\theta} \right) \right) \right) \tag{2.2}$$

where t denotes the number of attack iterations, whereas η is the step size. $\mathbf{x}^t$ is the created example at the step t and $\mathbf{x} + S$ is the allowable input domain. $\prod_{\mathbf{x}+S}$ is the projection operation to $\mathbf{x} + S$ domain, which ensures that the perturbation remains within the ℓ_∞ ball with radius ϵ centered at $\mathbf{x}$. For instance, Figure 2.2 illustrates that ℓ_2 ball with the radius of ϵ in 2D space.

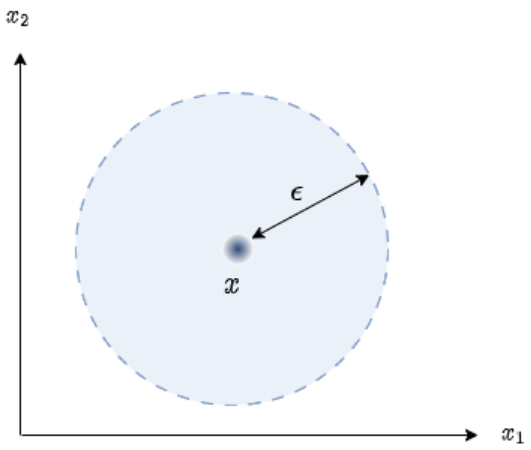


Figure 2.2. Allowable perturbation space within the ℓ_2 ball with radius ϵ center at $\mathbf{x}$.

The cumulative nature of the attack, with the amount η increasing at each step, may cause the total perturbation amount to be exceeded. The projection operator ensures that the total amount of perturbation will remain in the set S. Projection can be executed with clipping operators for ℓ_∞ attacks. The pixel values of the perturbed input might also need clipping to stay in the input domain. It is important to note that although the PGD attack is more successful in generating adversarial examples, it is also very time-consuming compared to one-step attacks due to the high number of backpropagation operations required for multi-iterations to generate the attack [30].

2.1.3. Carlini and Wagner Attack (CW)

Carlini-Wagner(CW) [16] is another widespread white-box attack that aims to generate imperceptible adversarial examples with high success rates against targeted DNNs. The CW attack is formulated as an optimization problem that minimizes a weighted sum of two objective functions as the distance between the original and adversarial example, and a loss function that encourages the adversarial example to be misclassified as the target class, which is expressed as

$$\begin{aligned} &\text{minimize } |\boldsymbol{\delta}|_p + c \cdot f(\boldsymbol{x} + \boldsymbol{\delta}, t) \\ &\text{such that} \quad \boldsymbol{x} + \boldsymbol{\delta} \in [0,1]^n \end{aligned} \tag{2.3}$$

where $\boldsymbol{x}$ denotes the original input, $\boldsymbol{\delta}$ represents the perturbation added to $\boldsymbol{x}$, t is the target class, c is a constant that controls the trade-off between the distance and the loss can be found with binary search. Distance metric, $|\cdot|_p$, is usually specified in terms of ℓ_p norm (e.g., ℓ_1, ℓ_2, ℓ_∞), and $f(\cdot)$ is the loss function.

2.1.4. AutoAttack

Despite the development of various defense mechanisms, evaluating the robustness of DNNs against adversarial attacks remains a challenging task. Current evaluation methods often lack diversity among attacks and rely on limited hyperparameter tuning, leading to unreliable results that may overestimate the robustness of models and provide an inaccurate impression of robustness. To address this issue, Croce and Hein propose the ensemble attack method called AutoAttack [42], which combines four different attacks without the need for hyperparameter tuning.

Specifically, they developed Auto-PGD, a variant of the PGD attack, to overcome its shortcomings, such as fixed step size, ignorance of budget, and agnosticism to optimization trends. Auto-PGD is a budget-aware and step-size-free version of PGD that only requires tuning the number of iterations. The authors then combine Auto-

PGD with two existing complementary attack methods, the white-box FAB attack [43], and the black-box Square attack [44], to construct the AutoAttack ensemble. The attacks are chosen for their diversity in terms of optimization strategies, loss functions, and perturbation norms. According to experimental results on various datasets with various adversarial training methods, the AutoAttack ensemble performs successfully in terms of both reliability and sensitivity.

2.2. Adversarial Robustness

Prior to Szegedy's [14] initial discovery of adversarial examples, DNNs were mainly evaluated based on their performance on natural images. A number of metrics, including recall, accuracy, precision, and F1 score, are used to assess how well DNNs perform on a given task. Several studies [14, 15], [21] showed that DNNs are vulnerable to adversarial attacks, which have a considerable impact on the accuracy and reliability concerns of DNNs. To address this issue, researchers started using adversarial robustness as a new metric to evaluate the performance of DNNs. Adversarial robustness measures the ability of DNNs to resist adversarial attacks and maintain their accuracy against these attacks. This metric is crucial for ensuring the reliability and safety of DNNs in safety-critical applications.

In order to mitigate the vulnerability of DNNs against adversarial examples, several adversarial defense mechanisms [21], [30], [34], [45], [46] have been recently proposed recently. Among these mechanisms, adversarial training has been shown to be a successful approach for improving the robustness of DNNs, with the aid of various adversarial attack generation methods. Adversarial training can employ various types of attacks, including gradient-based attacks, generative-based attacks, and other techniques such as mixup.

2.2.1. Adversarial Training with Gradient-based Attacks

Goodfellow et al. [15] introduce an AT approach with a single-step attack known as FGSM. Their approach involves augmenting the training samples in a mini-batch with adversarial samples generated using and optimizing the network parameters using cross-entropy terms computed separately for natural and adversarial samples as the following objective function

$$\mathcal{L}_{\text{Total}}(\boldsymbol{\theta}, \mathbf{x}, y) = \phi\mathcal{L}(\boldsymbol{\theta}, \mathbf{x}, y) + (1-\phi)\mathcal{L}(\boldsymbol{\theta}, \mathbf{x} + \epsilon\,\text{sign}(\nabla_x\mathcal{L}(\boldsymbol{\theta}, \mathbf{x}, y)), y) \tag{2.4}$$

where $\boldsymbol{x}$ denotes the natural images, and $\mathcal{L}$ is the cross-entropy loss function in adversarial training of the model parametrized with $\boldsymbol{\theta}$. ϕ is a hyperparameter of the proportion of the adversarial term and is selected as 0.5, as suggested by Goodfellow et al. [15].

Despite the regularization effect of this approach, it is not robust against iterative attacks due to label leaking phenomena. Therefore, Madry et al. [21] proposed an iterative attack, named PGD as expressed with the AT scheme formalized in Equation 1.1. The mathematical representation of AT involves a robust optimization problem that adopts the structure of a minimax problem.

In the minimax problem of Equation 1.1, inner maximization seeks to find the optimal adversarial perturbation that maximizes the loss of the neural network when applied to a given data point $\mathbf{x}$ input. On the other hand, the outer minimization problem then aims to find the optimal model parameters that minimize the adversarial loss produced by the inner maximization problem. Such a framework can be seen as a competition between adversarial attack and defense, where the adversary attacks the target model continuously, and the target model develops the ability to withstand all of the attacks.

The PGD adversarial training methodology is widely considered to be a highly

effective defense mechanism against adversarial attacks. However, it has been noted that this technique suffers from suboptimal generalization capabilities as a result of the limited diversity present in PGD attacks [47]. The academic discourse surrounding robustness frequently discusses the trade-off that arises between robustness and generalization when the PGD attack methodology is employed in AT. As a consequence, numerous research efforts have emerged with the aim of exploring and developing more robust and generalized defense mechanisms.

Nevertheless, there are several challenges that must be addressed. For instance, label leaking [26], gradient masking [27], catastrophic overfitting [23], [28], and robust overfitting [29] can lead to an inaccurate sense of robustness. Although PGD attacks have been shown to significantly improve the robustness of deep models when compared to FGSM attacks, their use in large-scale problems is not feasible due to the significant increase in the number of projected gradient descent steps required to solve the inner maximization problem.

The trade-off between adversarial robustness and the computationally expensive nature of PGD-based adversarial training has spurred the creation of AT methods that rely on single-step attacks, employing techniques such as accumulative perturbation and perturbation initialization to increase efficiency and robustness. Shafahi et al. [30] introduced "free" AT based on a single backpropagation step that simultaneously updates both the model weights and the input perturbation, thereby minimizing any additional computational overhead typically associated with AT methods. Despite claims made in earlier studies [15], [21] that FGSM-based AT only provide robustness against small ℓ_∞ perturbation, Wong et al. [23] proposed "fast AT" as a solution, incorporating random initialization before the FGSM step. In their work, the authors also identified the problem of FGSM-based AT, the phenomena of catastrophic overfitting where the model's robustness against PGD attack suddenly and drastically drops to almost 0% while its robustness against FGSM attacks rapidly increases after a few epochs. To mitigate this issue, the authors proposed the use of random initialization in FGSM-based AT [23].

Andriushchenko et al. [28] have argued that the main purpose of random initialization is to reduce the magnitude of the perturbation against catastrophic overfitting and catastrophic overfitting is associated with the behavior of a single filter leading to a nonlinear model with single-layer convolutional networks. Based on that discovery, they proposed a novel regularizer, GradAlign, which maximizes the gradient alignment inside the perturbation in order to generate stronger adversarial examples. Kim et al. [48] have further explored the phenomenon of catastrophic overfitting in FGSM-based AT, attributing it to distortions in decision boundaries and a highly curved loss surface. Based on their observations [48], maximizing gradient alignment can encourage the model to learn a loss surface that is more linear, which in turn reduces the likelihood of having a distorted decision boundary. They have also proposed a simple method that prevents decision boundary distortion by searching for an appropriate step size to create the adversarial example. In their method, they gradually increase the magnitude of perturbations while considering the loss surface of each image. Moreover, Sriramanan et al. [49] proposed a Guided Adversarial Margin Attack (GAMA), utilizing a relaxation term to smooth the loss surface initially and decreasing the weight of the term over further iterations.

The empirical results of the aforementioned research studies suggest that AT based on single-step attacks still do improve robustness against adversarial attacks, their generalization performance on natural samples remains limited. The robustness of these approaches does not demonstrate a significant improvement over the PGD adversarial training method developed by Madry et al. [21]. While AT based on single-step attacks can provide some degree of defense against adversarial attacks, their limited capacity for generalization suggests that more advanced methods such as the PGD method should be used for optimal robustness.

2.2.2. Adversarial Training with Generative-based Attacks

As an alternative to traditional adversarial training based on gradient-based attacks, several publication [25], [32], [47], [50] are proposed AT based on the generative-

based attacks with generator architecture. Wang and Yu [50] proposed GAN-based defense mechanism, training jointly discriminative and generative network as a min-imax game. Jang et al. [25] develop a recursive generator for generating adversarial examples, and they ensure diversity by incorporating an extra loss term. However, the method cannot reach the state-of-the-art results of traditional AT. Furthermore, Jeddi et al. [45] introduce Learn2Perturb which involves learned injecting perturbations into each layer of the DNN to increase uncertainty. This method leverages a learnable standard deviation, which is derived from a zero-mean normal distribution, to achieve this objective.

Despite the fact that generative-based methods offer more potent and diverse attacks for AT, the majority of the current techniques still fail to attain the desired levels of both robustness and natural accuracy. Nevertheless, the ability of these methods to explore various attack scenarios is a valuable asset that can be leveraged to improve the effectiveness of AT. Further research is required to investigate the full potential of generative-based AT methods and their potential impact on the field of adversarial attacks and defenses.

2.2.3. Adversarial Training with Mixup

Mixup-based techniques have generated considerable interest in the field of adversarial robustness, as they hold the potential to improve the generalization of adversarial training approaches. Firstly, Zhang et al. [33] presented a novel data augmentation technique called "Mixup" which generated the new training examples as a linear combination of pairs found in the existing input and their labels. As a result, the model is trained on a linearly interpolated mixture of the images and their corresponding labels. Mixup builds training scheme as follows:

$$\begin{aligned} \tilde{\boldsymbol{x}} &= \lambda \boldsymbol{x_i} + (1-\lambda)\boldsymbol{x_j} \\ \tilde{\boldsymbol{y}} &= \lambda \boldsymbol{y_i} + (1-\lambda)\boldsymbol{y_j} \end{aligned} \tag{2.5}$$

where $\lambda \sim$ Beta (α, α) for $\alpha \in (0, \infty)$ for the two random training data pairs $(\boldsymbol{x_i}, \boldsymbol{y_i})$ and $(\boldsymbol{x_j}, \boldsymbol{y_j})$ respectively.

The literature indicates that the use of adversarial training with mixup augmentation may yield superior performance to that of standard adversarial training, while also enhancing the model's ability to generalize well. In that regard, Manifold-Mixup [51], introduced by Verma et al., is a regularization method that encourages DNNs to predict less confidently with the interpolations of hidden representations in the feature space and to have smooth decision boundaries. Manifold-Mixup provides robustness against single-step adversarial attacks but suffers from multi-step attacks due to gradient masking [51]. Furthermore, Saehyung et al. [46] extended the mixup technique and introduced Adversarial Vertex Mixup, which is a soft-labeled data augmentation approach. This method involves generating a virtual vector in the adversarial direction and using linear interpolation between the virtual vector and the raw input vector to expand the training distribution. The raw input and adversarial vertex, which is γ times far away from raw x in the same direction of adversarial example x_{adv}, are used for adversarial Mixup training.

In addition, Yun et al. proposed CutMix [52], which combines the Mixup technique with a cut-and-paste operation. Unlike Cutout [53], which removes pixels from images, CutMix replaces a randomly cut-out patch from one image with a patch from another image. The ground truth labels are also mixed proportionally to the number of pixels of combined images. This process encourages the model to learn more localizable and robust features while reducing overfitting [52]. Zhang and Xu [54], also presented the adversarial interpolation training scheme to create adversarial samples by minimizing the cosine distance between the representations of natural and perturbed examples.

2.2.4. Other Adversarial Training Methods

To alleviate the challenge of achieving both standard accuracy and adversarial robustness, Zhang et al. presented a multi-step adversarial defense method, referred to as TRADES defense, which maximizes the Kullback-Leibler (KL) Divergence between the softmax outputs of adversarial and natural images for attack generation, and minimizes the term of KL divergence loss and cross-entropy loss on clean samples during the training process. Furthermore, Wu et al. [55] made further progress by introducing Adversarial Weight Perturbation (AWP) to the training process. Their research revealed that the weight loss landscape plays a crucial role in achieving robust generalization, and they proposed a double-perturbation approach that perturbs both inputs and weights to flatten the loss surface.

This thesis draws inspiration from the Feature-Scattering [34] based AT technique proposed by Zhang and Wang. The Feature-Scattering-based AT approach focuses on capturing the inter-sample structure with a single-step attack configuration, which is a departure from previous natural and adversarial training methods that treat training samples independently of each other. The authors argued that perturbing each image individually disregards the inter-sample relationships, which could limit the effectiveness of both attack generation and defense. To address this issue, they proposed the Feature-Scattering-based AT method, which utilizes the optimal transport distance to define inter-sample relationships between adversarial and natural samples within the mini-batch. The unsupervised acquisition of perturbations that maximize the OT distance between natural and adversarial examples is followed by model training using the obtained adversarial examples with those perturbations. As a result of various experiments, the method outperforms in terms of generalization and robustness by preventing gradient masking and label leaking. The same authors also introduced a method called bilateral adversarial training [56]. This approach involves attacking both the images and labels during training using a single-step targeted attack, with the aim of targeting the most confusing class.

2.3. Shift-Invariance Approaches

Although adversarial training is a methodology that has been developed to enhance the generalization and robustness of deep neural networks against adversarial attacks, it may not be adequate on its own to accomplish high levels of generalization. On the other hand, shift-invariance approaches have been proven to be successful in enhancing DNNs' capacity for generalization up to a certain point. In this thesis, we aim to combine strategies by incorporating shift-invariance methods into AT. Our objective is to develop a cutting-edge method that can boost DNNs' robustness against adversarial attacks while improving their generalization performance on unseen data. By doing this, we aim to tackle the limitations of the current adversarial training techniques and contribute to the development of robust DNNs by making them more robust to adversarial attacks. In that regard, state-of-the-art shift invariance methods are explained in the following paragraphs to provide a comprehensive understanding of the proposed approach.

Convolutional neural networks (CNNs) are known to be sensitive to small shifts in input images, which can result in significant changes in the output [57, 58]. To address this issue, various techniques have been developed to improve the shift-invariance behavior of CNNs. One such technique is data augmentation, where the input images are randomly translated, rotated, or scaled during training. More recently, a number of advanced techniques have been developed to improve the shift-invariance behavior of CNNs. These approaches may not only improve their shift-invariance behavior but also slightly increase their accuracy [35].

Shift-equivariance [35] refers to where shifting the input results in an equal shift in the output. In other words, the function is commutable with respect to shifting and feature extraction, as in the following equation,

$$\text{Shift}_{\Delta h,\Delta w}(\tilde{\mathcal{F}}(X)) = \tilde{\mathcal{F}}\left(\text{Shift}_{\Delta h,\Delta w}(X)\right) \quad \forall(\Delta h, \Delta w). \tag{2.6}$$

Shift invariance [35] can be defined as a property where an identical representation is obtained after shifting the input by a certain amount as following

$$\widetilde{\mathcal{F}}(x) = \widetilde{\mathcal{F}}\left(\text{Shift}_{\Delta h, \Delta w}(x)\right) \quad \forall(\Delta h, \Delta w). \tag{2.7}$$

For both Equation (2.6) and Equation (2.7) [35], $\tilde{\mathcal{F}}$ denotes upsampling map function and input x exposed to shifting with the amount of height and weight, Δh and Δw respectively. One of the primary factors contributing to the lack of shift-invariance in CNNs is attributed to downsampling, or stride, which is a linear operation that samples evenly spaced image pixels located at fixed positions on the grid and discards the rest [35, 58].

The well-known solution in signal processing to prevent aliasing is to apply low-pass filtering before downsampling. However, simply inserting this module into deep networks degrades performance [35]. To mitigate this issue, various methods [35,36,58] have been developed to make the DNN shift-invarient.

Zhang [35] elucidated that even though convolution operation is shift equivalent, pooling builds up shift-invariance. For instance, the max-pooling effect on the DNNs in terms of shift-variance is explained in the toy example as seen in custom Figure 2.3. The anti-aliasing CNNs method [35] proposed anti-aliased max-pooling by combining a dense max-pooling layer and strided blurred pooling to boost shift invariance and shift equivariance. The main concern about the anti-aliasing method is its capability to cause information loss. As opposed to the expectation, accuracy improvement is validated on different modern networks as it provides an additional regularization effect. However, there is still a risk of information loss due to over-blurring, and the effectiveness of the method is limited and dependent on the non-linear activation function's behavior [36].

On the other hand, Chaman and Dokmanic introduced Adaptive Polyphase Sampling (APS), which can provide 100% shift-invariance with circular padding. This

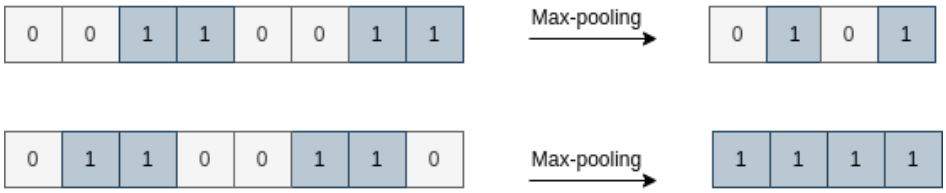


Figure 2.3. Toy example of max-pooling operation to see the effect on shift-invariance performance of the CNNs.

method addresses the main reason for shift-variance to the usage of the fixed grid in sub-sampling(stride), and based on this; the networks always tend to use the first polyphase component as the output of sampling. The method proposes a non-linear downsampling method named APS, which adopts selecting a polyphase which has the highest energy in terms of ℓ_p norm [36]. The potential concern with the APS method is the possibility of multiple polyphase components having identical norms, which could impact its performance. However, it is important to note that such an event is practically impossible.

3. METHOD

In this chapter, we propose three methods aimed at improving the robustness and generalization of DNNs against adversarial attacks. The first method is an adversarial training solution called Perturbation Direction Augmentation for Adversarial Training with Diverse Attacks (PDA-AT), which is introduced and explained in detail. PDA-AT leverages the inter-sample relationship of adversarial examples within the mini-batch to identify the principal perturbation direction. By augmenting the principal perturbation direction within a 45-degree range, PDA-AT enhances the model's robustness against adversarial attacks without requiring any additional back-propagation. The effectiveness of PDA-AT is evaluated through empirical experiments on benchmark datasets.

The second method outlined in this chapter involves the incorporation of shift-invariance methods into the adversarial training process with the objective of enhancing the robustness of DNNs further. We implement and analyze the shift-invariant methodology in adversarial training to enable the model to handle adversarial attacks better, as shift-invariance is a property of neural networks that allows them to recognize patterns regardless of their location in the input space. The third and final method proposed in this study leverages the generalization issue in adversarial training by incorporating mixup, which is the data augmentation method based on convex interpolation.

3.1. Robustness with PDA-AT

AT is one of the successful defense methods that augment each mini-batch of training data with adversarial examples as given the optimization problem of Equation (1.1). In this training scheme, the adversarial perturbation, $\boldsymbol{\delta}$, is obtained through inner maximization using adversarial directions obtained from a single or multi-step attack [21]. While adversarial training has been shown to be effective, it can also suffer from several problems that can negatively impact its performance. These problems

include catastrophic overfitting [23], [28], robust over-fitting [29], label leaking [26], and gradient masking [27]. One of the underlying reasons for these problems is the use of cross-entropy loss to obtain adversarial directions, which can be inadequate to create strong and diverse attacks.

One of the principal challenges associated with traditional AT methods is their inability to generalize effectively to natural samples. Specifically, as the adversarial robustness of models increases, the accuracy of natural samples during testing can decrease below the state-of-the-art performance level, which is an undesirable outcome and affect the widespread usage of these methods. Consequently, we argue that the primary objective of any adversarial training should be to maintain state-of-the-art performance on natural images while simultaneously enhancing the model's robustness to adversarial attacks. Additionally, the adversarial training scheme must be scalable, meaning that the process of generating adversarial samples should not require multiple backpropagation steps, which can be computationally expensive.

In this thesis, we propose an adversarial training framework, PDA-AT, as illustrated in Figure 3.1. In this AT scheme, the primary adversarial directions are generated via Optimal Transport (OT) distance, and they are augmented to support the generalization. As a result, the proposed method improves both accuracies on clean and perturbed images.

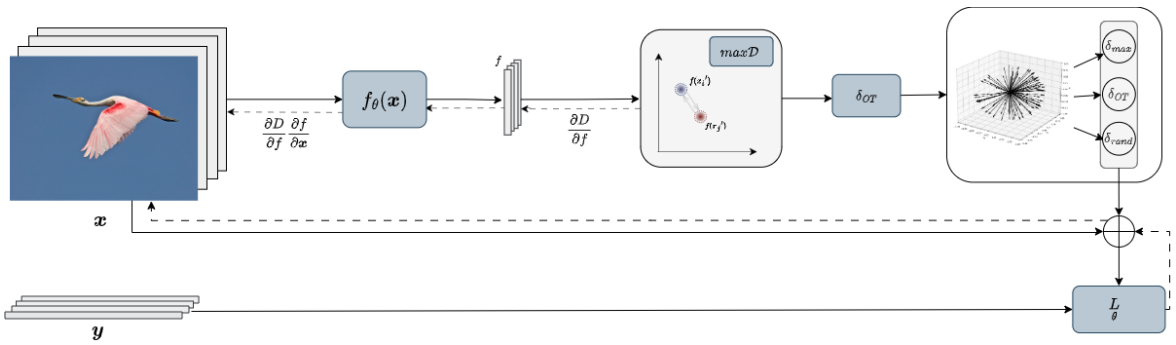


Figure 3.1. Proposed PDA-AT method.

Euclidean distance is one of the commonly used metrics as a straight-line distance between pair of points. On the other hand, Kullback-Leibler (KL) or Jensen-Shannon (JS) divergence both measures the difference between two probability distributions. These methods have limitations in providing useful gradients when two empirical distributions are disjoint or have a non-empty intersection contained in a set of measure zero [59]. An alternative to traditional distance metrics is the OT distance, also known as the Wasserstein distance [60] or earth mover's distance. This method measures the minimum cost of transforming one probability distribution into another, providing another way to quantify the distance between two distributions. The OT distance [61], [62, 63] can be defined as follows:

$$\mathcal{D}(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\gamma} c(\boldsymbol{x}, \boldsymbol{y}). \tag{3.1}$$

The notation $\Pi(\mu, \nu)$ refers to the set of joint distributions $\gamma(x, y)$ whose marginals are $\mu(x)$ and $\nu(y)$. The cost function $c(x, y)$ represents a $K \times K$ transport cost matrix C, where $C_{i,j} = c(x_i, y_j)$ specifies the cost of transporting the i-th data vector in the mini-batch $\boldsymbol{X}$ to the j-th data vector in the mini-batch $\boldsymbol{Y}$.

One of the crucial advantages of the OT distance approach is its ability to consider the spatial distribution of probability mass [61]. This indicates that it is sensitive to both the geometric structure of the distributions and variations in the probability values [61]. Specifically, it is worth noting that even if two distributions have the same probability values, they may exhibit significant differences in terms of their shapes, which can result in a significant OT distance [61], [64]. In this regard, the current study leverages the OT distance to generate adversarial examples.

The AT based on Feature Scattering proposed by Zhang and Wang [34], leverages the inter-sample relationship by utilizing feature matching distance via OT loss. Feature scattering encourages diversity in the data without drastically changing the data manifold with its unsupervised AT scheme. The method substitutes a multi-step PGD attack on cross-entropy loss with a single-step attack on OT distance. Specifically, to

generate adversarial images through maximization of OT loss, the perturbed images of the mini-batch are computed by adding uniform random perturbation within the ϵ-ball and OT loss is measured via OT distance between natural and obtained perturbed samples. In essence, the standard feature scattering AT technique involves the alteration of inter-sample relationships between a natural image and its randomly shifted version of the mini-batch to deepen the discrepancy between natural and adversarial distributions and leads to the creation of an adversarial direction to be used in training.

We argue that the maximization of the OT distance between natural and randomly perturbed samples may charge a constraint on the variety of adversarial perturbations created during each iteration. To address this limitation, we introduce a novel approach that concerns computing the OT distance between the same mini-batch, as specified in the equation. The proposed method based on the feature-matching distance between adversarial couples can be formulated as follows,

$$\delta_{\mathrm{OT}} = \arg\max_{\mathbf{x}} \mathcal{D}\left(\boldsymbol{\nu}_{\mathrm{adv}}, \boldsymbol{\nu}_{\mathrm{adv}}\right) \tag{3.2}$$

$$\mathcal{D}\left(\boldsymbol{\nu}_{\mathrm{adv}}\,, \boldsymbol{\nu}_{\mathrm{adv}}\,\right) = \min_{\mathbf{T}\in \Pi(\mathbf{v_{adv}}, \mathbf{v_{adv}})} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{T}_{ij} \cdot C\left(f(\mathbf{x}_{\mathrm{adv}})_i, f\left(\mathbf{x}_{\mathrm{adv}}\right)_j\right) \tag{3.3}$$

where $\boldsymbol{\nu_{adv}}$ is perturbed version of $\boldsymbol{\nu}$, as defined before. Also, $\mathbf{T} \in \Pi\left(\mathbf{v_{adv}}, \mathbf{v_{adv}}\right) = \left\{\mathbf{T} \in \mathbb{R}_{+}^{n\times n} \mid \mathbf{T}\mathbf{1}_n = \mathbf{u}, \mathbf{T}^{\top}\mathbf{1}_n = \mathbf{v}\right\}$. 1_n is an n-dimensional all-one vector. Transportation cost matrix denoted by $C\left(f(\mathbf{x}_{\mathrm{adv}})_i, f\left(\mathbf{x}_{\mathrm{adv}}\right)_j\right)$ and formulized as follows,

$$C\left(f(\mathbf{x}_{\mathrm{adv}})_i, f\left(\mathbf{x}_{\mathrm{adv}}\right)_j\right) = \left\| f(\mathbf{x}_{\mathrm{adv}})_i - f\left(\mathbf{x}_{\mathrm{adv}}\right)_j \right\|_2^2 \tag{3.4}$$

where $C\left(f(\mathbf{x}_{\mathrm{adv}})_i, f\left(\mathbf{x}_{\mathrm{adv}}\right)_j\right)$ has the element of zero if the pairs include same samples. The transport cost is described as the cosine distance between image features.

The objective of the proposed method for generating adversarial attacks is to improve the exploration of adversarial directions that can modify the distribution of

adversarial samples in a more flexible manner. The method achieves this by eliminating the constraint of intensifying the discrepancy between natural samples and their randomly perturbed versions, as is the case in traditional feature scattering, and by focusing on decision boundaries using multi-step attacks, as in traditional AT methods [34]. Beyond that, our hypothesis is that the adversarial directions must vary to disturb the current mini-batch distribution as the learned representations are updated at each iteration. Therefore, we can enhance attack diversity by avoiding the constraint of feature matching between natural and instead leveraging it via the feature matching distance found in the mini-batch, as proposed. By doing so, it can examine a larger region in the input space and explore different ways of perturbing the samples by utilizing the distance between the adversarial example representations. Our empirical analysis demonstrates that the proposed modification leads to increased attack diversity and improved robustness.

3.1.1. Adversarial Attack Augmentation

In the literature of adversarial robustness, it is well-established that introducing a diverse set of perturbations during model training can help to reduce overfitting to specific types of perturbations. This study presents a novel and intuitive method for augmenting adversarial perturbations. The perturbation δ_{OT} represents the primary adversarial direction which is produced by maximizing feature matching distance via the OT loss. We can leverage the information contained in δ_{OT} to generate new adversarial directions without requiring additional backpropagation steps and significant additional computational costs. Specifically, we propose the random sampling of a new adversarial direction within a region spanning 45-degree from the primary direction via the cosine similarity of normalized δs , formulated as follows:

$$\delta_{\text{rand}} = \cos(\alpha)\delta_{\text{OT}} + \sin(\alpha)\delta_{\perp} \tag{3.5}$$

where α is a hyperparameter that controls the value of a random angle within the range of 0 to 45 degrees, and $\delta_{\perp}$ denotes a randomly generated unit perturbation vector

perpendicular to δ_{OT}. δ_{rand} is generated so as to have cosine-similarity of angle, α with δ_{OT} using normalized $\delta_{\perp}$. In addition to δ_{rand}, one more adversarial direction vector represented by δ_{45} is obtained when α is set to 45 degrees, the maximum allowable deviation angle from the primary attack direction. 45 degrees span can be defined as the hyperparameter of the training.

Consequently, three adversarial attack directions are obtained, as shown in Figure 3.2, with the aim of increasing the variation of adversarial attacks without incurring additional computational costs at each iteration. In Figure 3.2, δ_{rand} and δ_{max} are illustrated with their respective angles, α_{rand} and α_{max} in 2-dimensional space relative to δ_{OT}.

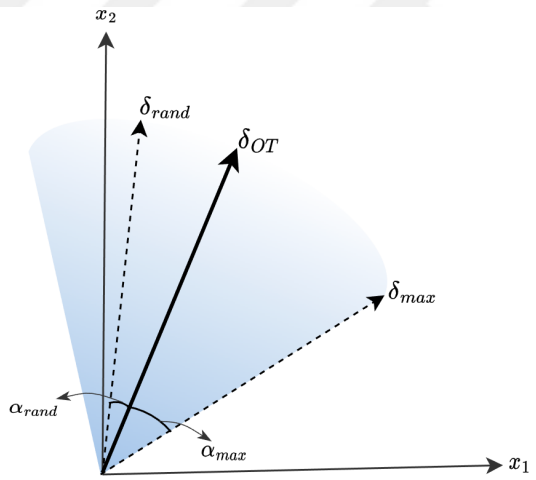


Figure 3.2. Perturbation augmentation in 2-Dimensional space.

Our hypothesis is that neighboring regions of the primary adversarial direction would also lead to an increase in the loss by exploring the broader regions in the decision boundaries with one than one adversarial direction for each natural sample. Thus, it is reasonable to investigate potential new adversarial attack directions that are similar to the gradient of the loss in order to enhance the overall effectiveness of the attack and explore the wider region.

As a result, the model parameters are updated with the mini-batch, which is enriched with augmented perturbed images as follows:

$$\min_{\theta} \frac{1}{N} \left[\sum_{i=1}^{N/3} \mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{\mathrm{OT}}; \theta), y) + \sum_{i=1}^{N/3} \mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{\mathrm{rand}}; \theta), y) + \sum_{i=1}^{N/3} \mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{\mathbf{45}}; \theta), y \right] \quad (3.6)$$

where N denotes the total number of examples found in the mini-batch, and the rest of the notations are the same as indicated before. The aforementioned equation demonstrates that the model parameters, θ, are updated by utilizing samples perturbed with both the primary and augmented adversarial directions. In order to remain within the ϵ-ball and the input domain, the perturbations and perturbed images are clipped. Figure 3.3 outlines the proposed PDA-AT algorithm.

Require: $\mathcal{D}$, OT distance
  for t = 1, ... , T **do**
    for mini-batch $\{\mathbf{x}_i, \boldsymbol{y}_i\}_{i=1}^{n} \sim$ D **do**
      $\boldsymbol{\nu}_{adv} \leftarrow \sum_i v_i \boldsymbol{\delta}_{\mathbf{x}'_i}, \qquad \mathbf{x}' \sim \mathcal{B}(\mathbf{x}, \epsilon)$
      $\boldsymbol{\delta}_{\mathrm{OT}} \leftarrow \arg\max_{\mathbf{x}} \mathcal{D}(\boldsymbol{\nu}_{\mathrm{adv}}, \boldsymbol{\nu}_{\mathrm{adv}}) \qquad \alpha \sim \mathcal{U}(0, \alpha_{max})$
      $\boldsymbol{\delta}_{\mathrm{rand}} \leftarrow \cos(\alpha)\boldsymbol{\delta}_{\mathrm{OT}} + \sin(\alpha)\boldsymbol{\delta}_{\perp}$
      $\boldsymbol{\delta}_{\mathrm{max}} \leftarrow \cos(\alpha_{max})\boldsymbol{\delta}_{\mathrm{OT}} + \sin(\alpha_{max})\boldsymbol{\delta}_{\perp}$
      $\mathbf{x}' \leftarrow [\mathbf{x}' + \boldsymbol{\delta}_{OT}, \mathbf{x}' + \boldsymbol{\delta}_{rand}, \mathbf{x}' + \boldsymbol{\delta}_{max}]$
      model update:
      $\theta \leftarrow \theta - \tau \cdot \frac{1}{n} \sum_{i=1}^{n} \nabla_{\theta} \mathcal{L}(\mathbf{x}', \boldsymbol{y}; \theta)$
    end for
  end for

Figure 3.3. PDA-AT Framework. Robustness for T epochs, D dataset, θ network parameters, ϵ a perturbation budget, τ learning rate, n mini-batch size, α_{max} maximum deviation angle.

3.2. Robustness with Shift-Invariance Approaches

Despite its high efficacy against adversarial attacks, there are limitations of AT techniques. Therefore, exploring alternative approaches could pave the way for a better understanding of adversarial attacks and improve the overall robustness of the models. We investigate the potential of state-of-the-art shift invariance methods, which primarily strive to eliminate the influence of minor shifts in the inputs, on AT for the purposes of improving adversarial robustness. In that regard, we propose adversarial training utilizing low-pass filtering and adaptive polyphase sampling. A comprehensive understanding of these methods is provided in the subsequent sections.

3.2.1. Low-Pass Filtering in Shift-Invariance

The study of Zhang [35] highlights the role of max-pooling operation in causing shift-variance, as depicted in Figure 2.3. The study explained that the max-pooling process involves two architectural operations - the max operator that densely extracts features and the naive subsampling process. To address this issue, Zhang proposes a solution that involves incorporating an additional blurring kernel between these operations. This approach helps to mitigate the effects of the subsampling process and improve shift-equivariance with minimal additional computation.

The correct positioning of the low-pass filter right before subsampling can be applied to widely-used downsampling techniques, such as max-pooling, average-pooling, and strided-convolution. Architectural modifications are given below:

$$\text{MaxPool}_{\text{k, s}} \rightarrow \text{Supsampling}_{\text{s}} \circ \text{Blur}_{\text{m}} \circ \text{Max}_{\text{k}} = \text{BlurPool}_{\text{m, s}} \circ \text{Max}_{\text{k}} \quad (3.7)$$

$$\text{Relu} \circ \text{Conv}_{\text{k, s}} \rightarrow \text{BlurPool}_{\text{m, s}} \circ \text{Relu} \circ \text{Conv}_{\text{k, 1}} \quad (3.8)$$

$$\text{AvgPool}_{\text{k, s}} \rightarrow \text{BlurPool}_{\text{m, s}} \quad (3.9)$$

where k is pooling kernel with stride s of $\text{MaxPool}_{\text{k, s}}$ and anti-aliasing filter with kernel shape of m $\times$ m represented with Blur_{m}. $\circ$ is used to indicate the sequential flow of

operations, it means that the output of one operation is serving as the input to the next operation.

We hypothesize that making DNNs more shift-invariant might also boost adversarial robustness. The proposed method involves the utilization of anti-aliasing, which aims to make the output of the DNN irrelevant to rotation and translation. The proposed method is expected to lead to anti-aliased DNNs may learn more robust features than traditional models. In that regard, traditional feature-scattered based [34] AT can be trained with shift-invariant DNNs by implementing a blur kernel before the subsampling layer. Through empirical evaluation, the thesis aims to provide an analysis that the proposed method can effectively enhance the adversarial robustness of DNNs.

In the AT scheme, adversarial attack direction, δ_{FS} is obtained via OT distance loss between natural and adversarial counterparts of them.

$$\delta_{\mathrm{FS}} = \arg\max_{\mathbf{x}} \mathcal{D}\left(\boldsymbol{\nu}, \boldsymbol{\nu}_{\mathrm{adv}}\right) \tag{3.10}$$

where notations are kept as before. The model parameters, θ are updated with the mathematical formulation as follows

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{\mathrm{FS}}; \theta), y). \tag{3.11}$$

Figure 3.4 summarize the traditional feature scatter adversarial training scheme. The AT algorithm will update the parameters of the model, which utilizes a blurring kernel before the subsampling operation.

Require: $\mathcal{D}$ OT distance

for t = 1, ... , T **do**

for mini-batch $\{\mathbf{x}_i, \boldsymbol{y}_i\}_{i=1}^{n} \sim$ D **do**

$\boldsymbol{\nu}_{adv} \leftarrow \sum_i v_i \boldsymbol{\delta}_{\mathbf{x}_i'}, \qquad \mathbf{x}' \sim \mathcal{B}(\mathbf{x}, \epsilon)$

$\boldsymbol{\delta}_{\mathrm{FS}} \leftarrow \arg\max_{\mathbf{x}} \mathcal{D}(\boldsymbol{\nu}, \boldsymbol{\nu}_{\mathrm{adv}})$

$\mathbf{x}' \leftarrow \mathbf{x}' + \boldsymbol{\delta}_{FS}$

model update:

$\theta \leftarrow \theta - \tau \cdot \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \mathcal{L}(\mathbf{x}', \boldsymbol{y}; \theta)$

end for

end for

Figure 3.4. Traditional Feature Scatter [34] Framework. Adversarial Training for T epochs, D dataset, θ network parameters, ϵ a perturbation budget, τ learning rate, n mini-batch size.

3.2.2. Adaptive Polyphase Sampling in Shift-Invariance

Chaman and Dokmanic [36] point out that downsampling in the form of stride or pooling can lead to loss of shift-invariance. They have expressed their concerns regarding the effectiveness of existing anti-aliasing methods in achieving shift-invariance. They argue that anti-aliased CNNs have certain limitations due to non-linear activation functions such as ReLu in their study. Moreover, they suggest anti-aliasing beyond a certain threshold may lead to over-blurring and a decrease in accuracy.

The conventional approach of downsampling involves sampling image pixels at fixed positions on the grid by discarding the rest. Specifically, if the image is shifted, the pixel intensities located on the fixed grid will change, leading to a different subsampled output. As a measure of this, APS proposes choosing of sampling grid adaptively with the highest ℓ_p norm corresponding to the highest energy. This technique ensures that 100% shift-invariance and the subsampled output remains consistent even when the

image is shifted, resulting in identical outcomes for shifted and unshifted samplings.

The shift-invariant result of the subsampling operation of the APS method a, y_{APS}, is computed based on [36] as follows

$$
\begin{aligned}
y_{\text{APS}} &= y_{i_1 j_1}, && (3.12)\\
\text{s.t. } i_1, j_1 &= \underset{i,j}{\operatorname{argmax}} \left\{ \|y_{ij}\|_p \right\}_{i,j=0}^{s-1}, && (3.13)\\
\text{where } \{y_{ij}\} &= \mathbf{x}\left(sn_1 + i, sn_2 + j\right) = \{\mathbf{x}_{k,ij}\}_{k=1}^{C}. && (3.14)
\end{aligned}
$$

Here, candidate outcomes of sampling as polyphase components are denoted with $\{y_{ij}\}_{i,j=0}^{1}$. The value of p determines the type of norm being used. $\mathbf{x}$ is the input, C is the channel number, and s represents the stride. n_1 and n_2 are integer variables that represent the position of the subsampling operation within the input image. If the two outputs of subsampling have the same norm, there is a possibility to have inconsistency in shift-invariance, yet it is a nearly impossible event in practice when considering the continuous distribution of images.

We hypothesize that combining APS with adversarial training might provide additional improvement in robustness and generalization. This is because the perfect shift-invariance property of APS reduces sensitivity to small shifts in the input, while its cooperation with AT might help to improve the model's ability to handle more complex perturbations by reducing overfitting to specific attacks. As a result, the combination of these techniques can lead to better generalization and robustness performance. We aim to tackle the limitations of the current AT techniques and contribute to the development of robust DNNs by taking advantage of generalization performance of APS.

Furthermore, although anti-aliasing is not necessary for shift invariance, using it before downsampling can be beneficial in terms of increasing accuracy. In this regard, we will demonstrate the effects of APS with and without LPS in the adversarial training scheme through various experiments. The shift-invariance model by adopting APS will be trained with traditional feature-scatter [34] based AT scheme as presented in the

algorithm of Figure 3.4. We investigate the solution of APS for adversarial robustness and generalization building upon the existing FS model.

3.2.3. Robustness with Mixup-based Adversarial Training

Mixup is the state-of-the-art data augmentation technique that uses the convex combination of data and its label to enhance the generalization. The straightforward idea, formulized in Equation (2.5), is proven that the improve the robustness of the network 2.7 times against FGSM white-box attack and 1.25 times against black-box FGSM attack [33]. In this thesis, we aim to combine mixup and AT strategies by incorporating mixup methodology into feature scattering-based AT for seeking to enhance the robustness and generalization performance of the DNNs. In that regard, four different ways of integration into the AT scheme are developed.

3.2.3.1. Mixup + Feature Scatter. In this approach, the main natural input and targets are exposed to the mixup before entering the AT scheme. With the mixup operation, natural images denoted by x and their corresponding labels y are interpolated with the shuffled mini-batch of it, x' and y', resulting in the production of $\tilde{x}$ and $\tilde{y}$. Adversarial training is performed on the obtained images. For the AT, OT distance loss is used to get adversarial direction as a single-step attack, and the final cross-entropy loss of AT is calculated between the prediction of the model and $\tilde{y}$. The details of the proposed algorithm are given in Figure 3.5. As seen in the Figure 3.5, mixup operations are followed by traditional feature scatter based training operations.

3.2.3.2. Feature Scatter + Mixup. In the first approach, linear interpolation of inputs is carried on at the beginning of the training experiment. On the other hand, mixup operations can be applied after adversarial examples are obtained utilizing OT distance loss. This approach can be put into training by obtaining adversarial direction for a particular mini-batch, then executing the mixup operation using the acquired adversarial images. The process is completed by updating the model parameters. This

approach employs the same operations given in Figure 3.5, but executes the mixup operation after adversarial sample generation.

Require: $\mathcal{D}$ OT distance
  for t = 1, ... , T **do**
    for mini-batch $\{\mathbf{x}_i, \boldsymbol{y}_i\}_{i=1}^{n} \sim$ D **do**
      mixup operation:
      $\tilde{\boldsymbol{x}} \leftarrow \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{x'} \qquad \lambda \sim \mathcal{D}(0,1)$
      $\tilde{\boldsymbol{y}} \leftarrow \lambda \boldsymbol{y} + (1-\lambda)\boldsymbol{y'}$
      adversarial direction obtainment:
      $\tilde{\boldsymbol{\nu}_{adv}} \leftarrow \sum_i v_i \boldsymbol{\delta}_{\tilde{\mathbf{x}_i'}}, \qquad \tilde{\mathbf{x}'} \sim \mathcal{B}(\tilde{\mathbf{x}}, \epsilon)$
      $\boldsymbol{\delta}_{\mathrm{FS}} \leftarrow \arg\max_{\mathbf{x}} \mathcal{D}(\tilde{\boldsymbol{\nu}}, \tilde{\boldsymbol{\nu}}_{\mathbf{adv}})$
      $\tilde{\mathbf{x}'} \leftarrow \tilde{\mathbf{x}'} + \boldsymbol{\delta}_{FS}$
      model update:
      $\theta \leftarrow \theta - \tau \cdot \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \mathcal{L}(\tilde{\mathbf{x}'}, \tilde{\boldsymbol{y}}; \theta)$
    end for
  end for

Figure 3.5. Mixuped Feature Scatter [34] Framework. Adversarial Training for T epochs, D dataset, θ network parameters, ϵ a perturbation budget, τ learning rate, n mini-batch size.

3.2.3.3. Mixup in OT Distance. Other than the straightforward integrations of mixup into feature scatter based AT, we propose to utilize mixup in attack direction generation by measuring the distance between mixuped natural samples and their non-mixuped randomly perturbed correspondences of the original mini-batch. The approach uses distance following formula:

$$\delta = \arg\max_{\mathbf{x}} \mathcal{D}\left(\tilde{\boldsymbol{\nu}}, \boldsymbol{\nu}_{\mathrm{adv}}\right) \tag{3.15}$$

where $\tilde{\boldsymbol{\nu}}$ demonstrate that the marginal distribution of the inputs exposed to the mixuped with Equation (2.5) whereas $\boldsymbol{\nu}_{\mathrm{adv}}$ is the marginal distribution of perturbed original

input images. The rest of the adversarial training uses the non-mixuped original batch to perturb the obtained adversarial training.

3.2.3.4. Mixup in OT Distance and CE Loss. Besides utilizing mixup in OT Distance, we can also leverage it in Cross-Entropy (CE) loss function of it. After the adversarial direction is obtained with Equation 3.15, the obtained adversarial examples are subjected to mixup with the same λ parameter and index shuffling used in the first mixup operation in OT loss distance calculation.

In this thesis, we proposed three different approaches for improvement on robustness and generalization performance of AT. PDA-AT is AT technique based on the inter-sample relationship via OT distance within the randomly perturbed mini-batches. This method incorporates data augmentation through the generation of new perturbation directions without requiring additional backpropagation steps. Additionally, we propose AT approaches that leverage shift-invariant techniques, such as APS and LPS, to enhance AT performance. Furthermore, we propose to enhance AT performance through mixup-based data augmentation. The subsequent chapter provides detailed explanations of the experimental results, implementation details, and result discussions.

4. EXPERIMENTS

In this chapter, we present the experimental setup and results for the analysis of the proposed methods. We conducted various expensive experiments to analyze and validate the effectiveness of our proposed methods thoroughly. We also present a detailed analysis of the experimental results to gain a deeper understanding of the factors that affect the performance of our proposed methods. Overall, this chapter provides a comprehensive overview of the experiments conducted and the results obtained, which demonstrate the efficacy of our proposed techniques.

4.1. Datasets

This study was conducted on five publicly available datasets with different configurations. The datasets used in the study are CIFAR-10 [65], CIFAR-100 [65], SVHN [66], Tiny-Imagenet [7] , and GTSRB [67].

CIFAR (Canadian Institute For Advanced Research) has created two widely-used image classification datasets called CIFAR-10 and CIFAR-100. The CIFAR-10 and CIFAR-100 are datasets created among 80 million small images collected by Alex Krizhevsky et al. [65]. The CIFAR datasets are widely utilized in the field of computer vision due to the low resolution of the images they contain. This feature of the datasets reduces the computational complexity of experiments, making them more manageable in terms of computational resources. As a result, the CIFAR datasets have become the predominant benchmark datasets in the field of computer vision.

The CIFAR-10 dataset is considered a standard benchmark for adversarial training and contains $60,000$ RGB images of size 32×32 distributed across ten different object classes, each containing $6,000$ images. Similarly, the CIFAR-100 dataset contains the same number of images and dimensions as CIFAR-10 but includes 100 object classes, each containing 600 images. CIFAR-100 is grouped into 20 superclasses. More-

over, CIFAR-100 is organized into 20 superclasses, with each image having both a fine label (i.e. the specific class to which it belongs) and a coarse label (i.e. the superclass to which the fine label belongs). Both datasets are split into $50,000$ training images and $10,000$ test images.

The Street View House Number (SVHN) dataset [66] is a widely used benchmark dataset in computer vision and consists of approximately $100,000$ labeled images of digits. The images were collected from Google Street View house numbers with varying sizes and orientations. In this dataset, each image contains a digit ranging from 0 to 9, and the digits are positioned horizontally in the image. The dataset is divided into training and test sets, with the training set containing $73,257$ images and the test set containing $26,032$ images. Due to its real-world nature and the diverse range of conditions under which the images were captured, the SVHN dataset presents significant challenges for computer vision models.

Tiny-ImageNet [7] is a dataset that comprises 110,000 images divided into 200 object categories, each containing 500 training images and 50 validation images. The dataset was created as a more manageable version of the ImageNet dataset [6], which contains over 14 million images across $1,000$ object categories. To achieve this, a subset of 200 diverse object categories was randomly selected from ImageNet while still maintaining the diversity of the original dataset. The images in the Tiny-ImageNet dataset have varying sizes and resolutions, with a majority having dimensions of 64×64 pixels. Its smaller size and diversity of object categories make it an efficient alternative to the larger and more computationally expensive ImageNet dataset for developing and evaluating image classification models.

The German Traffic Sign Recognition Benchmark (GTSRB) [67] dataset is a publicly available dataset of traffic signs collected from real-world scenarios in Germany. There are $39,209$ training images and $12,630$ test images in the dataset, which consists of 43 types of traffic signs. Many lighting setups and elaborate backgrounds can be found in the images.

4.2. Deep Neural Networks Architectures

We employed two different DNN architectures in our experiments to evaluate the proposed methods. Specifically, the WideResNet-28 is employed for the experiments of PDA-AT, and ResNet-18 is utilized for the shift-invariant based AT experiments and AT method cooperated with mixup-based data augmentation. These models have proven to be highly effective in various computer vision tasks and are widely used as benchmark models for evaluating novel techniques.

ResNet (Residual Network) [10] was introduced to address the issue of vanishing gradients in DNNs. This problem occurs when gradients become too small to be useful for propagating information through the model, resulting in poor performance. ResNet overcomes this issue by introducing residual connections, which allow for alternate shortcut paths for gradient flow. ResNet [10] can be scaled up to 152 layers with high accuracy. Resnet-18 is a variant of ResNet that contains 72-layer architecture with 18 deep layers. It comprises convolution layers followed by batch normalization and ReLU activation, with residual connections between each block of layers.

Wide Residual Networks (WideResNet) [68] are a variant of ResNets. WideResNets have decreased depth and increased width when compared to ResNet, which enhances the model's representation power of features. WideResNet increases the number of filters in each convolutional layer. Also, ReLu and batch norm operations come first from the convolution layer in WideResnet blocks, which is introduced as pre-activation. WideResnet is controlled by the parameter defined as the width factor, which represents the number of filters in each layer. In our experiments, it set to 28 corresponds to WideResnet28.

4.3. Training Scheme

All experiment of the thesis is performed with the DNN architecture of WideResNet-28 and ResNet-18. WideResNet-28 is utilized for the PDA-AT, and mixup-based ad-

versarial training experiments, whereas ResNet-18 is preferred for the experiments of AT based on shift-invariance enhancement. Regarding that, all models are trained up to 200 epoch with a batch size of 60. Stochastic gradient descent (SGD) is utilized as an optimizer with weight decay of 2×10^{-4} and momentum of 0.9. The initial learning rate is set to 0.1 The learning rate value used during the experiments is different for the SVHN dataset compared to the other datasets. Specifically, a learning rate of 0.01 is used for the SVHN dataset, while a value of 0.1 is used for the other datasets following the suggestions in [34]. At the 30th and 60th epochs, the learning rate is decayed by a factor of 0.10. For the adversarial training, the maximum allowable perturbation amount, ϵ, is set to 8/255.

The CIFAR and Tiny-Imagenet datasets are augmented with flipping and random cropping during the training process. The input images are cropped randomly in the shape of 32×32 with padding size 4, except for the Tiny-Imagenet dataset, which is cropped into 64×64. Label smoothing is applied to the input labels with a factor of 0.5. The Sinkhorn algorithm is utilized for the experiments that include the computation of OT distance, where the regularization parameter is set to 0.01 in a one-step attack configuration [34].

The proposed methods are implemented using codebases obtained from multiple sources, including [40], [69–72]. PyTorch, a popular deep learning framework, is used to implement all the methods by modifying existing codebases.

4.4. Results of PDA-AT

The method of PDA-AT aims to leverage the robustness of the DNNs without sacrificing natural accuracy when compared to state-of-the-art methods. PDA-AT framework increases the attack diversity by adopting adversarial examples generation based on the OT distance between a mini-batch of randomly perturbed versions of natural samples. Furthermore, the framework also improves attack diversity by exploring alternative perturbation directions within a 45-degree spanning region. In this regard,

various extensive experiments were carried out to analyze the proposed method. The models trained on five different benchmark datasets are evaluated against white-box and black-box attacks.

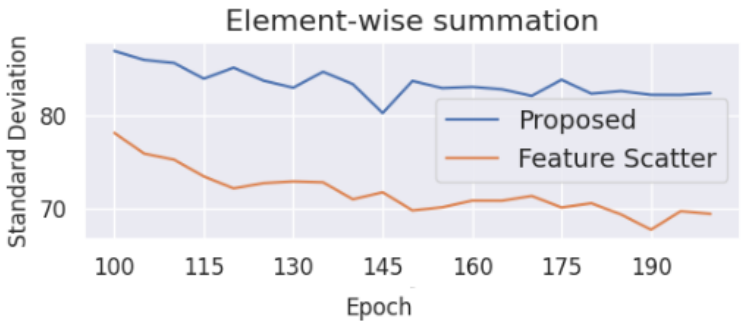
4.4.1. Baseline Studies

The baseline techniques for the evaluation process are selected from the state-of-the-art gradient-based AT methodologies. It is worth noting that neither the proposed method nor the selected baseline methods utilize synthetically generated data during training for the sake of an equivalent comparison. The performance of the listed state-of-the-art AT methods is reported:

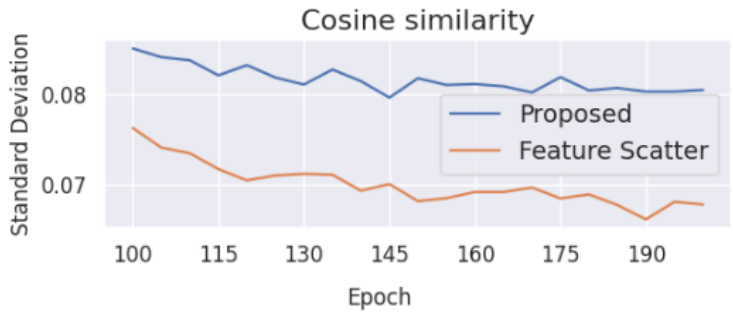
- Natural: The standard training of natural images.
- Madry: AT of images perturbed with PGD attack proposed by Madry et al. [21]. The method is widely recognized as one of the most effective defense methods and is regarded as a pioneering study in the field.
- Bilateral: Training of both adversarially perturbed image and label [56]. It utilizes a single-step targeted attack by targeting the most confusing class.
- Feature Scatter (FS): AT framework with adversarial example generation based on the maximization of feature matching distance between natural and randomly perturbed images via OT distance loss [34].
- Adv-Interp: AT scheme employs adversarial interpolation for the creation of adversarial examples and their adversarial examples [54]. It adopts maximization of cosine distance between features of perturbed and natural different mini-batch for adversarial image generation.
- AV-Mixup: AT of soft-labeled data augmentation approach [46]. Adversarial vertexes are generated to be used in AT, in the direction of perturbation to extend the training distribution.

4.4.2. Attack Diversity

One of the primary objectives behind this thesis and the proposed method is to augment the diversity of attacks employed in the context of adversarial training. The proposed methodology centers on examining the interconnections between adversarial examples within a given mini-batch to augment the attack and increase the diversity. We scrutinize the extent of attack diversity by drawing comparisons between the proposed method and the traditional Feature Scatter [34] AT approach outlined in prior literature.



(a) Element-wise summation.



(b) Cosine similarity.

Figure 4.1. Diversity of adversarial attack directions in training. The element-wise sum of the elements of the adversarial direction tensor is shown as a change in standard deviations in the top plot. Standard deviations of cosine similarities between random reference direction and adversarially generated direction is illustrated in bottom plot.

To this end, adversarial directions of CIFAR-10 training images are saved at several epochs of training for both the proposed method and Feature Scatter initially. Subsequently, a reference vector is sampled to calculate the cosine similarities between the reference vector and adversarial directions. In the bottom graph of Figure 4.1, the standard deviations of cosine similarities at various epochs are visualized. Additionally, the diversity of attacks is evaluated with the element-wise sum of the gradient direction tensor's elements across each training sample. The standard deviation of the element-wise summation of adversarial directions over epochs is plotted in the top graph of Figure 4.1. We envisage that the proposed attack approach utilizing OT distance between randomly perturbed images as outlined in Equation (3.2) and Equation (3.3) yield a higher standard deviation as compared to Feature Scatter approach for both diversity comparison. The graphical representations in Figure 4.1 serve as empirical evidence, indicating that the proposed methodology can result in a more expansive and diverse set of attack directions than the traditional one-step OT distance attack.

4.4.3. Results against White-Box Attacks

The test performance of the proposed PDA-AT framework is evaluated against various white-box attacks, namely FGSM, PGD, and CW, on five datasets which includes CIFAR-10, CIFAR-100, SVHN, Tiny-Imagenet, and GTSRB. In the following tables, we evaluate the proposed models and baselines to demonstrate our contribution.

Our evaluation includes accuracy metrics for clean test images that are identified as natural, while also denoting iterative attacks as PGDT and CWT, where T represents the number of attack iterations. Specifically, the models are subjected to attacks consisting of 10, 20, and 100 iterations. In experiments, iterative attacks are implemented with a step size of 2/255, and single-step attacks with a step size of 8/255, utilizing randomized initialization. Adversarial budget indicated with ϵ is set to 8/255.

In Table 4.1, Table 4.2, and Table 4.3 listed the accuracy of the baselines and proposed methods for the datasets of CIFAR-10, CIFAR-100 and SVHN, respectively.

In the tables of accuracy evaluation, the asterisk symbol (*) denotes outcomes from scratch training, while dagger (†) symbols indicate results sourced from the study of Baytas and Deb [47]. All other values of baseline models are the best results reported by the primary studies, and those marked with a hyphen (-) are currently unavailable.

In those tables, δ_{OT} indicates the AT with one-step attack given in Equation (3.2) whereas $\delta_{\mathrm{OT,\ rand,\ 45}}$ denotes the version of this AT with adversarial attack augmentation acquired by 3.5. In our evaluation comparing the results obtained from the Feature Scatter approach [34] and our proposed method, δ_{OT} serves as experimental evidence for the utilization of intersample relationships within the perturbed mini-batch, as opposed to comparing the natural and perturbed batches. Furthermore, our comparative analysis between the proposed methods effectively highlights the additional benefits of the perturbation augmentation technique.

Table 4.1. Robustness comparison of proposed and baseline models against white-box attacks on CIFAR-10 dataset. The asterisk symbol (*) denotes results from scratch training, while dagger (†) symbols indicate results sourced from the study of Baytas and Deb [47]. All other values of baseline models are the best results reported by the primary studies, and those marked with a hyphen (-) are currently unavailable.

CIFAR-10								
Defences	Natural	FGSM	PGD10	PGD20	PGD100	CW10	CW20	CW100
Natural*	96.08	34.7	0.01	0.00	0.00	0.01	0.00	0.00
Madry† [21]	87.25	62.64	47.33	45.91	45.29	-	46.99	46.54
Bilateral [56]	91.00	70.70	63.00	57.80	55.20	-	56.20	53.80
FS [34]	90.0	78.40	70.9	70.5	68.6	62.6	62.4	60.6
Adv-Interp [54]	90.30	78.00	-	73.50	73.00	-	69.70	68.70
AV-Mixup [46]	93.24	78.25	62.67	58.23	-	53.63	-	-
Ours - δ_{OT}	90.24	78.11	73.97	72.29	71.1	62.11	59.66	57.52
Ours - $\delta_{\mathrm{OT,\ rand,\ 45}}$	92.83	91.54	74.16	71.83	66.74	64.35	60.6	53.85

In the case of CIFAR-10, δ_{OT} performed better against iterative attacks such as PGD and CW. Specifically, it achieves 71.1% and 57.52% accuracy against PGD100 and CW100, respectively. However, its performance was found to be relatively comparable to the baselines for natural images and images subjected to FGSM attack. Meanwhile, $\delta_{\mathrm{OT,\ rand,\ 45}}$ demonstrated enhanced performance in terms of natural accuracy and accuracy against FGSM-based attacks while maintaining robust accuracy at a level similar to that of the baselines. The proposed method achieved the highest scores among the baselines evaluated in the study for robustness against FGSM and PGD10 attacks, with accuracies of 91.54% and 74.16%, respectively, while performing a successful natural accuracy of 92.83%.

Table 4.2. Robustness comparison of proposed and baseline models against white-box attacks on CIFAR-100 dataset. The asterisk symbol (*) denotes results from scratch training, while dagger (†) symbols indicate results sourced from the study of Baytas and Deb [47]. All other values of baseline models are the best results reported by the primary studies, and those marked with a hyphen (-) are currently unavailable.

CIFAR-100								
Defences	Natural	FGSM	PGD10	PGD20	PGD100	CW10	CW20	CW100
Natural*	79.78	5.57	0.00	0.00	0.00	0.00	0.00	0.00
Madry† [21]	59.78	32.70	23.49	22.78	22.44	-	23.05	22.87
Bilateral [56]	66.20	31.30			22.40	-	-	20.00
FS [34]	73.9	61.0	47.6*	47.2	46.2	30.76*	34.60	30.60
Adv-Interp [54]	73.6	58.3	-	41.00	40.2	-	32.4	31.2
AV-Mixup [46]	74.81	62.76	-	38.49		-	-	-
Ours - δ_{OT}	71.8	49.31	41.5	40.83	25.08	23.18	22.48	24.46
Ours - $\delta_{\mathrm{OT,\ rand,\ 45}}$	75.73	73.04	49.19	48.83	48.77	35.16	34.64	33.86

As for CIFAR-100, δ_{OT} approach failed to achieve state-of-the-art performance, which can be attributed to the complexity of the dataset. CIFAR-100 is recognized as one of the challenging benchmark datasets because it consists of 100 classes with 6000 images. It makes it difficult to represent the intersample relationship of adversarial images within the batch. On the other hand, when the attack diversity is increased through the approach of $\delta_{\mathrm{OT,\ rand,\ 45}}$, the robust accuracies outperform against FGSM, PGD, and CW attacks with the highest scores among the baselines. Furthermore, it obtains the highest natural accuracy among the listed ATs.

Regarding SVHN dataset, δ_{OT} demonstrated an ability to achieve higher robust accuracies when faced with an increased number of iterative attacks, similar to the results observed for the CIFAR-10 dataset. In particular, it achieves 55.43% and 48.43% robust accuracy against PGD100 and CW100, respectively. Contrarily, the $\delta_{\mathrm{OT,\ rand,\ 45}}$ approach demonstrated superior performance in classifying natural images, achieving an accuracy of 97.15%, and in robustness against FGSM and PGD10 attacks, with adversarial accuracies of 97.33% and 72.48%, respectively.

Overall, perturbations based on the distribution of the perturbed mini-batch rather than expanding the gap between mini-batch distributions of the natural and adversarial samples can enhance robustness in the CIFAR-10 and SVHN datasets. On the contrary, δ_{OT} without perturbation augmentation does not contribute to the robustness of CIFAR-100. The difference in performance can be attributed to the larger number of classes in CIFAR-100, which may limit the generalization of perturbing different mini-batches at each iteration. Furthermore, the proposed perturbation augmentation leads to significant improvements in both adversarial and natural accuracies, surpassing the performance of the baselines. Specifically, in comparison with the highest baseline accuracy, the proposed method achieves an increase of 13.14%, 10.28%, and 5.82% in accuracy against FGSM attacks on the CIFAR-10, CIFAR-100, and SVHN datasets, respectively.

Table 4.3. Robustness comparison of proposed and baseline models against white-box attacks on SVHN dataset. The asterisk symbol (*) denotes results from scratch training, while dagger (†) symbols indicate results sourced from the study of Baytas and Deb [47]. All other values of baseline models are the best results reported by the primary studies, and those marked with a hyphen (-) are currently unavailable.

SVHN								
Defences	Natural	FGSM	PGD10	PGD20	PGD100	CW10	CW20	CW100
Natural*	96.33	45.29	1.59	0.62	0.39	1.04	0.63	0.48
Madry† [21]	90.74	64.50	44.23	41.38	40.37		42.46	41.60
Bilateral [56]	94.10	69.80	-	53.90	50.30	-	-	48.90
FS [34]	96.20	83.50	55.40*	62.90	52.00	58.18*	61.30	50.80
Adv-Interp [54]	94.10	81.83	-	-	-	-	-	-
AV-Mixup [46]	95.59	91.51	37.97	-	-	-	-	-
Ours - δ_{OT}	96.31	95.78	71.74	66.38	55.43	67.35	60.75	48.43
Ours - $\delta_{\text{OT, rand, 45}}$	97.15	97.33	72.48	65.02	50.61	65.13	56.58	42.45

Table 4.4 and Table 4.5 display the performances achieved by the proposed method of PDA-AT on the Tiny-Imagenet and GTSRB datasets, respectively. Due to the unavailability of baseline results for these datasets on the original studies, we present the results by comparing them with naturally trained model and the Feature Scatter [34] model, which were trained from scratch. All models used in these comparisons were trained with the same training setup and attack configurations as the other datasets. Although the Tiny-Imagenet dataset presents a significant challenge due to size of images and the increased number of classes to 200, the robustness provided by the proposed method is not significantly different from that of Feature Scatter. In contrast, our proposed method shows a clear improvement in the robustness of the GTSRB dataset, as can be observed from the results in Table 4.5.

Table 4.4. Accuracy comparison of models for Tiny-Imagenet dataset. All models are trained from scratch.

Tiny-Imagenet								
Defences	Natural	FGSM	PGD10	PGD20	PGD100	CW10	CW20	CW100
Natural	69.04	2.28	0.02	0.01	0.00	0.00	0.00	0.00
FS	57.42	24.42	12.54	11.25	10.33	9.11	7.97	7.55
Ours-$\delta_{\text{OT, rand, 45}}$	57.34	25.31	12.93	11.73	11.24	11.14	10.31	9.97

Table 4.5. Accuracy comparison of models for GTSRB dataset. All models are trained from scratch.

GTSRB								
Defences	Natural	FGSM	PGD10	PGD20	PGD100	CW10	CW20	CW100
Natural	98.62	63.34	19.94	14.31	11.29	17.29	13.22	11.36
FS	96.5	91.81	87.82	84.56	78.52	76.01	72.02	67.42
Ours-$\delta_{\text{OT, rand, 45}}$	97.01	94.86	90.46	88.27	82.66	78.56	74.76	69.35

4.4.4. Results against Black-Box Attacks

In addition, the performance of the proposed model is assessed under transfer-based black-box attack settings [17]. To achieve this, the models trained solely on natural and adversarial samples with PGD attacks [21] are employed for the generation of gradient-based black-box attacks. The AT model is trained with PGD7. The proposed model and FS [34] models are evaluated using PGD and CW attacks generated under black-box settings. The models of the Feature Scatter approach are trained from scratch, serving to establish a clear baseline for evaluating the performance improvements of the proposed approach.

The results presented in Tables 4.6 and 4.7 demonstrate the adversarial accuracies obtained by the proposed PDA-AT approach and the Feature Scatter AT [34]

on CIFAR-10 and CIFAR-100 datasets, respectively. As seen in those tables, the robustness achieved by the proposed approach can generalize to black-box attacks and exceeds the performance of the Feature Scatter [34] AT.

Table 4.6. Results against black-box attack transferred from natural model on CIFAR-10.

Attacks	**Defences**	
	FS	Proposed
PGD20	89.42	91.94
PGD100	89.33	92.14
CW20	89.41	91.87
CW100	89.34	91.88

Table 4.7. Results against black-box attack transferred from natural model on CIFAR-100.

Attacks	**Defences**	
	FS	Proposed
PGD20	76.60	78.40
PGD100	76.46	77.94
CW20	77.71	78.86
CW100	77.44	78.76

Moreover, the proposed approach was also evaluated under black box settings, where the model is subjected to attack transferred from an adversarially trained model. The results of the evaluation can be seen in Table 4.8. The inclusion of the Feature Scatter model's performance in the same table serve as a benchmark for comparison and demonstrate the performance enhancement achieved by the proposed approach.

Table 4.8. Results against black-box attack transferred from AT [21] model on CIFAR-10.

Attacks	**Defences**	
	FS	Proposed
PGD20	89.42	91.94
PGD100	89.33	92.14
CW20	89.41	91.87
CW100	89.34	91.88

In addition to white-box and black-box attacks, we evaluated the model's performance against AutoAttack [42], which is an ensemble of four diverse attacks. The δ_{OT} model was found to achieve a 35.79% robust accuracy against AutoAttack, while the $\delta_{\text{OT, rand, 45}}$ model underperformed, achieving only 10.59%. As a baseline, Feature Scatter [34] achieved an adversarial accuracy of 36.69% against AutoAttack. Thus, the proposed approach does not provide any additional robustness against AutoAttack.

4.4.5. Variations of Perturbation Augmentations

In this section, we investigate the optimum combination of perturbation augmentation setup to further enhance the attack performance. We first identify the primary direction using the ordinary optimal transport distance and then augment it within a cosine similarity region of 45 spanning degrees. Through a number of experiments and analyses, we have found that the most effective attack perturbation involves a combination of the original perturbation direction and the augmented directions generated by both random and maximum allowable perturbations based on the results demonstrated in the Table 4.9. It is worth noting that the perturbation direction, δ_{FS}, is obtained by computing the optimal transport distance between the natural and adversarial examples as original study of [34] hat were augmented in the experiments. This approach enables us to observe the independent effect of the proposed perturbation augmentation.

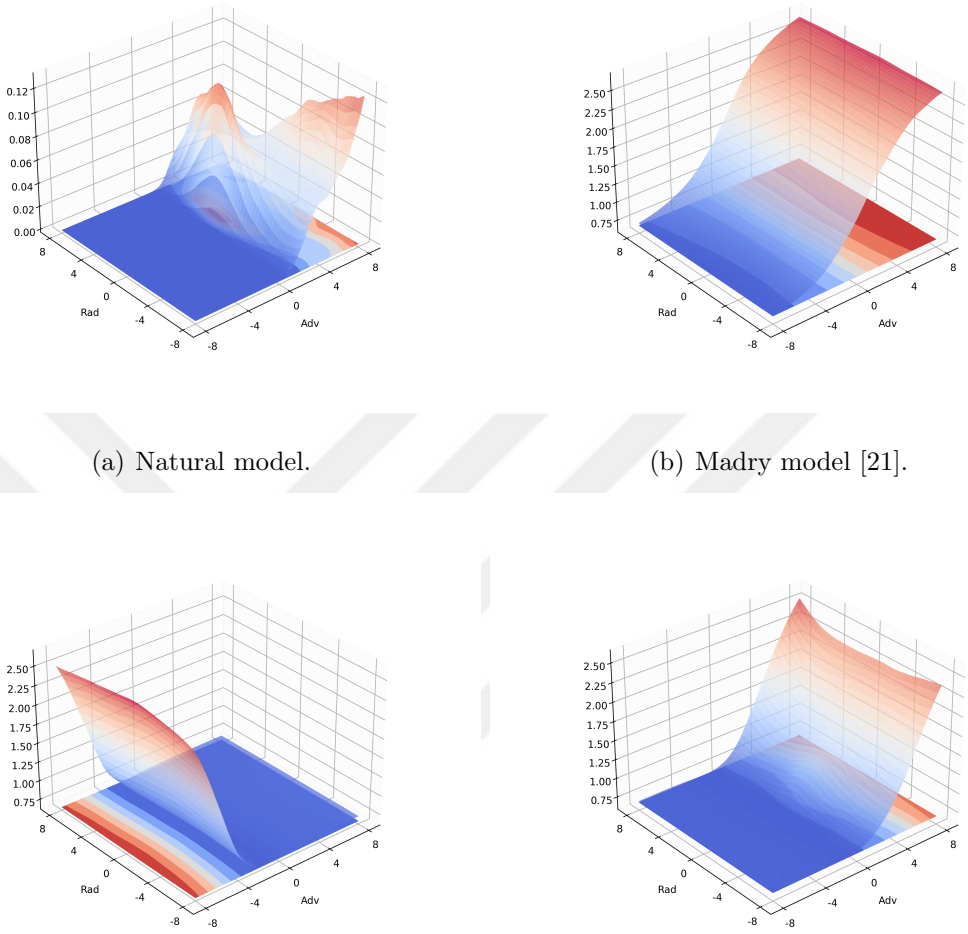
Table 4.9. Adversarial robustness performance of various perturbation augmentation permutations.

Augmentations	**Natural**	**FGSM**	**PGD10**	**PGD20**	**CW10**	**CW20**
δ_{rand}	94.51	93.28	73.93	68.16	69.11	64.36
$\delta_{FS}\delta_{rand}$	92.98	92.54	68.47	62.83	63.70	58.6
$\delta_{FS}\delta_{rand}\delta_{rand}$	92.44	89.27	65.52	59.74	63.08	57.40
$\delta_{FS}\delta_{rand}\delta_{45}$	93.68	94.33	73.58	68.08	66.58	61.88

4.4.6. Visualization of the Loss Landscape

Figure 4.2 consists of loss landscapes of four models against FGSM attack on one of the random images in CIFAR-10. We generated plots of the cross-entropy loss function projected onto the adversarial direction and a random (Rademacher) direction, using an ϵ value of 8. The visualizations are created based on the codebase provided by [73]. The x-axis of the plot represents the magnitude of the adversarial perturbation obtained by FGSM attack applied to the input image, while the y-axis represents the magnitude of the random Rademacher perturbation applied to the input image. A Rademacher distribution is a discrete probability distribution that takes values +1 or -1 with equal probability [74]. The color of each point on the plot represents the corresponding loss value, with darker red colors indicating higher loss values.

The loss landscape is a 2D plot that shows how the loss function changes with changes in the input data. As expected, the model trained with natural images has a complex loss landscape, as seen in Figure 4.2(a). On the other hand, AT models have similar loss landscapes with minor differences in terms of the loss amount. Our analysis shows that the models are vulnerable to adversarial attacks when the adversarial perturbation is large, while the random Rademacher perturbation have a more noticeable effect on the network's performance, but not as significant as the impact of the adversarial attack.



(a) Natural model.

(b) Madry model [21].

(c) Feature Scatter model [34].

(d) PDA-AT.

Figure 4.2. Loss landscapes of proposed and baselines. The visualizations are created based on the codebase provided by [73].

4.5. Results of Shift-Invariance Approaches

The shift-invariance approaches in DNNs models have been shown to enhance generalization by reducing vulnerability to small shifts and transitions. To further improve the robustness of models, we propose the integration of shift-invariance techniques with AT based on the Feature Scatter. Specifically, we adopt models enhanced with low-pass filtering (LPF) and adaptive polyphase sampling (APS) during the AT

training process. In order to evaluate the efficacy of our proposed models against white-box attacks, we conducted a series of experiments using ResNet-18 with three different padding types and sizes, namely circular, reflect, and zeros padding with padding sizes of 3, and 5. We also trained a base AT model without the application of any shift-invariance techniques to analyze the independent impact of shift-invariance in AT. The results of these experiments provide valuable insights into the effectiveness of our proposed models in combating adversarial attacks.

4.5.1. Results of LPF Models

Table 4.10 displays the results of performance evaluations for shift-invariant models utilizing anti-aliasing. In Table 4.10, the model names are listed in the first column, followed by the size of the blurring kernel used for LPF indicated by a hyphen next to LPF. The second column displays the padding type employed in the model architecture. The natural and adversarial accuracies of each model are presented in the remaining columns of the table.

The evaluations carried out have indicated that circular padding adversely affects the generalization performance and robustness of the models, as evidenced by the reduction in adversarial accuracy. Conversely, the models that utilized LPF with a blurring kernel of size 3 and 5, along with zero and reflect padding, displayed slightly improved performance. However, due to the random initialization of the attacks, it is not conclusive that this method effectively enhances generalization performance. As a potential future work, the evaluation experiment can be repeated with more times, and the mean natural and adversarial accuracies can be calculated, along with their standard deviation, to provide more evidence of any potential improvements.

4.5.2. Results of APS Models

The performance evaluation results of shift-invariant models using APS and anti-aliasing are presented in Table 4.11. The first column of the table specifies the model

Table 4.10. Adversarial accuracy comparison of adversarially trained shift-invariant models by LPS.

	Padding	**Natural**	**FGSM**	**PGD20**	**PGD100**	**CW20**	**CW100**
Base	zeros	88.68	69.51	51.92	47.54	47.52	43.9
LPF-3	zeros	90.5	69.98	52.94	49.23	49.77	46.24
LPF-3	reflect	89.62	68.52	49.66	45.37	49.05	45.86
LPF-3	circular	89.08	66.13	46.63	42.76	44.92	41.31
LPF-5	zeros	90.05	70.24	53.69	49.68	49.2	45.35
LPF-5	reflect	89.35	69.1	47.92	42.92	46.09	42.23
LPF-5	circular	89.27	66.16	48.09	44.12	45.56	41.93

that was evaluated. If the APS shift-invariance method was combined with a low-pass filtering kernel, the kernel size is indicated next to APS with a hyphen. The second column of the table provides information about the padding type used in the model architecture. The natural and adversarial accuracies of the models are listed in the remaining columns. Notably, the APS shift-invariance method can be combined with LPF, as suggested in the original study, to enhance generalization.

Based on the evaluations conducted, it is apparent that circular padding harms the generalization performance and robustness by causing a decrease in natural accuracy and adversarial accuracy. On the other hand, the models that employed APS with a blurring kernel of size 3 showed marginally better performance. Yet, this improvement is not sufficient to conclude that the method enhances the generalization performance due to the random initialization of the attacks. As a future work, the evaluation experiment can be repeated at least 10 times, and mean natural and adversarial accuracies can be calculated with their standard deviation to show that there is an improvement.

Table 4.11. Adversarial accuracy comparison of adversarially trained shift-invariant models with APS and anti-aliasing.

Model	**Padding**	**Natural**	**FGSM**	**PGD20**	**PGD100**	**CW20**	**CW100**
Base	zeros	88.68	69.51	51.92	47.54	47.52	43.9
APS	zeros	88.91	69.44	51.79	46.96	45.53	41.14
APS	reflect	88.94	68.68	50.23	44.77	43.91	39.43
APS	circular	88.1	65.61	46.79	42.4	42.61	38.81
APS-3	zeros	90.09	70.45	53.96	48.8	49.6	44.92
APS-3	reflect	89.63	69.69	51.01	46.13	48.32	44.32
APS-3	circular	88.97	65.69	46.19	41.49	44.24	40.83
APS-5	zeros	89.89	70.57	53.53	48.7	49.43	44.46
APS-5	reflect	89.17	68.73	49.3	44.53	48.15	44.28
APS-5	circular	89.19	66.31	47.63	42.27	44.48	40.74

4.6. Results of Mixup-based Approaches

While AT techniques have proven highly effective in combating adversarial attacks, their limitations have become apparent. Therefore, it is crucial to investigate alternative methods that can contribute to a deeper understanding of adversarial attacks and enhance model robustness. We investigate the mixup-based data augmentation utilization for AT techniques through various experiments. Mixup employs convex combinations of data and their corresponding labels to effectively enhance a model's generalization abilities, as demonstrated in Equation (2.5). In that regard, we developed four distinct approaches to integrate mixup data augmentation with AT based on the FS. The WideResNet-18 is trained with those approaches on CIFAR-10.

The developed ideas are evaluated against various white-box attacks, which have been thoroughly discussed in the preceding sections. The results of these experiments are presented in Table 4.12, providing a summary of the adversarial accuracies achieved by our proposed methodologies. The results of the model, which is trained solely with

FS [34] approach, are also provided to clearly demonstrate the impact of the proposed ideas.

Table 4.12. Adversarial accuracy comparison of proposed models enhanced with mixup methodology on CIFAR-10 dataset.

	Natural	FGSM	PGD10	PGD20	PGD100	CW10	CW20	CW100
FS	90.3	77.31	70.79	67.6	64.8	64.04	60.82	58.43
Mixup + FS	94.13	92.68	18.86	9.31	2.65	13.25	7.31	2.95
FS + Mixup	91.83	73.41	51.53	43.88	37.83	45.54	39.86	35.35
Mixup in OT	90.65	77.96	70.82	67.87	65.23	60.2	57.35	55.37
Mixup in OT & CE	92.7	65.42	34.69	26.77	22.45	32.58	27.56	24.55

It is evident that the implementation of Mixup followed by FS can lead to catastrophic overfitting and label leaking by demonstrating robustness only against a certain attack. Among the proposed approaches, only utilizing mixup in OT distance loss calculation have marginally better adversarial accuracies against FGSM and PGD variants. Consequently, our findings suggest that the developed approaches, which use interpolated inputs in different steps of the Feature Scattered based AT scheme, fail to boost the robustness and generalization, and also, in most cases, they cannot achieve the state-of-the-art level.

5. CONCLUSION

DNNs have been demonstrated to be susceptible to adversarial attacks, which pose a significant threat to their trustworthiness and security. Adversarial training is a widely used technique to improve the robustness of DNNs against such attacks. This thesis presents three new ideas for adversarial training aimed at enhancing the robustness of DNNs against adversarial attacks. These ideas can be summarized as:

- The PDA-AT framework, which involves the acquisition of alternative adversarial directions within a region that spans a 45-degree angle from the principal adversarial direction. This direction is generated by the gradient of the optimal transport distance between perturbed mini-batch samples.
- The integration of LPF and APS shift-invariant approaches into the AT framework and investigation of the performance effect in terms of generalization and robustness.
- Employment of mixup-based data augmentation into AT and an analysis of its effectiveness in enhancing the robustness and generalization of AT.

The proposed PDA-AT framework generates the principal adversarial direction based on the increase in optimal transport distance examples found in perturbed mini-batches, which deviates from the traditional Feature Scatter AT [34] method that takes into account the inter-sample relationship between natural and randomly perturbed images within a batch. We empirically show that the proposed primary attack generation expands the diversity of attacks to be used in AT scheme. Our experiments show that this PDA-AT framework enhances the robustness of DNNs against adversarial attacks, and the performance is further improved by augmenting the adversarial directions. We present extensive experimental results on CIFAR-10, CIFAR-100, SHVN, GTSRB, and Tiny-Imagenet datasets against various well-known white-box and black-box attacks. As a result, the proposed attack generation method without the perturbation augmentation achieves to boost the robustness of datasets that have a limited num-

ber of classes. Moreover, the perturbation augmentation in the PDA-AT framework has demonstrated a substantial enhancement in the adversarial robustness of DNNs in specific scenarios. The performance gap between the natural accuracy and robust accuracy against adversarial samples generated through the FGSM attack method has been considerably diminished.

Furthermore, this thesis investigates the performance of two shift-invariant approaches, namely LPF [35] and APS [36], integrated into the AT framework. These approaches aim to enhance the generalization of traditional AT, but their effectiveness is limited in certain configurations. Specifically, among the experiments, we observed that the anti-aliased network with blurring kernel sizes 3 and 5 with zero padding showed only minor improvement. In addition, mixup-based [33] data augmentation is employed to improve the generalization and robustness of AT. However, the experimental results demonstrate that these approaches fall short of achieving state-of-the-art performance levels.

Although this thesis has proposed and evaluated three new ideas for enhancing the robustness and generalization of AT, there is still room for improvement and further investigation. One possible area for future work could be to integrate the perturbation augmentation operation of the PDA-AT framework into other AT methods to determine its usefulness as a general base. Also, while the optimal transport distance has been employed in the proposed PDA-AT framework, it is conceivable that other metrics could yield even more diverse and effective adversarial directions.

The integration of shift-invariant approaches into the AT framework might have promise for improving generalization, but their effectiveness with LPF and APS approaches is limited in certain configurations. Investigating alternative shift-invariant approaches or adapting these methods to other AT techniques could be a better suit for future research. Additionally, while mixup-based data augmentation has demonstrated effectiveness in natural training, further research could explore ways to integrate it into the AT framework. Furthermore, the scope of our investigation only covered the tradi-

tional mixup method. However, other mixup techniques, including adversarial vertex mixup, cut-mix, and manifold mixup, have shown promise and could potentially be integrated into various AT schemes to enhance their robustness and generalization capabilities further. These future research directions have the potential to contribute significantly to improving the robustness and security of DNNs.

REFERENCES

1. Krizhevsky, A., I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Vol. 1, Red Hook, NY, USA, pp. 1097–1105, 2012.

2. Hannun, A., C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates and A. Y. Ng, "Deep Speech: Scaling Up End-to-end Speech Recognition", ArXiv:1412.5567 [cs], 2014.

3. Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is All You Need", *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, pp. 6000–6010, 2017.

4. Bojarski, M., D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao and K. Zieba, "End to End Learning for Self-driving Cars", ArXiv:1604.07316 [cs], 2016.

5. Gulshan, V., L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega and D. R. Webster, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs", *JAMA*, Vol. 316, No. 22, pp. 2402–2410, 2016.

6. Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge", *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211–252, 2015.

7. Le, Y. and X. Yang, "Tiny Imagenet Visual Recognition Challenge", *CS 231N*, Vol. 7, No. 7, p. 3, 2015.

8. Geiger, A., P. Lenz, C. Stiller and R. Urtasun, "Vision meets robotics: The KITTI dataset", *The International Journal of Robotics Research*, Vol. 32, No. 11, pp. 1231–1237, 2013.

9. Simonyan, K. and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", *3rd International Conference on Learning Representations,*, San Diego, CA, USA, pp. 1–14, 2015.

10. He, K., X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 770–778, 2016.

11. Tan, M. and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97, Long Beach, California, USA, pp. 6105–6114, 2019.

12. Kingma, D. P. and J. Ba, "Adam: A Method for Stochastic Optimization", *3rd International Conference on Learning Representations*, San Diego, CA, USA, pp. 1–15, 2015.

13. You, Y., J. Li, S. J. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer and C. Hsieh, "Large Batch Optimization for Deep Learning: Training BERT in 76 minutes", *8th International Conference on Learning Representations, ICLR*, Addis Ababa, Ethiopia, p. 962, 2020.

14. Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow and R. Fergus, "Intriguing Properties of Neural Networks", *2nd International Conference on Learning Representations*, Banff, AB, Canada, p. 1312, 2014.

15. Goodfellow, I. J., J. Shlens and C. Szegedy, "Explaining and Harnessing Adver-

sarial Examples", *3rd International Conference on Learning Representations*, San Diego, CA, USA, p. 1412, 2015.

16. Carlini, N. and D. Wagner, "Towards Evaluating the Robustness of Neural Networks", *2017 IEEE Symposium on Security and Privacy (SP)*, Los Alamitos, CA, USA, pp. 39–57, 2017.

17. Papernot, N., P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik and A. Swami, "Practical Black-Box Attacks against Machine Learning", *Proceedings of the ACM on Asia Conference on Computer and Communications Security*, Abu Dhabi, United Arab Emirates, pp. 506–519, 2017.

18. Ma, X., Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey and F. Lu, "Understanding Adversarial Attacks on Deep Learning based Medical Image Analysis Systems", *Pattern Recognition*, Vol. 110, No. 0031-3203, p. 107332, 2021.

19. Eykholt, K., I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno and D. Song, "Robust Physical-World Attacks on Deep Learning Visual Classification", *Computer Vision and Pattern Recognition*, Salt Lake City, Utah, pp. 1625–1634, 2018.

20. Fursov, I., M. Morozov, N. Kaploukhaya, E. Kovtun, R. Rivera-Castro, G. Gusev, D. Babaev, I. Kireev, A. Zaytsev and E. Burnaev, "Adversarial Attacks on Deep Models for Financial Transaction Records", ArXiv:2106.08361 [cs], 2021.

21. Madry, A., A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks", *6th International Conference on Learning Representations*, Vancouver, BC, Canada, p. 1706, 2018.

22. Wang, Y., D. Zou, J. Yi, J. Bailey, X. Ma and Q. Gu, "Improving Adversarial Robustness Requires Revisiting Misclassified Examples", *8th International Conference on Learning Representations*, Addis Ababa, Ethiopia, p. 2201, 2020.

23. Wong, E., L. Rice and J. Z. Kolter, "Fast is Better than Free: Revisiting Adversarial Training", *8th International Conference on Learning Representations*, Addis Ababa, Ethiopia, p. 2001, 2020.

24. Shaham, U., Y. Yamada and S. Negahban, "Understanding Adversarial Training: Increasing Local Stability of Supervised Models Through Robust Optimization", *Neurocomputing*, Vol. 307, No. 0925-2312, pp. 195–204, 2018.

25. Jang, Y., T. Zhao, S. Hong and H. Lee, "Adversarial Defense via Learning to Generate Diverse Attacks", *CVF International Conference on Computer Vision*, Seoul, Korea, pp. 2740–2749, 2019.

26. Kurakin, A., I. J. Goodfellow and S. Bengio, "Adversarial Machine Learning at Scale", *5th International Conference on Learning Representations*, Toulon, France, p. 1611, 2017.

27. Athalye, A., N. Carlini and D. A. Wagner, "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples", ArXiv:1802.00420 [cs], 2018.

28. Andriushchenko, M. and N. Flammarion, "Understanding and Improving Fast Adversarial Training", *Neural Information Processing Systems*, Vancouver, Canada, pp. 16048–16059, 2020.

29. Rice, L., E. Wong and J. Z. Kolter, "Overfitting in Adversarially Robust Deep Learning", *International Conference on Machine Learning*, Vienna, Austria, pp. 8093–8104, 2020.

30. Shafahi, A., M. Najibi, A. Ghiasi, Z. Xu, J. P. Dickerson, C. Studer, L. S. Davis, G. Taylor and T. Goldstein, "Adversarial Training for Free!", *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, pp. 3358–3369, 2019.

31. Ilyas, A., S. Santurkar, D. Tsipras, L. Engstrom, B. Tran and A. Madry, "Adversarial Examples Are Not Bugs, They Are Features", *Neural Information Processing Systems*, Vancouver, BC, Canada, pp. 125–136, 2019.

32. Poursaeed, O., T. Jiang, H. Yang, S. J. Belongie and S.-N. Lim, "Robustness and Generalization via Generative Adversarial Training", *International Conference on Computer Vision*, Montreal, Canada, pp. 15691–15700, 2021.

33. Zhang, H., M. Cissé, Y. N. Dauphin and D. Lopez-Paz, "Mixup: Beyond Empirical Risk Minimization", *6th International Conference on Learning Representations*, Vancouver, BC, Canada, pp. 1–13, 2018.

34. Zhang, H. and J. Wang, "Defense Against Adversarial Attacks Using Feature Scattering-based Adversarial Training", *Advances in Neural Information Processing Systems*, Vancouver, British Columbia, Canada, pp. 1831–1841, 2019.

35. Zhang, R., "Making Convolutional Networks Shift-Invariant Again", *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, CA, USA, pp. 7324–7334, 2019.

36. Chaman, A. and I. Dokmanić, "Truly Shift-Equivariant Convolutional Neural Networks with Adaptive Polyphase Upsampling", *55th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, pp. 3773–3783, 2021.

37. Akhtar, N. and A. Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey", ArXiv:1801.00553 [cs], 2018.

38. Ren, K., T. Zheng, Z. Qin and X. Liu, "Adversarial Attacks and Defenses in Deep Learning", *Engineering*, Vol. 6, No. 3, pp. 346–360, 2020.

39. Huang, Z., Y. Fan, C. Liu, W. Zhang, Y. Zhang, M. Salzmann, S. Süsstrunk and J. Wang, "Fast Adversarial Training with Adaptive Step Size", ArXiv:2206.02417 [cs], 2022.

40. Lippe, P., "UvA Deep Learning Tutorials", 2022, `https://uvadlc-notebooks.readthedocs.io/en/latest/`, accessed on March 21, 2023.

41. Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library", *Advances in Neural Information Processing Systems 32*, Vancouver, BC, Canada, pp. 8026–8037, 2019.

42. Croce, F. and M. Hein, "Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-Free Attacks", *Proceedings of the 37th International Conference on Machine Learning*, Vienna, Austria, pp. 2206–2216, 2020.

43. Croce, F. and M. Hein, "Minimally Distorted Adversarial Examples with a Fast Adaptive Boundary Attack", *Proceedings of the 37th International Conference on Machine Learning*, Vienna, Austria, pp. 2196–2205, 2020.

44. Andriushchenko, M., F. Croce, N. Flammarion and M. Hein, "Square Attack: A Query-Efficient Black-Box Adversarial Attack via Random Search", *European Conference on Computer Vision*, Glasgow, UK, pp. 484–501, 2020.

45. Jeddi, A., M. J. Shafiee, M. Karg, C. Scharfenberger and A. Wong, "Learn2Perturb: an End-to-end Feature Perturbation Learning to Improve Adversarial Robustness", *Proceedings of the CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 1238–1247, 2020.

46. Lee, S., H. Lee and S. Yoon, "Adversarial Vertex Mixup: Toward Better Adversarially Robust Generalization", *2020 CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 269–278, 2020.

47. İnci M. Baytaş and D. Deb, "Robustness-via-Synthesis: Robust Training with

Generative Adversarial Perturbations", *Neurocomputing*, Vol. 516, No. C, pp. 49–60, 2023.

48. Kim, H., W. Lee and J. Lee, "Understanding Catastrophic Overfitting in Single-step Adversarial Training", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vancouver, Canada, pp. 8119–8127, 2021.

49. Sriramanan, G., S. Addepalli, A. Baburaj and V. B. R., "Guided Adversarial Attack for Evaluating and Enhancing Adversarial Defenses", *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada, pp. 20297–20308, 2020.

50. Wang, H. and C.-N. Yu, "A Direct Approach to Robust Deep Learning Using Adversarial Networks", ArXiv:1905.09591 [cs], 2019.

51. Verma, V., A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz and Y. Bengio, "Manifold Mixup: Better Representations by Interpolating Hidden States", *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, CA, USA, pp. 6438–6447, 2019.

52. Yun, S., D. Han, S. J. Oh, S. Chun, J. Choe and Y. Yoo, "CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features", *International Conference on Computer Vision*, Seoul, Korea, pp. 6022–6031, 2019.

53. DeVries, T. and G. W. Taylor, "Improved Regularization of Convolutional Neural Networks with Cutout", ArXiv:1708.04552 [cs], 2017.

54. Zhang, H. and W. Xu, "Adversarial Interpolation Training: A Simple Approach for Improving Model Robustness", 2020, `https://openreview.net/forum?id=Syejj0NYvr`, accessed on February 21, 2023.

55. Wu, D., S.-T. Xia and Y. Wang, "Adversarial Weight Perturbation Helps Robust Generalization", *Advances in Neural Information Processing Systems*, Vancouver,

British Columbia, Canada, pp. 2958–2969, 2020.

56. Wang, J., "Bilateral Adversarial Training: Towards Fast Training of More Robust Models Against Adversarial Attacks", *CVF International Conference on Computer Vision*, Munich, Germany, pp. 6628–6637, 2018.

57. Engstrom, L., D. Tsipras, L. Schmidt and A. Madry, "A Rotation and a Translation Suffice: Fooling CNNs with Simple Transformations", ArXiv:1712.02779 [cs], 2017.

58. Azulay, A. and Y. Weiss, "Why Do Deep Convolutional Networks Generalize So Poorly to Small Image Transformations?", ArXiv:1805.12177 [cs], 2018.

59. Arjovsky, M., S. Chintala and L. Bottou, "Wasserstein Generative Adversarial Networks", *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, pp. 214–223, 2017.

60. Xie, Y., X. Wang, R. Wang and H. Zha, "A Fast Proximal Point Method for Computing Exact Wasserstein Distance", *Conference on Uncertainty in Artificial Intelligence*, New Orleans, Louisiana, USA, pp. 433–453, 2018.

61. Villani, C., *Optimal Transport: Old and New*, Springer, Berlin Heidelberg, 2008.

62. Salimans, T., H. Zhang, A. Radford and D. N. Metaxas, "Improving GANs Using Optimal Transport", *6th International Conference on Learning Representations*, Vancouver, BC, Canada, p. 1803, 2018.

63. Cuturi, M., "Sinkhorn Distances: Lightspeed Computation of Optimal Transport", *Advances in Neural Information Processing Systems*, Lake Tahoe, Nevada, United States, pp. 2292–2300, 2013.

64. Solomon, J., "Optimal Transport on Discrete Domains", ArXiv:1801.07745 [math.OC], 2018.

65. Krizhevsky, A. and G. Hinton, *Learning Multiple Layers of Features from Tiny Images*, Tech. rep., University of Toronto, Toronto, Ontario, 2009.

66. Netzer, Y., T. Wang, A. Coates, A. Bissacco, B. Wu and A. Y. Ng, "Reading Digits in Natural Images with Unsupervised Feature Learning", *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, Granada, Spain, pp. 12–17, 2011.

67. Stallkamp, J., M. Schlipsing, J. Salmen and C. Igel, "The German Traffic Sign Recognition Benchmark: A Multi-Class Classification Competition", *Proceedings of the International Joint Conference on Neural Networks*, California, USA, pp. 1453–1460, 2011.

68. Zagoruyko, S. and N. Komodakis, "Wide Residual Networks", *Proceedings of the British Machine Vision Conference*, York, UK, pp. 1–12, 2016.

69. Zhang, H., "Feature Scatter", 2018, `https://github.com/Haichao-Zhang/FeatureScatter`, accessed on November 15, 2022.

70. Chaman, A., "Truly Shift-Invariant Convolutional Neural Networks", 2021, `https://github.com/achaman/truly_shift_invariant_cnns`, accessed on February 1, 2023.

71. Zhang, R., "Making Convolutional Networks Shift-Invariant Again", 2021, `https://github.com/adobe/antialiased-cnns`, accessed on February 10, 2023.

72. facebookresearch, "Mixup-CIFAR10", 2018, `https://github.com/facebookresearch/mixup-cifar10`, accessed on January 5, 2023.

73. Yeong, P. G., "Reliably Fast Adversarial Training via Latent Adversarial perturbation", 2021, `https://github.com/ParkGeonYeong/SLAT`, accessed on March 3, 2023.

74. Hitczenko, P. and S. Kwapień, "On the Rademacher Series", *Probability in Banach Spaces*, Boston, MA, USA, pp. 31–36, 1994.