HOKKAIDO UNIVERSITY

Title	A study on deep learning-based adversarial defense by introducing robust features
Author(s)	張, 加煥
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第16012号
Issue Date	2024-03-25
DOI	https://doi.org/10.14943/doctoral.k16012
Doc URL	https://hdl.handle.net/2115/92322
Type	doctoral thesis
File Information	Zhang_Jiahuan.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

A thesis for the degree of Doctor of Philosophy

A Study on Deep Learning-based Adversarial Defense by Introducing Robust Features

ロバスト特徴の導入による深層学習ベースの敵対的防御に関する研究

Jiahuan Zhang

Graduate School of Information Science and Technology

Hokkaido University

March, 2024

Contents

1 Introduction **9**

1.1 Background 9

1.2 Proposition in This Thesis 11

1.3 Organization of This Thesis 12

2 Related Works **15**

2.1 Introduction 15

2.2 Adversarial Attack & Defense for Image Classification 15

2.2.1 Adversarial Attack for Image Classification 15

2.2.2 Adversarial Defense for Image Classification 21

2.2.2.1 Adversarial Training 21

2.2.2.2 Adversarial Example Detection 23

2.2.2.3 Processing of Input Data 25

2.2.2.4 Regularization 26

2.3 Adversarial Attack for Image Captioning 27

2.4 Conclusions 30

3 Integration of Robust Regularization-based Enhanced Multi-stage Fusion Features into Image Classification System **32**

3.1 Introduction 32

3.2 Methodology 33

3.2.1 Multi-Stage Features Enhancement (MSFE) 33

3.2.2 Multi-Stage Features Fusion (MSF^2) 34

3.2.3 Regularization 37

3.3 Dataset & Attack Methods 38

3.3.1 Dataset: CIFAR-10 38

3.3.2 Adversarial Attacks 38

3.4 Comparison Experiments 39

3.4.1 Comparison Methods 39

3.4.2 Experimental Results under ℓ_∞-norm Attacks 41

3.4.3 Experimental Results under ℓ_2-norm Attacks . . . 42
3.5 Ablation Studies . . . 47
3.5.1 Performance on Baseline, MSF^{2}Net, and EMSF^{2}Net . . . 47
3.5.2 Performance on Each Class of CIFAR-10 . . . 48
3.5.3 Visualization by Grad-CAM . . . 53
3.5.4 Visualization by t-SNE . . . 54
3.6 Conclusions . . . 61

4 Integration of Robust Caption Features from Stable RNN into Image Captioning System 63
4.1 Introduction . . . 63
4.2 Proposed Approach . . . 64
4.2.1 Overall Framework . . . 64
4.2.2 Details of LRNN . . . 65
4.2.3 Gate LRNN . . . 67
4.3 Experiments . . . 67
4.3.1 Dataset & Attack Method . . . 67
4.3.2 Baseline . . . 68
4.3.3 Evaluation Metrics . . . 68
4.3.4 Experimental Results . . . 70
4.4 Conclusions . . . 70

5 Integration of Robust Caption Features from Pre-trained Models into Image Classification System 71
5.1 Introduction . . . 71
5.2 Proposed Method . . . 74
5.2.1 Extraction of Heterogeneous Features . . . 74
5.2.2 Calculation of Multi-purpose Loss Functions . . . 76
5.3 Experiments . . . 79
5.3.1 Experimental Settings . . . 79
5.3.1.1 Datasets . . . 79
5.3.1.2 Attack Settings . . . 79
5.3.1.3 Comparison Methods . . . 80
5.3.1.4 Ablation Methods . . . 80
5.3.1.5 Parameter Settings . . . 81
5.3.2 Results and Discussion . . . 81

5.4 Conclusions 82

6 Conclusions **91**

6.1 Summary of This Thesis 91

6.2 Future Research Directions 92

Acknowledgements **94**

Bibliography **95**

Achievements of the Author **108**

List of Figures

2.1 A clean image from the dataset CIFAR-10 and its corresponding adversarial example. 16

2.2 A map of related works shown in Chapter 2. 31

3.1 The framework of EMSF^{2}Net. Serving as its backbone component, ResNet50 [1] is adopted. STAGEs 0 to 4 in this figure and those in the baseline share the same structures, and the structural specifics of STAGEs 0 to 4 are detailed in Table 3.1. Furthermore, the block of feature enhancement is denoted by FE Block, the operation of concatenation is signified by Concat, and within the FE Block, the global average pooling is represented by GAP. 35

3.2 Some visual examples of each class in the CIFAR-10 dataset. 40

3.3 Results of the robust accuracy of the baseline, MSF^{2}Net, and EMSF^{2}Net. The adversarial attacks utilized were set as ℓ_∞-norm, with the evaluation on the dataset of CIFAR-10. 49

3.4 Results of the robust accuracy of the baseline, MSF^{2}Net, and EMSF^{2}Net. The adversarial attacks utilized were set as ℓ_2-norm, with the evaluation on the dataset of CIFAR-10. 50

3.5 The accuracy results per class without any attacks and under the adversarial attacks with ℓ_∞-norm, with the evaluation of baseline, MSF^{2}Net, and EMSF^{2}Net on the dataset of CIFAR-10. In these sub-figures, the X-axis denotes the robust accuracy, while the Y-axis indicates the respective class name within the CIFAR-10 dataset. 51

3.6 The accuracy results per class under the adversarial attacks with ℓ_2-norm, with the evaluation of baseline, MSF^{2}Net, and EMSF^{2}Net on the dataset of CIFAR-10. In these sub-figures, the X-axis denotes the robust accuracy, while the Y-axis indicates the respective class name within the CIFAR-10 dataset. 52

3.7 Visualization results using Grad-CAM, with the utilization of the ℓ_∞-norm adversarial attacks for baseline, MSF^{2}Net, and EMSF^{2}Net. Original data was selected from CIFAR-10. The darker red areas indicate that the models pay more attention to those parts. 55

3.8 Visualization results using Grad-CAM, with the utilization of the ℓ_2-norm adversarial attacks for baseline, MSF^{2}Net, and EMSF^{2}Net. Original data was selected from CIFAR-10. The darker red areas indicate that the models pay more attention to those parts. 56

3.9 Visualization results using t-SNE, with the condition of no attacks for the baseline, MSF^{2}Net, and EMSF^{2}Net. Experiments were performed on CIFAR-10. . . . 57

3.10 Visualization results using t-SNE, with the utilization of the less complex ℓ_∞-norm attacks for the baseline, MSF^{2}Net, and EMSF^{2}Net. Experiments were performed on CIFAR-10. 58

3.11 Visualization results using t-SNE, with the utilization of the more complex ℓ_∞-norm attacks for the baseline, MSF^{2}Net, and EMSF^{2}Net. Experiments were performed on CIFAR-10. 59

3.12 Visualization results using t-SNE, with the utilization of the ℓ_2-norm attacks for the baseline, MSF^{2}Net, and EMSF^{2}Net. Experiments were performed on CIFAR-10. 60

4.1 The proposed retrieval-based image captioning system, with the design of implementing the defense against the adversarial attack with noun type. In addition, the visual semantic embedding is represented by VSE in the figure. 66

5.1 The outputs of ResNet-50 and ClipCap when the inputs are adversarial examples. 73

5.2 Framework of HFFN, where Up. signifies the upsampling operation, Concat denotes the operation of concatenation, and $\ell_2(\cdot)$ corresponds to ℓ_2 normalization. 75

5.3 Accuracy results among HFFN and four ablation methods on the CIFAR-10 dataset. 87

List of Tables

3.1 Details of the baseline architecture. Specifically, $\mathcal{L}_{PC}$ denotes the regularization loss function, whereas the classic cross-entropy loss is represented by $\mathcal{L}_{CE}$. . . . 36

3.2 Accuracy results in clean and FGSM attack scenarios on the CIFAR-10 dataset, with the rightmost column displaying accuracies from the proposed method. In addition, the top-performing results are highlighted in bold. 43

3.3 Results of robust accuracy under the less complex iteration-based adversarial attacks with ℓ_∞-norm on the dataset of CIFAR-10. The top-performing results are highlighted in bold. 44

3.4 Results of robust accuracy under the more complex iteration-based adversarial attacks with ℓ_∞-norm on the dataset of CIFAR-10. The top-performing results are highlighted in bold. 45

3.5 Results of robust accuracy under adversarial attacks with ℓ_2-norm on the dataset of CIFAR-10. ε represents attack strength, IN denotes iteration number (*steps*), and the top-performing results are highlighted in bold. 46

4.1 Some original captions from the dataset of MS-COCO and some adversarial examples (captions) generated using the noun-type attack method. The words in bold represent the original key nouns in the original caption, while the words in both bold and italic indicate the adversarial key nouns resulting from the noun-type attack on the original caption. 69

4.2 Experimental results of the proposed method (PM) and the baseline, where PM represents the retrieval-based image captioning system with the Gate LRNN, while baseline represents the retrieval-based image captioning system with the modified vanilla RNN. 70

5.1 Accuracy results of HFFN and three comparison methods in clean and FGSM attack scenarios on the CIFAR-10 dataset. 83
5.2 Accuracy results among HFFN and comparison methods in the scenarios of PGD^{10}-ℓ_∞ and PGD^{20}-ℓ_∞ attacks on the CIFAR-10 dataset. 84
5.3 Accuracy results among HFFN and comparison methods in the scenarios of MI-FGSM^{20} and Jitter^{40} attacks on the dataset CIFAR-10. 85
5.4 Accuracy results among HFFN and comparison methods in the scenarios of PGD^{20}-ℓ_2 and PGD^{40}-ℓ_2 attacks on the dataset CIFAR-10. 86
5.5 Accuracy results of HFFN, comparison methods, and ablation approaches in clean and FGSM attack scenarios on the CIFAR-100 dataset. 88
5.6 Accuracy results of HFFN, comparison methods, and ablation approaches in the scenarios of the PGD^{10}-ℓ_∞, PGD^{20}-ℓ_∞, MI-FGSM^{20}, and Jitter^{40} attacks on the CIFAR-100 dataset. 89
5.7 Accuracy results of HFFN, comparison methods, and ablation approaches in the scenarios of the PGD^{20}-ℓ_2 and PGD^{40}-ℓ_2 attacks on the CIFAR-100 dataset. . . . 90

Chapter 1

Introduction

The research background about this thesis, the proposition in this thesis, and the organization of this thesis are shown in this chapter.

1.1 Background

In recent years, deep learning technologies [1–17] have emerged as dominant forces in the realm of artificial intelligence. However, current research indicates that the prevalent deep learning-based architectures represented by image classification and image captioning systems exhibit the significant susceptibility to the malicious data represented by adversarial examples. Adversarial examples, which are barely noticeable to the human, can lead deep learning models to make erroneous decisions. In addition, the utilization of these adversarial examples by an attacker to attack the image classification and image captioning systems could lead to the severe repercussions. Consequently, the emergence of adversarial attacks [18–42] has developed into a significant impediment in the development of reliable image classification and image captioning systems. Conclusively, it has become increasingly urgent to construct corresponding defense methods for these systems to enable them to defend against attacks from adversarial examples.

Numerous contributions have been made in the field of adversarial defense for the image classification system. The first is adversarial training-based approaches [43–51]. These methods can introduce slight perturbations to the input images throughout the training phase, compelling the image classification system to adapt them and thereby augmenting its defense performance against adversarial attacks. However, the approaches based on adversarial training

are incapable of defending against widespread adversarial attacks and require substantial time investment throughout the training phase. The second is adversarial example detection-based approaches [52–60]. In the majority of instances, these approaches primarily concentrate on developing the architecture that identifies whether input images contain adversarial perturbations, as opposed to classifying them into correct classes. Thirdly, some researchers concentrate on processing the input data [61–70] to eliminate the perturbations or lessen the effects of such adversarial noises. Nonetheless, given the substantial variance in the distribution of diverse adversarial noises, it is challenging for the current deep learning-based image classification system to acquire a distribution function capable of fully addressing every form of adversarial noises. Consequently, adversarial defense methodologies for processing the input data exhibit certain limitations. Finally, it has been hypothesized by certain researchers that the subpar effectiveness of the deep learning-based image classification system against adversarial examples could be attributed to overfitting. Consequently, methods centered on regularization for adversarial defense [71–78] have been suggested as the solution to mitigate the effects of overfitting and bolster the generalization capabilities of the image classification system.

Although significant attention has been given by a number of related researchers to the security and robustness of the existing deep learning-based models, the following issues remain to be addressed.

- Firstly, within the domain of adversarial defense strategies for the image classification system, approaches centered on regularization are gaining traction owing to their efficiency and reduced computational demands. Nonetheless, these approaches lack a comprehensive analysis regarding the types of features that would be more effective for regularization to enhance adversarial robustness further. In other words, the existing regularization-centered methods do not introduce sufficiently robust features for the image classification system.

- Secondly, multimodal deep learning models are gaining significant traction for their closer approximation to reality, among which the image captioning system is a notable example. However, it has been observed that the features within the existing image captioning system lack robustness, rendering it particularly vulnerable to adversarial attacks.

- Lastly, from a holistic perspective, the majority of current adversarial defense methods for

the image classification system only utilize the information in the input image and have not incorporated the robust features from other modalities, thereby limiting the enhancements in robustness.

1.2 Proposition in This Thesis

The purpose of this thesis is to realize adversarial defense for deep learning models. To achieve this goal, this thesis proposes a method for introducing robust features that are not easily affected by adversarial examples and can ensure the system stability and effectiveness. The proposed method includes the following three aspects.

The first aspect proposes a multi-stage feature fusion network with the feature regularization operation for the image classification system, which can introduce the adversarially robust regularization-based enhanced multi-stage fusion features into this system. In this architecture, the enhanced multi-stage fusion features effectively represent and maintain the global information from each channel. Furthermore, the proposed network can further improve the regularization-based adversarial defense approaches for the image classification system.

In addition, to construct the adversarially robust image captioning system, the second aspect integrates a novel recurrent neural network (RNN) structure with robustness and stability, which is grounded in the dynamical systems theory, into the system. The recurrent unit within this novel structure demonstrates a significant ability to represent the hidden states effectively. This means that such an RNN structure can provide the system with robust caption features.

Since the importance of robust caption features for deep learning models has been demonstrated in the second aspect, naturally, the third aspect integrates the robust caption features into the image classification system and fuses them with the image features to improve the robustness of the system. To mitigate the disparity between the heterogeneous features, the hetero-center loss function, which effectively minimizes the class center distances between these modalities, is introduced. Additionally, the adversarial robustness of the image classification system can be deeply improved with the integration of the guided complement entropy loss function, which constrains the predictive probabilities on classes that do not align with the ground truth.

Conclusively, the primary focus of this thesis is the introduction of robust features into the deep learning-based models for adversarial defense, which is the novelty of this thesis.

1.3 Organization of This Thesis

The subsequent chapters of this thesis are structured as follows.

In Chapter 2, the relevant works about adversarial attack and defense approaches for the image classification system are presented, the related research on adversarial attack methods targeting the image captioning system is also introduced.

In Chapters 3, 4, and 5, where these chapters constitute the central parts of this thesis, solutions to the three problems mentioned in Section 1.1 are presented, respectively. In the following, the brief introduction to each of these three chapters is provided in sequence.

Chapter 3

In this chapter, we propose the Enhanced Multi-Stage Feature Fusion Network (EMSF^{2}Net), a novel framework for the image classification system, featuring the multi-stage fusion structure combined with the feature regularization operation. The EMSF^{2}Net framework integrates three core components: multi-stage feature enhancement (MSFE), multi-stage feature fusion (MSF^2), and the regularization operation. The MSFE part endeavors to acquire more expressive and enhanced features at every stage through the multiplication of features within every channel, while the MSF^2 part is designed to fuse the enhanced features from various stages, thereby augmenting the richness of the feature information. Furthermore, during the training phase, the regularization component effectively imposes constraints on both the original and fused features, ensuring the performance of the robustness. EMSF^{2}Net has demonstrated that incorporating the regularization operation for the enhanced multi-stage fusion feature significantly boosts the adversarial robustness of the convolutional neural network (CNN). Experimental results from applying the proposed method to the image classification dataset CIFAR-10 [79] demonstrate its considerable effectiveness and adversarially robustness in defending adversarial attacks with white-box setting.

Chapter 4

In this chapter, we propose a novel defense framework for the retrieval-based image captioning system, employing a kind of novel RNN structure with robustness and stability to defend against the image captioning attacks. This advanced RNN is developed from the

dynamical system theory, endowing its recurrent unit with a superior capacity for representing hidden states. Theoretically, this structure exhibits global exponential stability. To evaluate the robust performance of the proposed framework, experiments involving noun-type attacks are conducted on the retrieval-based image captioning system, contrasting a modified vanilla RNN with the improved RNN. The results of the experiments comprehensively demonstrate the superior effectiveness and robustness of the integration of improved RNN into the retrieval-based image captioning system.

Chapter 5

In this chapter, we introduce the heterogeneous feature fusion network (HFFN), a potent method for adversarial defense within the realm of the image classification system. Drawing inspiration from the human ability to employ multimodal information for object perception, robust caption features are integrated into conventional CNNs, fusing with the image features. To mitigate the disparity in heterogeneous feature modalities, the hetero-center loss is introduced, serving to limit the class center distances between different modalities. Moreover, the incorporation of guided complement entropy loss into the HFFN framework serves to limit the prediction probability for the classes that do not align with the ground truth, thereby significantly bolstering the adversarial robustness of the fused heterogeneous features. To validate the adversarial robustness and effectiveness of the HFFN framework, three advanced comparative approaches are introduced, along with the design of four distinct ablation methods. Furthermore, for a comprehensive evaluation of these methods, a substitute model is employed to produce a diverse array of adversarial examples. Conducting extensive black-box adversarial attack experiments using these adversarial examples on the CIFAR-10 and CIFAR-100 datasets [79] emphatically confirmed the superiority of the proposed HFFN framework.

To summarize, this thesis introduces three robust features for deep learning models to achieve the effective adversarial defense: the robust regularization-based enhanced multi-stage fusion features for the image classification system, the robust caption features generated from the stable RNN structure for the image captioning system, and the robust caption features generated by the pre-trained models for the image classification system.

Additionally, in Chapter 6, the conclusions derived from the methodology parts are comprehensively discussed, and the directions for the potential future investigations and studies are subsequently delineated.

In closing, the correspondence between each chapter in the methodology parts of this thesis and the research achievements of the author is presented. Specifically, Chapter 3 is constructed based on [A-1] (Sensors 2022), while Chapters 4 and 5 are formulated from [B-1] (GCCE 2021) and [B-2] (ICASSP 2023), respectively.

Chapter 2

Related Works

2.1 Introduction

In this chapter, the research relevant to this thesis in terms of the image classification system and image captioning system is presented. Specifically, the mainstream adversarial attack and defense methods for the image classification system are presented in Section 2.2. Among them, some popular adversarial attack approaches for the image classification system are shown in Subsection 2.2.1, while in Subsection 2.2.2, some adversarial defense methods for the image classification system are presented in four main directions: adversarial training, adversarial example detection, processing of input data, and regularization. In addition, the research advancements on adversarial attack methods against the image captioning system are only discussed in Section 2.3, given that there are few adversarial defense approaches developed for the image captioning system at present.

2.2 Adversarial Attack & Defense for Image Classification

2.2.1 Adversarial Attack for Image Classification

First of all, the adversarial examples within the image classification system are introduced intuitively. As mentioned in Chapter 1, the adversarial examples have the capability to easily deceive CNNs into making incorrect judgments (misclassifications for the image classification task), while they are not easily perceptible to the human eyes. Fig. 2.1 [80] illustrates both a clean image and its associated adversarial example. Where the adversarial example is generated by ResNet-50 [1] through the FGSM attack [81] with the attack strength $\varepsilon = 2/255$. This figure

reveals that, given a clean image labeled as “bird”, the CNN is capable of classifying it correctly before the addition of any adversarial noises. However, after slight adversarial perturbations are introduced to the original image, the CNN is easily fooled into misclassifying it as “dog”, even though humans can still identify it correctly.

Figure 2.1: A clean image from the dataset CIFAR-10 and its corresponding adversarial example.

Next, a succinct definition of adversarial examples in the image classification system is provided from a mathematical perspective. Given a classification model $\boldsymbol{F}$, and an original image $\boldsymbol{I}_{ori}$ accompanied by its label $\boldsymbol{y}_{ori}$, the adversarial example $\boldsymbol{I}_{adv}$ can be defined in the following equation:

$$\begin{aligned} \boldsymbol{I}_{adv} = \max_{\Delta} \quad & \mathcal{L}\left[\boldsymbol{F}\left(\boldsymbol{I}_{ori}+\Delta\right), \boldsymbol{y}_{ori}\right] \\ \text{s.t.} \quad & \mathcal{D}\left(\boldsymbol{I}_{ori}, \boldsymbol{I}_{ori}+\Delta\right) \leqslant \varepsilon, \end{aligned} \tag{2.1}$$

where $\mathcal{L}[\cdot,\cdot]$ signifies the loss function targeted for maximization, Δ is the perturbations introduced to $\boldsymbol{I}_{ori}$. Under the distance measure $\mathcal{D}(\cdot,\cdot)$, the discrepancy of the perturbed image from $\boldsymbol{I}_{ori}$ remains unobvious, not surpassing the perturbation strength ε.

Next, some commonly used adversarial attack approaches for the image classification system are introduced.

Reference [81]

Fast Gradient Sign Method (FGSM), a kind of classic adversarial attack strategy, produces perturbations by leveraging the gradient of the loss function associated with the original image $\boldsymbol{I}_{ori}$. The produced adversarial example $\boldsymbol{I}_{adv}$ can be described in the following equa-

tion:

$$\boldsymbol{I}_{adv} = \boldsymbol{I}_{ori} + \varepsilon \cdot \text{sign}\left[\nabla_{\boldsymbol{I}_{ori}} \mathcal{L}\left(\boldsymbol{F}(\boldsymbol{I}_{ori}), \boldsymbol{y}_{ori}\right)\right], \tag{2.2}$$

where ε denotes the strength of the attack, while sign$(\cdot)$ refers to the sign function and is defined as follows:

$$\text{sign}(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{if } x = 0 \cdot \\ -1, & \text{if } x < 0 \end{cases} \tag{2.3}$$

Reference [82]

Since FGSM achieves adversarial attacks by only one iteration, it cannot generate highly effective adversarial examples when facing non-linear CNN models. Therefore, Madry et al. proposed Projected Gradient Descent (PGD), an adversarial attack method based on iteration, which can also be regarded as an iterative version of FGSM. Specifically, PGD generates adversarial perturbations through multiple steps of iteration, and the perturbations can be confined within a certain range after each step of iteration. The formulation of PGD at the $(k+1)$-th step is represented as follows:

$$\boldsymbol{I}_{adv}^{(0)} = \boldsymbol{I}_{ori} + \mathcal{U}(-\varepsilon, \varepsilon), \tag{2.4}$$

$$\boldsymbol{I}_{adv}^{(k+1)} = \mathcal{P}\left\{\boldsymbol{I}_{adv}^{(k)} + \alpha \cdot \text{sign}\left[\nabla_{\boldsymbol{I}_{adv}^{(k)}} \mathcal{L}\left(\boldsymbol{F}\left(\boldsymbol{x}_{adv}^{(k)}\right), \boldsymbol{y}_{ori}\right)\right]\right\}, \tag{2.5}$$

where $\mathcal{U}(\cdot, \cdot)$ represents the uniform distribution, with α signifying the step size. Through the projection function $\mathcal{P}\{\cdot\}$, it is ensured that the produced adversarial example $\boldsymbol{I}_{adv}$ remains within the bounds of an ε-ball centered around $\boldsymbol{I}_{ori}$, where ε denotes the radius.

Reference [83]

Unlike the traditional iteration-based FGSMs [82, 84], Momentum Iterative Fast Gradient Sign Method (MI-FGSM, MIM) incorporates the momentum term into the interactive process of the adversarial perturbation generation, thereby helping to stabilize the direction of updates and produce more effective adversarial examples. The formulae for MI-FGSM at

the $(k+1)$-th step are presented as follows:

$$\begin{aligned}
&\boldsymbol{g}^{(0)} = 0,\ \boldsymbol{I}_{adv}^{(0)} = \boldsymbol{I}_{ori}, \\
&\boldsymbol{g}^{(k+1)} = \mu \cdot \boldsymbol{g}^{(k)} + \frac{\nabla_{\boldsymbol{I}_{adv}^{(k)}} \mathcal{L}\left(\boldsymbol{F}\left(\boldsymbol{I}_{adv}^{(k)}\right), \boldsymbol{y}_{ori}\right)}{\left\|\nabla_{\boldsymbol{I}_{adv}^{(k)}} \mathcal{L}\left(\boldsymbol{F}\left(\boldsymbol{I}_{adv}^{(k)}\right), \boldsymbol{y}_{ori}\right)\right\|_1}, \\
&\boldsymbol{I}_{adv}^{(k+1)} = \mathcal{P}\left\{\boldsymbol{I}_{adv}^{(k)} + \alpha \cdot \operatorname{sign}\left(\boldsymbol{g}^{(k+1)}\right)\right\},
\end{aligned} \tag{2.6}$$

where μ denotes the decay factor about the direction of the gradient, $\boldsymbol{g}$ is the value of momentum, α represents the step size, and $\mathcal{P}\{\cdot\}$ signifies the projection function, which is capable of projecting the produced adversarial example $\boldsymbol{I}_{adv}$ within the confines of the ε-ball.

Reference [85]

Inspired by the data augmentation [3, 86], DI^2-FGSM (Diverse Inputs Iterative Fast Gradient Sign Method) enhances the transferability of adversarial examples by introducing the input diversity. In this method, a transformation function with random, designed for the inputs, is applied throughout the iterations of adversarial example generation. It should be noted that a kind of momentum-based DI^2-FGSM attack is used in the experiments of this thesis, and the expressions of this kind of DI^2-FGSM attack at the $(k+1)$-th step are presented as follows:

$$\begin{aligned}
&\boldsymbol{I}_{adv}^{(0)} = \boldsymbol{I}_{ori} + \mathcal{U}(-\varepsilon, \varepsilon), \\
&\boldsymbol{g}^{(k+1)} = \mu \cdot \boldsymbol{g}^{(k)} + \frac{\nabla_{\boldsymbol{I}_{adv}^{(k)}} \mathcal{L}\left(\boldsymbol{F}\left(\mathcal{T}\left(\boldsymbol{I}_{adv}^{(k)}; P\right)\right), \boldsymbol{y}_{ori}\right)}{\left\|\nabla_{\boldsymbol{I}_{adv}^{(k)}} \mathcal{L}\left(\boldsymbol{F}\left(\mathcal{T}\left(\boldsymbol{I}_{adv}^{(k)}; P\right)\right), \boldsymbol{y}_{ori}\right)\right\|_1}, \\
&\boldsymbol{I}_{adv}^{(k+1)} = \mathcal{P}\left\{\boldsymbol{I}_{adv}^{(k)} + \alpha \cdot \operatorname{sign}\left(\boldsymbol{g}^{(k+1)}\right)\right\}, \\
&\mathcal{T}\left(\boldsymbol{I}_{adv}^{(k)}; P\right) = \begin{cases} \mathcal{T}\left(\boldsymbol{I}_{adv}^{(k)}\right), & \text{with probability } P \\ \boldsymbol{I}_{adv}^{(k)}, & \text{with probability } 1-P \end{cases},
\end{aligned} \tag{2.7}$$

where $\mathcal{U}(\cdot,\cdot)$ represents the uniform distribution, while the parameters μ, $\boldsymbol{g}$, α, and $\mathcal{P}\{\cdot\}$ keep the definitions assigned in Eq. (2.6). In addition, $\mathcal{T}(\cdot;\cdot)$ denotes the transformation

function with random, with P representing the probability of transformation.

Reference [87]

Drawing inspiration from the expectation over transformation (EOT) [88], an enhanced variant of PGD, termed averaged PGD (A-PGD, EOTPGD), was developed to achieve greater stability and effectiveness in adversarial attacks compared to the vanilla PGD. Specifically, this method incorporates expectation into the PGD attack, with the formulations at the $(k+1)$-th step of EOTPGD being delineated as follows:

$$\begin{aligned}
\boldsymbol{I}_{adv}^{(0)} &= \boldsymbol{I}_{ori} + \mathcal{U}(-\varepsilon, \varepsilon), \\
\boldsymbol{I}_{adv}^{(k+1)} &= \mathcal{P}\left\{\boldsymbol{I}_{adv}^{(k)} + \alpha \cdot \operatorname{sign}\left(\mathbb{E}\left[\nabla_{\boldsymbol{I}_{adv}^{(k)}} \mathcal{L}\left(\boldsymbol{F}(\boldsymbol{I}_{adv}^{(k)}), \boldsymbol{y}_{ori}\right)\right]\right)\right\},
\end{aligned} \tag{2.8}$$

where $\mathbb{E}[\cdot]$ represents the operation of expectation, while α signifies the step size.

Reference [89]

Carlini and Wagner (CW), which is designed based on the optimization, is an innovative attack approach. Unlike other adversarial attack methods that simply maximize the misclassification, the CW attack balances the strength of the perturbation and the ability to cause misclassification by optimizing the objective function. To produce adversarial examples with higher deceptiveness, variable $\boldsymbol{w}$ is introduced, which is optimized based on the following formulations:

$$\begin{aligned}
\boldsymbol{w}' &= \min_{\boldsymbol{w}} \left\| \frac{1}{2}(\tanh(\boldsymbol{w}) + 1) - \boldsymbol{I}_{ori} \right\|_2^2 + c \cdot \mathcal{G}\left(\frac{1}{2}(\tanh(\boldsymbol{w}) + 1)\right), \\
\boldsymbol{I}_{adv} &= \frac{1}{2}(\tanh(\boldsymbol{w}') + 1), \\
\mathcal{G}(\cdot) &= \max\left(\boldsymbol{F}(\cdot)_{\boldsymbol{y}_{ori}} - \max_{\boldsymbol{i} \neq \boldsymbol{y}_{ori}} \boldsymbol{F}(\cdot)_{\boldsymbol{i}}, -\kappa\right),
\end{aligned} \tag{2.9}$$

where c is a hyperparameter directly influencing the strength of the adversarial examples generated, while κ acts as a confidence hyperparameter, enhancing the probability of misclassification for the adversarial example $\boldsymbol{I}_{adv}$. The output probability corresponding to the true label $\boldsymbol{y}_{ori}$ is denoted by $\boldsymbol{F}(\cdot)_{\boldsymbol{y}_{ori}}$, and $\boldsymbol{F}(\cdot)_{\boldsymbol{i}}$ signifies the probability associated with the misclassification.

Reference [90]

Due to the drawback that too many pixels in the target image may be modified by the existing adversarial attack methods, potentially leading to perception by the human eyes, an adversarial attack algorithm called One Pixel was developed by the authors, capable of achieving good attack effects by altering a minimal number of pixel points, or even a single pixel point in the image. The optimization problem can be expressed in mathematical expressions as follows:

$$\begin{aligned} \max_{\Delta_{\boldsymbol{I}}} \quad & \boldsymbol{F}\left(\boldsymbol{I}_{\text{ori}} + \Delta_{\boldsymbol{I}}\right) \\ \text{s.t.} \quad & \|\Delta_{\boldsymbol{I}}\|_0 \leq d, \end{aligned} \tag{2.10}$$

where $\Delta_{\boldsymbol{I}}$ is the additional adversarial perturbation about $\boldsymbol{I}_{ori}$, and the value of d is small. If $d = 1$, it is a pure attack on one pixel. To address the optimization problem mentioned above, the authors adopted the differential evolution algorithm, a series of sub-images can be generated by iteratively modifying each pixel, then these sub-images are compared to the original image. To achieve the goal of the attack, the sub-image producing the most effective attack impact is chosen.

Reference [91]

The authors proposed an adversarial attack algorithm, called Universal Adversarial Perturbations (UAP), that generates the "omnipotent" perturbations. These perturbations are independent of the individual images and are generic. Thus, the objective of the UAP attack is to find the universal perturbations φ such that they satisfy the following equation for most images $\boldsymbol{I}_{ori}$ drawn from the distribution μ:

$$\boldsymbol{F}\left(\boldsymbol{I}_{ori} + \varphi\right) \neq \boldsymbol{F}\left(\boldsymbol{I}_{ori}\right), \tag{2.11}$$

where $\boldsymbol{F}$ represents the classifier network. Additionally, to ensure the quality and effectiveness of the generated adversarial examples, the universal adversarial perturbations φ

should also meet the following conditions:

$$\textit{Condition A}: \quad \|\varphi\|_p \leq \varsigma, \tag{2.12}$$

$$\textit{Condition B}: \quad \mathbb{P}_{\boldsymbol{I}_{ori}\sim\varphi}\left[\boldsymbol{F}\left(\boldsymbol{I}_{ori}+\varphi\right) \neq \boldsymbol{F}\left(\boldsymbol{I}_{ori}\right)\right] \geq 1-\delta, \tag{2.13}$$

where ς controls the strength of the perturbations φ, and δ is capable of regulating the extent to which images sampled from the distribution μ are deceived. In order to generate μ, the authors performed the iterative calculations on the sampled images, followed by the summary operation.

Reference [92]

In most of the traditional attack approaches with the setting of black-box, the adversarial examples are typically produced using alternative models before being employed to attack the target models. An effective adversarial attack method, which directly attacks the black-box target model to produce adversarial examples, was proposed. Since the gradients of the target black-box model cannot be obtained directly, this method employs the zeroth order optimization (ZOO) to estimate the gradients. Given input data $\boldsymbol{x}$, the strategy for estimating the gradients in the ZOO attack method is expressed as follows:

$$\boldsymbol{G}_i := \frac{\partial \mathcal{L}\left(\boldsymbol{x}\right)}{\partial \boldsymbol{x}_i} \approx \frac{\mathcal{L}\left(\boldsymbol{x}+h\boldsymbol{e}_i\right)-\mathcal{L}\left(\boldsymbol{x}-h\boldsymbol{e}_i\right)}{2h}, \tag{2.14}$$

$$\boldsymbol{H}_i := \frac{\partial^2 \mathcal{L}\left(\boldsymbol{x}\right)}{\partial^2 \boldsymbol{x}_{ii}^2} \approx \frac{\mathcal{L}\left(\boldsymbol{x}+h\boldsymbol{e}_i\right)-2\mathcal{L}\left(\boldsymbol{x}\right)+\mathcal{L}\left(\boldsymbol{x}-h\boldsymbol{e}_i\right)}{h^2}, \tag{2.15}$$

where $\boldsymbol{G}_i$ and $\boldsymbol{H}_i$ respectively denote the first-order gradient and the second-order gradient. h represents a constant with small value, while $\boldsymbol{e}_i$ is defined as a vector where the i-th element is 1, with all remaining elements set to 0.

2.2.2 Adversarial Defense for Image Classification

2.2.2.1 Adversarial Training

Reference [43]

In traditional adversarial training approaches, enhancing the model's adversarial robustness often comes at the expense of sacrificing its inherent accuracy. The authors identified

the non-smooth nature of the prevalent ReLU activation function as a significant contributor to the suboptimal results of adversarial training. Consequently, they introduced a methodology called Smooth Adversarial Training (SAT), which primarily improves the conventional ReLU activation function by smoothing it. The improved activation function, termed SmoothReLU, can be expressed in the following equation:

$$\text{SmoothReLU}(x, \theta) = \begin{cases} x - \frac{1}{\theta}\log(\theta x + 1), & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}, \tag{2.16}$$

where θ is a learnable positive parameter. Moreover, SmoothReLU can always maintain the good property of being differentiable:

$$\frac{\mathrm{d}}{\mathrm{d}x}\text{SmoothReLU}(x, \theta) = \begin{cases} \frac{\theta x}{1 + \theta x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}. \tag{2.17}$$

Reference [46]

In this paper, the authors mainly focused on the improvement of the single-step adversarial training approach. While the single-step training method has a shorter training time and lower computational complexity compared to the multi-step adversarial training approaches, the performance of the trained models is suboptimal. The authors argued that overfitting is likely to occur during the initial phase of single-step adversarial training. To address this issue, they proposed a novel strategy called Single-step Adversarial Training with Dropout Scheduling (SADS). Specifically, SADS mainly adds the dropout layer to each non-linear layer in the model requiring training. Experimental results demonstrate SADS can enhance the robustness of the single-step method of adversarial training without high computational costs.

Reference [44]

In this paper, the authors argued that the model capacity of the existing CNNs often falls short in adversarial training. Therefore, given the limited model capacity, the authors believed that not all samples should be treated equally. Consequently, they improved the traditional adversarial training strategy and proposed an approach called Geometry-Aware Instance-Reweighted Adversarial Training (GAIRAT). From the geometric perspective,

the adversarial robustness of samples is inversely proportional to their distance from the classification boundary, meaning samples closer to the classification boundary are less robust, while those farther away are more robust. Based on this fact, GAIRAT can assign larger weights to samples located near the classification boundary while allocating smaller weights to those situated further from the classification boundary.

Reference [49]

In this paper, enhancing the efficiency of adversarial training while ensuring performance was concentrated. In conventional adversarial training approaches, generating ideal adversarial examples typically requires extensive training time. To address this issue, the authors introduced a strategy known as Adversarial Training with Transferable Adversarial examples (ATTA). This method is proposed based on the observation that the adversarial examples generated in previous epochs retain their adversarial characteristics in subsequent epochs within the same training process. Specifically, during the training phase, ATTA continuously utilizes adversarial perturbations from the previous epochs and accumulates them, significantly reducing the number of iterations while maintaining the performance of adversarial examples.

Reference [47]

In this paper, enhancing the attack strategy during the adversarial training process was concentrated. The authors argued that traditional adversarial training methods rely on fixed or manually designed attack strategies to produce adversarial examples, thereby limiting the adversarial robustness in the face of adversarial attacks. Consequently, the authors introduced a framework named LAS-AT, which is innovative and more flexible. This framework can bolster the robust effectiveness and generalization capabilities of deep learning models by employing learnable, dynamic attack strategies instead of fixed, human-designed ones.

2.2.2.2 Adversarial Example Detection

Reference [52]

In this paper, to ascertain if the input data falls under the category of adversarial exam-

ple, the authors ingeniously constructed an additional detection module for the primary architecture used in image classification. The inputs to this detection network derive from the classification network, and this detection network can characterize the identification of adversarial examples as a binary classification task, outputting the probability value of the input data being adversarial. Experimental results indicate that under specific adversarial attack conditions, this method can detect the vast majority of adversarial examples.

Reference [54]

In this paper, the authors analyzed the adversarial examples from the perspective of manifolds in two scenarios: one in which the adversarial examples lie at a substantial distance from the decision boundary, and another where they are close to the boundary. Consequently, the authors introduced a novel approach called MagNet, designed to address both types of adversarial examples accordingly. This approach fundamentally comprises a "detector" and a "reformer". The former is employed to filter out adversarial examples that significantly deviate from the decision boundary. For those near the boundary, MagNet utilizes the "reformer" to find the approximate representation of the adversarial examples, which can be then used in the subsequent classification task.

Reference [55]

In this paper, in order to enhance the detection of adversarial examples, the authors made improvements in both the training and inference phases of the model. Specifically, during the training phase of the model, they introduced a novel loss function called the reverse cross-entropy (RCE) loss, which encourages the mapping of clean data into the low-dimensional space and facilitates the clustering of clean data of the same class. This approach simplifies the differentiation of adversarial examples. Additionally, in the inference stage of the model, the authors also implemented a thresholding strategy centered around the kernel density theory.

Reference [53]

In this paper, the authors delved into the adversarial examples from a statistical perspective. Specifically, they argued that the adversarial samples and clean data do not reside within the same data manifold. Consequently, they introduced methods of kernel density

estimates and Bayesian uncertainty estimates. Specifically, the kernel density estimates method is employed to identify data points that stray from the data manifold, while the Bayesian uncertainty estimates approach is utilized to detect low-confidence data that the kernel density estimates method fails to recognize.

Reference [59]

In this paper, the authors proposed an attack-agnostic strategy for detecting the adversarial examples, where they characterized the adversarial perturbations in images as the form of noises and combined this with classical image processing technologies for detecting them. Specifically, they introduced techniques of scalar quantization and smoothing spatial filter to better perceive the presence of adversarial perturbations. Additionally, the authors also introduced the concept of image entropy to adaptively adjust the parameters.

2.2.2.3 Processing of Input Data

Reference [61]

In this paper, the authors discovered that compressing the input adversarial images can remove the unnecessary information, thereby achieving the purpose of eliminating the adversarial noise. Therefore, they proposed a simple but effective image compression framework named ComDefend, which mainly consists of two parts: a compression module for removing adversarial noise and a reconstruction module for restoring clean images. Moreover, the overall architecture of ComDefend bears a resemblance to the framework of the autoencoder.

Reference [63]

In this paper, improving the method for removing the adversarial perturbations was mainly concentrated. It was found that in traditional methods of adversarial perturbation removal, the remaining adversarial noise often leads to misclassification by the classifier. To address this issue, an effective adversarial defense approach named High-level representation Guided Denoiser (HGD) was proposed. The essence of this method is the introduction of a new loss function that minimizes the distance between the output features of the clean image and the restored image, rather than directly computing the pixel distance between

the clean and restored images.

Reference [64]

In this paper, the authors constructed a kind of low-dimensional image space that resembles the space of clean images. This kind of low-dimensional space is designed based on the theory of convolutional sparse coding. By mapping both clean and adversarial images to this low-dimensional space, adversarial perturbations can be effectively removed, thereby significantly enhancing the adversarial robustness of the model.

Reference [69]

In this paper, the authors introduced the matrix estimation theory to the adversarial defense task for the first time. Specifically, they began by randomly deleting pixels from the image with a certain probability and then employed the matrix estimation for reconstruction. The authors believed that their proposed method can disrupt the structure of adversarial noise while preserving the global structure of the original image. Furthermore, they pointed out that as a preprocessing approach, this strategy can be easily integrated into other architectures and enhance the generalization of the model.

2.2.2.4 Regularization

Reference [71]

In this paper, the authors observed a decrease in both the curvature of the decision boundary and the loss surface as the adversarial robustness of the model improved. Consequently, they introduced a novel curvature regularization approach. Specifically, they devised a novel regularization term that effectively constrains the curvature of the decision boundary and promotes the local linearity, aiming to enhance the adversarial robustness of the model. And this regularization term can make the gradients around the data points the same.

Reference [74]

In this paper, the author believed that the local flatness of the loss surface is highly correlated with the adversarial robustness of the model. The authors defined the regularization term of local flatness as the maximum norm of the gradient in the vicinity of clean images.

Specifically, during the training process, the authors combined the local flatness regularization term with the classical loss function and minimized them.

Reference [76]

In this paper, the authors introduced the theory of information geometry into the adversarial defense task and extended the classical classification cross-entropy. Specifically, they added a regularization term to the classification cross-entropy, which is crafted utilizing the principles of the Fisher-Rao distance theory, and it ensures that the predictions of clean images and adversarial examples are close to each other.

2.3 Adversarial Attack for Image Captioning

Reference [93]

In this paper, in order to investigate whether the visual and language models represented by the image captioning system really mine the connotative relationship between the text and image modalities, the authors expanded the existing MS-COCO dataset [94] by adding well-designed erroneous image captions, thus creating a dataset with the name of FOIL-COCO. Based on this dataset, the authors verified the performance of the image captioning system on three tasks: caption classification, abnormal word detection, and abnormal word correction. They concluded that the image captioning system fails to adequately mine the relationships between the texts and images. This study is a pioneering work on the robustness of the image captioning system and provides a strong foundation for subsequent research on adversarial attack methods for the image captioning system.

Reference [41]

In this paper, a novel method, which was designed for generating adversarial examples for the image captioning system, was proposed for the first time. Specifically, the authors proposed an approach named Show-and-Fool, which is an optimization-based adversarial example generation method and is specifically implemented through the carefully designed powerful loss functions. This approach effectively misguides the image captioning system into generating captions or keywords chosen at random. Furthermore, the authors have

also uncovered that the adversarial examples crafted using the Show-and-Fool method exhibit a certain degree of transferability. Specifically, the authors found that the adversarial examples generated for the generation-based image captioning system Show-and-Tell can also deceive other image captioning systems.

Reference [42]

In this paper, a kind of precise adversarial attack strategy, which was designed for the partial image captions, was proposed for the first time. In the generation-based image captioning system, the outputs are often structured sequences. Therefore, the authors believed that it is necessary to design the adversarial attack approach specifically tailored for the generation-based image captioning system. Specifically, they described the precision adversarial attack as the structured output learning model that incorporates latent variables. Furthermore, two structured optimization methods, namely the generalized expectation maximization and the structural support vector machines with latent variables, were proposed.

Reference [95]

In this paper, a novel adversarial example generation strategy called *Show, Mask and Tell* for the image captioning system was proposed. This method primarily involves adding meticulously crafted adversarial noises to the input image, leading to the removal of target words from image captions. Remarkably, the accuracy of the remaining captions is well preserved. This method is mainly implemented by optimizing a loss function, which contains three regularization terms, namely word removal regularization term used to limit the frequency of target words, captioning accuracy regularization term used to ensure the accuracy of captions, and noise filtering regularization term used to remove noises generated by adversarial perturbations. In addition, the authors also proposed a metric called "Attack Quality" to comprehensively evaluate the success rate of the attack and the accuracy of the captions, thereby more comprehensively evaluating the performance of the attack.

Reference [96]

In this paper, the construction of adversarial examples for the image captioning system in the complex domain was pioneered. Then an adversarial attack method called Fooled-by-

Imagination was proposed. Specifically, they designed a loss function mainly based on the complex domain, in which the semantic features corresponding to the adversarial examples and the semantic features corresponding to the original images are used as the imaginary part and the real part of the loss function respectively, and this loss function is the modulus of the complex formed by the semantic features of the adversarial examples and the semantic features of the original images. In addition, the author also designed an ℓ_2-norm loss function to maintain the similarity between the generated adversarial examples and the original images.

Reference [97]

In this paper, improving the generation speed and attack capabilities of adversarial examples for the image captioning system was mainly concentrated. Specifically, the authors designed a generative model to generate adversarial examples. At the same time, the authors also specially designed three loss functions for this generation model to ensure the quality and transfer ability of the generated adversarial examples, namely the image feature distortion loss, the local-global mismatching loss, and the language diversity loss. Specifically, the purpose of the image feature distortion loss is to make the distance between the features of original images and the features of the corresponding adversarial examples in the image domain as large as possible. The local-global mismatching loss is utilized to decrease the matching degree between the adversarial images and the corresponding ground truth captions. Furthermore, the language diversity loss can ensure that the adversarial captions generated from adversarial images have significant differences from the ground truth captions.

Reference [98]

In this paper, the caption output efficiency of the Neural Image Caption Generation (NICG) model was utilized as the target of the adversarial attack by the authors, who adopted a different approach. Specifically, the authors designed a novel strategy for the image captioning system called NICGSlowDown. This method can generate adversarial perturbations for the original input images that are imperceptible to the human eyes, causing the image captioning system to consume significantly more computational resources, thereby greatly reducing the operational efficiency of the NICG model. Additionally, the authors also

constructed two evaluation metrics named I-Loop and I-Latency to measure the efficiency robustness of the NICG model. The larger the values of I-Loop and I-Latency, the lower the model's efficiency, which indicates a more successful attack.

Reference [99]

In this paper, the security of the image captioning system in the physical world was concentrated for the first time. Specifically, the authors proposed a practical adversarial attack algorithm named CAPatch, which can generate adversarial patches for the input images of the image captioning system, leading the system to generate captions that either lack key words or are entirely incorrect. CAPatch generates specific adversarial patches through four main steps: detection assurance, attention enhancement, caption alteration, and robustness improvement.

2.4 Conclusions

In this chapter, the research progresses on the adversarial attack and defense for the image classification system and the related works about the adversarial attack on the image captioning system have been introduced. These studies are closely related to the proposition of this thesis. Fig. 2.2 shows the research map of the related works presented in this chapter.

Adversarial Attack and Defense Methods			
Image Classification System		**Image Captioning System**	
Attack	**Defense**	**Attack**	**Defense**
[81] I. J. Goodfellow *et al.* (2015) [82] A. Madry *et al.* (2018) [83] Y. Dong *et al.* (2018) [85] C. Xie *et al.* (2019) [87] R. S. Zimmermann (2019) [89] N. Carlini *et al.* (2017) [90] J. Su *et al.* (2019) [91] S.-M. Moosavi-Dezfooli *et al.* (2017) [92] P.-Y. Chen *et al.* (2017)	[43] C. Xie *et al.* (2020) [46] B. Vivek *et al.* (2020) [44] J. Zhang *et al.* (2021) [49] H. Zheng *et al.* (2020) [47] X. Jia *et al.* (2022) [52] J. H. Metzen *et al.* (2017) [54] D. Meng *et al.* (2017) [55] T. Pang *et al.* (2018) [53] R. Feinman *et al.* (2017) [59] B. Liang *et al.* (2021) [61] X. Jia *et al.* (2019) [63] F. Liao *et al.* (2018) [64] B. Sun *et al.* (2019) [69] Y. Yang *et al.* (2019) [71] S.-M. Moosavi-Dezfooli *et al.* (2019) [74] J. Xu *et al.* (2020) [76] M. Picot *et al.* (2022)	[93] R. Shekhar *et al.* (2017) [41] H. Chen *et al.* (2018) [42] Y. Xu *et al.* (2019) [95] J. Ji *et al.* (2020) [96] S. Zhang *et al.* (2020) [97] H. Wu *et al.* (2022) [98] S. Chen *et al.* (2022) [99] S. Zhang *et al.* (2023)	

Proposed Methods

Chapter 3
Integration of robust regularization-based enhanced multi-stage fusion features into image classification system.

Chapter 5
Integration of robust caption features from pre-trained models into image classification system.

Chapter 4
Integration of robust caption features from stable RNN into image captioning system.

Figure 2.2: A map of related works shown in Chapter 2.

Chapter 3

Integration of Robust Regularization-based Enhanced Multi-stage Fusion Features into Image Classification System

3.1 Introduction

This chapter shows the Enhanced Multi-Stage Feature Fusion Network (EMSF^{2}Net), a novel CNN-based framework. This framework is built upon three fundamental components: MSFE (multi-stage feature enhancement), MSF^2 (multi-stage feature fusion), and regularization. Regarding the MSFE component, drawing inspiration from SENet [100], the global average pooling (GAP) operation is initially applied to the features from each stage, thereby acquiring the global features with channel level. Next, the global features with channel level are multiplied with the initial features to yield the enhanced features. Regarding the MSF^2 component, the multi-stage features after enhancement are initially flattened into the features with one dimension. Subsequently, these features are directly subjected to the operation of concatenation. Despite its simplicity, this operation shows high efficiency, as MSF^2 is capable of maintaining the channel-wise global information learned from MSFE. Lastly, in the training phase, the operation of regularization is applied to both the fused features and the initial multi-stage features. In particular, a loss function with regularization is employed as the means of regularization for the proposed framework. Introducing the term of regularization into the enhanced multi-stage fusion features,

as demonstrated by the proposed framework, can substantially bolster the capability of CNN to defend against the adversarial attacks. The findings further indicate an increased suitability of the enhanced multi-stage fusion features for regularization. Moreover, in contrast to current adversarial defense methods that rely on global information, the technique of regularization is incorporated into the fused global features. This integration operation shows that applying regularization to these fused global features enhances the capability of CNN to defend against the adversarial attacks.

The key contributions in this chapter are outlined below,

- We propose EMSF^{2}Net, a novel framework wherein the global information across all channels is effectively represented and kept through the enhanced multi-stage fusion features, for the image classification system.
- This chapter demonstrates that the application of regularization to the enhanced multi-stage fusion features notably boosts the adversarial robustness. In other words, the robust regularization-based enhanced multi-stage fusion features are introduced into the image classification system.
- The robustness and effectiveness of the proposed approach are demonstrated by the comprehensive experimental results conducted across the attacks of white-box scenarios.

3.2 Methodology

The framework of the EMSF^{2}Net is shown in Fig. 3.1, while Table 3.1 depicts the structure of the baseline. As illustrated in Fig 3.1, the output features of STAGEs 2 to 4 of the standard ResNet50 [1], detailed in Table 3.1, are regarded as the multi-stage features. Three pivotal components primarily compose EMSF^{2}Net: MSFE, MSF^2, and the operation of regularization, each of which is introduced in the subsequent subsections.

3.2.1 Multi-Stage Features Enhancement (MSFE)

The MSFE component, with its specifics illustrated in the FE Block of Fig. 3.1, is explained in this subsection. In STAGE m (m = 2, 3, 4), the output features of the Conv Block, denoted

as $\boldsymbol{U}_m = \left[\boldsymbol{u}_m^1, \boldsymbol{u}_m^2, \cdots, \boldsymbol{u}_m^{C_m}\right] \in \mathbb{R}^{H_m \times W_m \times C_m}$, corresponds to Feature m (m = 2, 3, 4) shown in Fig. 3.1. The channel number, height, and width of Feature m are denoted as C_m, H_m, and W_m, respectively. Additionally, for each channel l (where $l = 1, \cdots, C_m$), $\boldsymbol{u}_m^l$ is indicative of the corresponding sub-feature within $\boldsymbol{U}_m$.

We initially apply the operation of GAP to the input $\boldsymbol{U}_m$, resulting in the acquisition of the global feature with channel level $\boldsymbol{Z}_m = [z_m^1, z_m^2, \cdots, z_m^{C_m}] \in \mathbb{R}^{1 \times 1 \times C_m}$. The specific operation for each channel l (where $l = 1, \cdots, C_m$) is described below:

$$z_m^l = GAP\left(\boldsymbol{u}_m^l\right) = \frac{1}{H_m \times W_m} \sum_{i=1}^{H_m} \sum_{j=1}^{W_m} \boldsymbol{u}_m^l(i, j). \tag{3.1}$$

Subsequently, the global feature $\boldsymbol{Z}_m$ with channel level undergoes a multiplication process with the initial input feature $\boldsymbol{U}_m$, serving as the operation of enhancement for the feature. And $\boldsymbol{U}'_m = \left[\boldsymbol{u}'^1_m, \boldsymbol{u}'^2_m, \cdots, \boldsymbol{u}'^{C_m}_m\right] \in \mathbb{R}^{H_m \times W_m \times C_m}$ denotes the feature after enhancement. The following equation shows the formulation for the operation on channel l (where $l = 1, \cdots, C_m$):

$$\boldsymbol{u}'^l_m = z_m^l \cdot \boldsymbol{u}_m^l. \tag{3.2}$$

After the GAP operation, the initial feature generates weights, which contain the receptive field with globality. By channel-wise fusion of feature weights with the original feature, each channel in the original feature acquires global information, thereby enhancing its richness and augmenting its expressiveness for realization. Ultimately, the Conv Block processes $\boldsymbol{U}'_m$, yielding $\tilde{\boldsymbol{U}}_m \in \mathbb{R}^{H_m \times W_m \times C_m}$, which is the final enhanced feature, as depicted in Fig. 3.1.

3.2.2 Multi-Stage Features Fusion (MSF^2)

After the feature after enhancing $\tilde{\boldsymbol{U}}_m \in \mathbb{R}^{H_m \times W_m \times C_m}$ ($m = 2, 3, 4$) for each stage is acquired, these features are performed by the operation of fusion. Initially, through the following operations, the features $\tilde{\boldsymbol{U}}_2$, $\tilde{\boldsymbol{U}}_3$, and $\tilde{\boldsymbol{U}}_4$ are flattened into $\boldsymbol{v}_2$, $\boldsymbol{v}_3$, and $\boldsymbol{v}_4$, respectively:

$$\boldsymbol{v}_m = F\left(AvgP\left(\tilde{\boldsymbol{U}}_m\right)\right) \in \mathbb{R}^{C_m}. \tag{3.3}$$

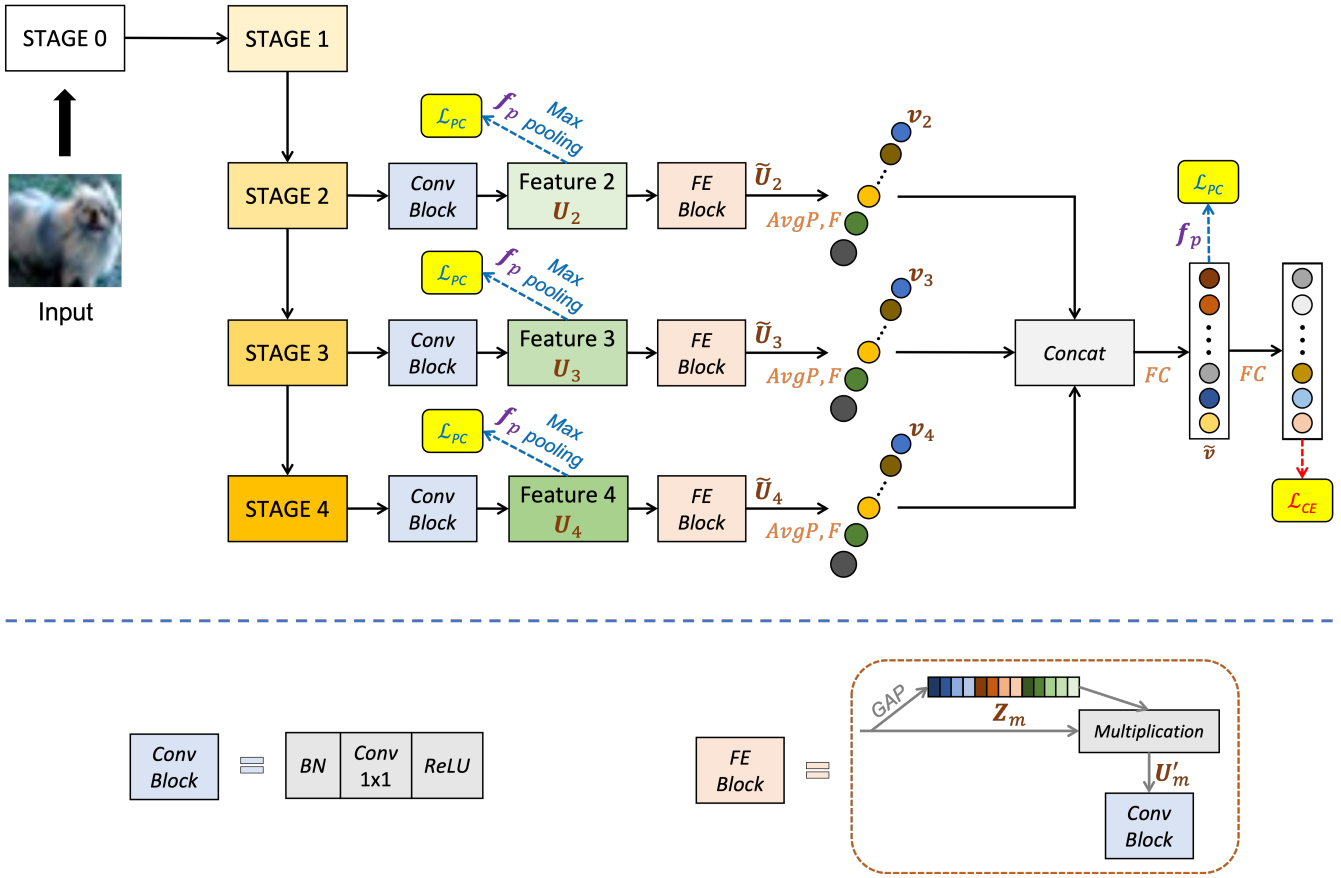


Figure 3.1: The framework of EMSF^{2}Net. Serving as its backbone component, ResNet50 [1] is adopted. STAGEs 0 to 4 in this figure and those in the baseline share the same structures, and the structural specifics of STAGEs 0 to 4 are detailed in Table 3.1. Furthermore, the block of feature enhancement is denoted by FE Block, the operation of concatenation is signified by Concat, and within the FE Block, the global average pooling is represented by GAP.

Table 3.1: Details of the baseline architecture. Specifically, $\mathcal{L}_{PC}$ denotes the regularization loss function, whereas the classic cross-entropy loss is represented by $\mathcal{L}_{CE}$.

Layer	**ResNet-50**
STAGE 0	Conv(64, 3×3), BN, ReLU
STAGE 1	$\left\{\begin{array}{c}\text{Conv}(64, 1 \times 1), \text{BN}, \text{ReLU} \\ \text{Conv}(64, 3 \times 3), \text{BN}, \text{ReLU} \\ \text{Conv}(256, 1 \times 1), \text{BN} \\ \text{shortcut}, \text{ReLU}\end{array}\right\} \times 3$
STAGE 2	$\left\{\begin{array}{c}\text{Conv}(128, 1 \times 1), \text{BN}, \text{ReLU} \\ \text{Conv}(128, 3 \times 3), \text{BN}, \text{ReLU} \\ \text{Conv}(512, 1 \times 1), \text{BN} \\ \text{shortcut}, \text{ReLU}\end{array}\right\} \times 4$ Max pooling $\rightarrow \mathcal{L}_{PC}$
STAGE 3	$\left\{\begin{array}{c}\text{Conv}(256, 1 \times 1), \text{BN}, \text{ReLU} \\ \text{Conv}(256, 3 \times 3), \text{BN}, \text{ReLU} \\ \text{Conv}(1024, 1 \times 1), \text{BN} \\ \text{shortcut}, \text{ReLU}\end{array}\right\} \times 6$ Max pooling $\rightarrow \mathcal{L}_{PC}$
STAGE 4	$\left\{\begin{array}{c}\text{Conv}(512, 1 \times 1), \text{BN}, \text{ReLU} \\ \text{Conv}(512, 3 \times 3), \text{BN}, \text{ReLU} \\ \text{Conv}(2048, 1 \times 1), \text{BN} \\ \text{shortcut}, \text{ReLU}\end{array}\right\} \times 3$
5	Average pooling $\rightarrow \mathcal{L}_{PC}$
6	FC(4096) $\rightarrow \mathcal{L}_{PC}$
7	FC(10) $\rightarrow \mathcal{L}_{CE}$

As illustrated in Eq. 3.3, the process begins with the application of average pooling ($AvgP$) to transform $\tilde{\boldsymbol{U}}_m$ into the dimension of $1 \times 1 \times C_m$. This is followed by the operation of flattening (F) which reshapes it into the dimension of C_m. Subsequently, the fusion operation is performed on the vectors after flattening from each stage, with the corresponding operation detailed as follows:

$$\tilde{\boldsymbol{v}} = FC\left([\boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4]\right) \in \mathbb{R}^{C'}. \tag{3.4}$$

The aforementioned equation illustrates that initially, a new vector is generated from the concatenation of $\boldsymbol{v}_2$, $\boldsymbol{v}_3$, and $\boldsymbol{v}_4$, followed by a fully connected layer (FC) for its mapping onto the dimension of C'.

Despite the apparent simplicity of this fusion operation, the channel-wise information with globality learned through the FE Block of each stage is effectively kept. Conversely, utilizing alternative fusion approaches could potentially compromise the information of the global features that have been learned.

3.2.3 Regularization

The prototype conformity loss $\mathcal{L}_{PC}$, proposed by Mustafa et al. [72], is chosen as the method of regularization in this chapter. Considering a classification scenario with k distinct classes and a set of images for training, we denote $\boldsymbol{f}_p$ as the output feature for the input $\boldsymbol{x}_p$ belonging to the class of $\boldsymbol{y}_p$. The formulation of $\mathcal{L}_{PC}$ is presented below:

$$\mathcal{L}_{PC} = \sum_{p} \left\{ \left\| \boldsymbol{f}_p - \boldsymbol{w}^c_{y_p} \right\|_2 - \frac{1}{k-1} \sum_{q \neq y_p} \left(\left\| \boldsymbol{f}_p - \boldsymbol{w}^c_q \right\|_2 + \left\| \boldsymbol{w}^c_{y_p} - \boldsymbol{w}^c_q \right\|_2 \right) \right\}, \tag{3.5}$$

in the above equation, the class centroid of the true class $\boldsymbol{y}_p$ is denoted as $\boldsymbol{w}^c_{y_p}$, while the centroids of all distinct classes excluding $\boldsymbol{y}_p$ are represented as $\boldsymbol{w}^c_q$. The aforementioned equation demonstrates that $\mathcal{L}_{PC}$ serves to augment the inter-class distance, while the distance of $\boldsymbol{f}_p$ and $\boldsymbol{w}^c_{y_p}$ is concurrently diminished, consequently leading to more pronounced boundaries between distinct classes.

When the representational information of the features from each class is abundant, $\mathcal{L}_{PC}$ more effectively perceives the variations across the features from diverse classes. Furthermore, the feature outputs from the proposed framework are characterized by global channel features rich

in information. Consequently, EMSF^{2}Net incorporates $\mathcal{L}_{PC}$ as the approach of regularization. Hence, the overall loss, $\mathcal{L}_{all}$, employed to train the proposed framework, is formulated below:

$$\mathcal{L}_{all} = \mathcal{L}_{CE} + \sum_{k=1}^{4} \mathcal{L}_{PC}^{k}, \tag{3.6}$$

in the given formula, the cross-entropy loss, denoted as $\mathcal{L}_{CE}$, serves to limit the outputs of classification of the proposed framework, while $\mathcal{L}_{PC}$ is introduced to regularize features across multiple stages, and features with enhanced multi-stage fusion. The sum of all items of $\mathcal{L}_{PC}$ in the proposed framework is denoted as $\sum_{k=1}^{4} \mathcal{L}_{PC}^{k}$. $\mathcal{L}_{all}$ exhibits its strength in augmenting the inter-class distances while concurrently reducing intra-class distances within the space of output.

3.3 Dataset & Attack Methods

The widely used dataset and the approaches of adversarial attack employed in this chapter for evaluating the robustness of the proposed framework are presented in this section.

3.3.1 Dataset: CIFAR-10

In this chapter, the CIFAR-10 dataset [79], one of the popular datasets in assessing adversarial defense strategies, served as the dataset for comparing the proposed framework with other methods with state-of-the-art approaches and conducting ablation analysis. The CIFAR-10 dataset is composed of 60,000 images, each measuring the pixels 32 by 32, split into a training set of 50,000 images and a test set comprising 10,000 images. CIFAR-10 comprises distinct classes with the number ten, namely “airplane”, “automobile”, “bird”, “cat”, “deer”, “dog”, “frog”, “horse”, “ship”, and “truck”. Fig. 3.2 shows the representative instances for each class within this dataset.

3.3.2 Adversarial Attacks

In this paper, seven popular adversarial attack methods were employed to evaluate the adversarial robustness and effectiveness of the proposed approach. These adversarial attack approaches can be classified into two distinct categories: those based on the ℓ_∞-norm and the ℓ_2-norm. The ℓ_∞-norm attack methods include FGSM [81], PGD-ℓ_∞ [82], MI-FGSM [83], DI^2-FGSM [85],

and EOTPGD [87]. It should be emphasized that the DI^2-FGSM attack utilized in this chapter is a kind of momentum-based DI^2-FGSM attack, and the formulations of this DI^2-FGSM attack utilized in this chapter have been presented in Eq. 2.7. Additionally, the ℓ_2-norm attack approaches used in this chapter comprise PGD-ℓ_2 [82] and CW [89]. It should be clarified that in the PGD-ℓ_2 attack used in this chapter, no slight perturbations are added to the initial value. This means that in this context, the initial value $\boldsymbol{I}_{adv}^{(0)}$ in Eq. 2.4 of PGD-ℓ_2 attack becomes $\boldsymbol{I}_{ori}$ rather than $\boldsymbol{I}_{ori} + \mathcal{U}(-\varepsilon, \varepsilon)$. Additionally, in the experiments, the toolbox employed for performing adversarial attacks is Torchattacks [101].

3.4 Comparison Experiments

3.4.1 Comparison Methods

Three state-of-the-art defense approaches were chosen for comparison to comprehensively validate the adversarial robustness and performance of the proposed framework.

MART [102]

This approach introduces a new loss function designed for adversarial defense, focusing more effectively on the data that is incorrectly classified to enhance the adversarial robustness of the model.

RobNet [103]

The authors of RobNet concentrated on the architecture of the network, employing the neural architecture search (NAS) technique for the adversarial defense task, enabling the discovery and design of the structures of the robust networks.

BPFC [104]

To mimic the process of human visual perception, a kind of novel regularizer, which ensures the representation with consistency across the features learned from images with various quantization, was introduced, thereby substantially enhancing the model robustness against adversarial attacks.

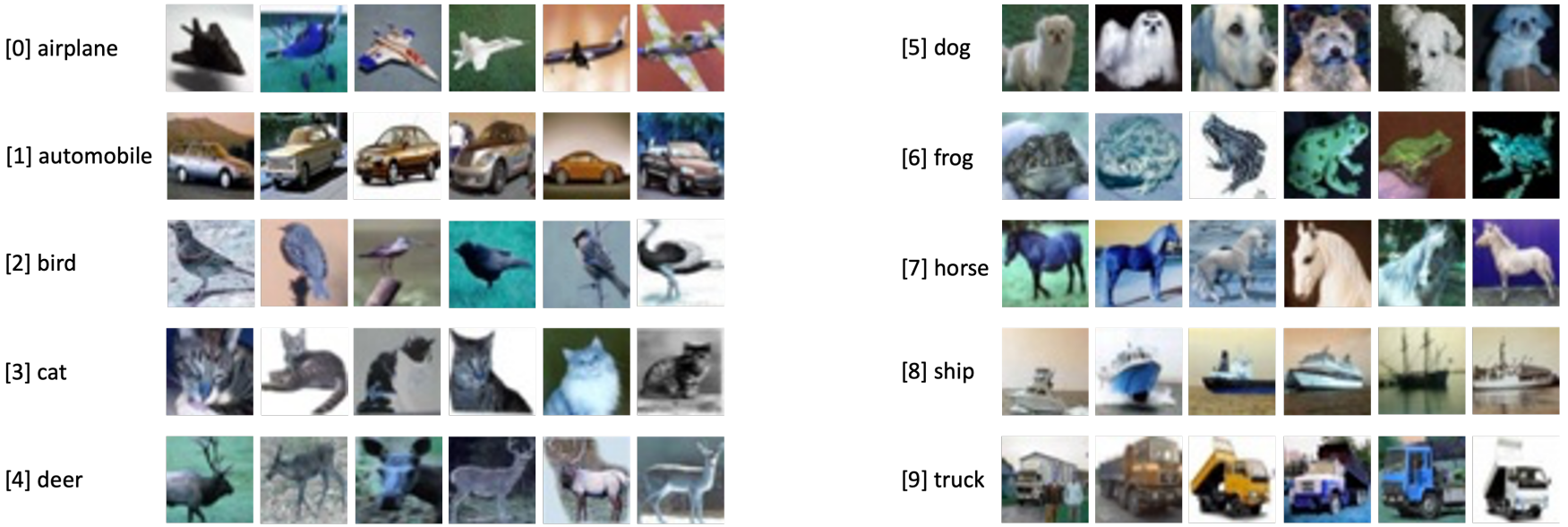


Figure 3.2: Some visual examples of each class in the CIFAR-10 dataset.

3.4.2 Experimental Results under ℓ_∞-norm Attacks

In this subsection, our focus is to present the results about the robust accuracy of EMSF^{2}Net compared to the three comparison methods when facing the ℓ_∞-norm adversarial attacks, based on the CIFAR-10 dataset. A variety of attack methodologies were employed to demonstrate the effectiveness of our proposed approach, including FGSM, PGD, MI-FGSM, DI^2-FGSM, and EOTPGD, each with varying levels of attack strengths. We configured these ℓ_∞-norm attacks in a white-box setting. The attack strength parameters, denoted as ε, were selected as the values of 2/255, 4/255, 8/255, and 16/255.

Initially, the clean accuracy results and the robust accuracy results under FGSM with single step on CIFAR-10 are presented, with the accuracies detailed in Table 3.2. Table 3.2 illustrates that, when subjected to the FGSM attack, the proposed EMSF^{2}Net demonstrates superior performance over the comparison methods while maintaining a high level of accuracy in classifying clean images.

Subsequently, the robust accuracy results under ℓ_∞-norm attacks based on less complex iterations (iteration number = 10) are presented. For ease of differentiation, the attacks were denoted as PGD^{10}-ℓ_∞, MI-FGSM^{10}, DI^2-FGSM^{10}, and EOTPGD^{10}, each being with the iteration value 10. $\varepsilon/10$ was selected as the step size, with ε representing the strength of attack. In the case of MI-FGSM^{10} and DI^2-FGSM^{10}, 0.5 was adopted as the value of momentum factor parameter. In addition, for EOTPGD^{10}, we configured the mean gradient estimation parameter to a value of 5. The accuracies are detailed in Table 3.3. As demonstrated in the table, the results indicate that, even under the challenging conditions of the adversarial attacks based on iteration, our proposed EMSF^{2}Net continues to exhibit significant superiority over the comparison methods. Moreover, as the attack strength denoted by ε progressively escalates, a widening disparity becomes evident between EMSF^{2}Net and comparison methodologies. This observation underscores the effectiveness and adversarial robustness of the approach we proposed.

Lastly, the performance about the adversarial robustness among EMSF^{2}Net and the comparison approaches under ℓ_∞-norm attacks based on more complex iterations (iteration number = 20) is shown. The attacks were denoted as PGD^{20}-ℓ_∞, MI-FGSM^{20}, DI^2-FGSM^{20}, and EOTPGD^{20}, each being with the iteration value 20. Apart from the number of iterations, the remaining parameters in the more complex attacks remain consistent with those in the less complex ones.

Table 3.4 displays the results regarding robust accuracy. The analysis of this table leads to the inference that, in the case of more complex adversarial attacks, there is a clear inverse relationship between ε and the results of the comparison methods. Conversely, EMSF^{2}Net continues to exhibit a high level of accuracy of classification.

3.4.3 Experimental Results under ℓ_2-norm Attacks

In Subsection 3.4.2, the performance of our proposed EMSF^{2}Net, alongside three advanced comparison methods, was evaluated under ℓ_∞-norm attacks employing white-box setting. The classification accuracies have demonstrated that EMSF^{2}Net possesses considerable robustness to the adversarial attacks with ℓ_∞-norm. In Subsection 3.4.3, a different yet prevalent form of adversarial attacks, the attacks based on ℓ_2-norm, are employed to more thoroughly and expansively evaluate the effectiveness of the proposed EMSF^{2}Net. We implemented PGD-ℓ_2 attacks with various iteration values, alongside CW attacks. Here, PGD^{10}-ℓ_2, PGD^{20}-ℓ_2, and PGD^{40}-ℓ_2 correspond to the PGD-ℓ_2 attacks based on 10, 20, and 40 iterations, respectively. Table 3.5 displays the accuracies of the comparison methods and proposed EMSF^{2}Net in terms of adversarial robustness, when defending against attacks with ℓ_2-norm varying in attack strengths denoted by ε or differing iteration numbers. In the implementation of PGD-ℓ_2 attacks, $\varepsilon/10$ was configured as the number of step size. Concurrently, for the CW attacks, 1.0 and 0 were respectively assigned as the values of box-constraint parameter c and confidence κ. The white-box setting was applied to these adversarial attacks with ℓ_2-norm.

Table 3.5 demonstrates that, in contrast to the comparison approaches, the proposed framework consistently achieves superior accuracy across various ℓ_2-norm attacks, varying in both attack strengths and iteration numbers. Specifically, under PGD-ℓ_2 attacks, the accuracy of comparison methods experiences a sharp decline, falling below 1.0% in certain instances, as the value of ε increases. Conversely, the proposed framework demonstrates the capability to sustain a considerably strong robustness against adversarial attacks. Furthermore, under the universally acknowledged difficult CW attack, the EMSF^{2}Net is also capable of sustaining the competitive performance.

Table 3.2: Accuracy results in clean and FGSM attack scenarios on the CIFAR-10 dataset, with the rightmost column displaying accuracies from the proposed method. In addition, the top-performing results are highlighted in bold.

Attack Method	Parameter (ε)	MART	RobNet	BPFC	EMSF^{2}Net (Ours)
No Attack	-	83.6%	82.7%	82.4%	**92.7%**
FGSM	2/255	78.4%	76.9%	73.7%	**83.3%**
	4/255	72.9%	70.6%	64.6%	**81.1%**
	8/255	61.6%	58.4%	50.1%	**73.3%**
	16/255	42.6%	38.0%	33.7%	**42.7%**

Table 3.3: Results of robust accuracy under the less complex iteration-based adversarial attacks with ℓ_∞-norm on the dataset of CIFAR-10. The top-performing results are highlighted in bold.

	Attack Strength			
	$\varepsilon = 2/255$	$\varepsilon = 4/255$	$\varepsilon = 8/255$	$\varepsilon = 16/255$
			PGD^{10}-ℓ_∞	
MART	79.7%	75.6%	65.5%	43.3%
RobNet	78.3%	73.4%	62.1%	38.6%
BPFC	75.7%	68.2%	52.0%	26.9%
EMSF^{2}Net (Ours)	**82.1%**	**80.6%**	**74.8%**	**53.8%**
			MI-FGSM^{10}	
MART	78.3%	72.4%	59.1%	32.9%
RobNet	76.8%	70.1%	55.7%	27.8%
BPFC	73.4%	63.2%	44.5%	20.5%
EMSF^{2}Net (Ours)	**82.1%**	**81.0%**	**77.5%**	**66.5%**
			DI^2-FGSM^{10}	
MART	79.9%	76.1%	66.6%	45.8%
RobNet	78.6%	74.0%	63.3%	40.9%
BPFC	76.1%	69.4%	53.6%	29.1%
EMSF^{2}Net (Ours)	**81.2%**	**79.5%**	**73.8%**	**57.2%**
			EOTPGD^{10}	
MART	79.7%	75.6%	65.4%	43.5%
RobNet	78.3%	73.3%	62.1%	38.4%
BPFC	75.7%	68.3%	52.0%	26.9%
EMSF^{2}Net (Ours)	**82.1%**	**80.7%**	**74.7%**	**54.1%**

Table 3.4: Results of robust accuracy under the more complex iteration-based adversarial attacks with ℓ_∞-norm on the dataset of CIFAR-10. The top-performing results are highlighted in bold.

	Attack Strength			
	$\varepsilon = 2/255$	$\varepsilon = 4/255$	$\varepsilon = 8/255$	$\varepsilon = 16/255$
			PGD^{20}-ℓ_∞	
MART	78.3%	72.2%	57.7%	27.7%
RobNet	76.7%	69.7%	53.9%	22.5%
BPFC	73.3%	61.9%	39.9%	13.0%
EMSF^{2}Net (Ours)	**81.4%**	**79.0%**	**70.2%**	**45.9%**
			MI-FGSM^{20}	
MART	78.3%	72.0%	57.3%	26.6%
RobNet	76.7%	69.6%	53.5%	21.8%
BPFC	73.2%	61.8%	39.7%	12.9%
EMSF^{2}Net (Ours)	**81.5%**	**79.0%**	**69.6%**	**50.5%**
			DI^2-FGSM^{20}	
MART	78.6%	73.0%	59.1%	30.1%
RobNet	77.0%	70.5%	55.5%	24.5%
BPFC	73.8%	63.5%	41.9%	14.5%
EMSF^{2}Net (Ours)	**80.2%**	**76.9%**	**68.2%**	**45.1%**
			EOTPGD^{20}	
MART	78.3%	72.2%	57.9%	27.6%
RobNet	76.7%	69.7%	53.7%	22.5%
BPFC	73.3%	62.0%	39.8%	13.1%
EMSF^{2}Net (Ours)	**81.5%**	**79.4%**	**70.5%**	**45.4%**

Table 3.5: Results of robust accuracy under adversarial attacks with ℓ_2-norm on the dataset of CIFAR-10. ε represents attack strength, IN denotes iteration number (*steps*), and the top-performing results are highlighted in bold.

Attack Method	Parameter	MART	RobNet	BPFC	EMSF^{2}Net (Ours)
PGD^{10}-ℓ_2	ε = 1.0	46.2%	42.8%	47.1%	**78.8%**
	ε = 2.0	17.2%	14.4%	23.0%	**72.4%**
	ε = 3.0	4.7%	3.9%	11.3%	**64.1%**
PGD^{20}-ℓ_2	ε = 1.0	37.5%	34.5%	41.7%	**74.1%**
	ε = 2.0	5.5%	4.9%	12.9%	**62.9%**
	ε = 3.0	0.4%	0.5%	3.3%	**52.0%**
PGD^{40}-ℓ_2	ε = 1.0	34.1%	31.5%	39.4%	**67.6%**
	ε = 2.0	3.1%	3.1%	8.7%	**50.7%**
	ε = 3.0	0.1%	0.1%	1.4%	**39.6%**
CW	IN=100	22.9%	12.9%	59.4%	**72.5%**
	IN=500	21.0%	10.8%	59.4%	**69.1%**
	IN=1000	20.9%	10.8%	59.4%	**67.7%**

3.5 Ablation Studies

A range of ablation studies was performed to more thoroughly demonstrate the adversarial robustness and effectiveness of the proposed framework in this section.

For conducting the ablation studies, two distinct methodologies were employed. The baseline method is the ResNet-50, as detailed in Table 3.1. To ensure an equitable comparison, three $\mathcal{L}_{PC}$ were integrated at the output of STAGEs 2 to 4. Additionally, to validate the effectiveness of the FE Block, we also developed a novel framework, Multi-Stage Feature Fusion Network (MSF^{2}Net). Diverging from EMSF^{2}Net, it excludes the FE Block at each stage but shares the same remaining structures with EMSF^{2}Net. During the training phase, the sum of $\mathcal{L}_{PC}$ and $\mathcal{L}_{CE}$ are utilized as the total loss functions for the baseline and MSF^{2}Net.

Initially, Subsection 3.5.1 distinctly presents the robust performance of the baseline, MSF^{2}Net, and EMSF^{2}Net when subjected to various adversarial attacks, each with differing parameter configurations, using line graphs for illustration. Subsequently, Subsection 3.5.2 aims to showcase, through histograms, the classification accuracy per class within the CIFAR-10 dataset when subjected to various attacks, using the three approaches. This subsection uncovers the susceptibility of specific classes whithin the CIFAR-10 dataset to misclassification under these methods. Additionally, Subsection 3.5.3 employs the interpretability tool, grad-cam, for visualizing the distinct stages (STAGEs 1–4) associated with these three approaches. This subsection also uncovers the specific features that these methods concentrate when subjected to adversarial attacks. Consequently, we can comprehend why the proposed framework maintains excellent robustness in the face of adversarial attacks. Ultimately, the distributions of the features of the three methodologies when subjected to variously configured adversarial attacks are demonstrated using another renowned interpretability method, t-SNE.

3.5.1 Performance on Baseline, MSF^{2}Net, and EMSF^{2}Net

Classification accuracies for the baseline, MSF^{2}Net, and EMSF^{2}Net are discussed in this subsection. The experiments regarding the dataset of CIFAR-10 were conducted under the adversarial attacks with both ℓ_∞-norm and ℓ_2-norm, and these attacks were set to the configuration of white box. Initially, Fig. 3.3 shows the accuracies against the adversarial attacks with ℓ_∞-norm.

For these attacks, 2/255, 4/255, 8/255, and 16/255 were respectively selected as the values of the parameters of attack strength, which is denoted as ε. In addition, the remaining parameters were consistent with those outlined in Subsection 3.4.2. Subsequently, Fig. 3.4 illustrates the robust classification results of these three methodologies when subjected to white-box attacks with the ℓ_2-norm configuration. For the PGD-ℓ_2 attacks, 1.0, 2.0, and 3.0 were selected as the values of ε. Concurrently, for CW attacks, 100, 500, and 1000 were selected as the parameter of iteration number. In addition, the remaining parameters were consistent with those outlined in Subsection 3.4.3.

As illustrated in Fig. 3.3, the disparities among the three methodologies are subtle when subjected to the FGSM attack; however, the superiority of the proposed framework becomes increasingly apparent under the attacks based on iterations. Furthermore, the EMSF^{2}Net, as proposed, maintains superior performance compared to both the baseline and MSF^{2}Net when subjected to ℓ_2-norm adversarial attacks with white-box setting. In particular, under the CW attack, both the baseline and MSF²Net demonstrate robust accuracy results below 50%, in contrast to the EMSF^{2}Net, which maintains the robust accuracy still exceeding 60%. Consequently, Figs. 3.3 and 3.4 distinctly verify the effectiveness of FE Block.

3.5.2 Performance on Each Class of CIFAR-10

The accuracies corresponding to each class of CIFAR-10 of the baseline, MSF^{2}Net, and EMSF^{2}Net were output to delve deeper into the effects of the attacks with white-box setting. Figs. 3.5 and 3.6 depict these results. The accuracy per class without any attacks and the accuracies per class under the ℓ_∞-norm adversarial attacks with white-box setting are presented in Fig. 3.5, while the accuracy results per class under the ℓ_2-norm white-box attacks are shown in Fig. 3.6. In Fig. 3.5, a uniform attack strength of $\varepsilon = 0.04$ was applied using various attack methods: FGSM, PGD^{10}-ℓ_∞, MI-FGSM^{20}, DI^2-FGSM^{10}, and EOTPGD^{20}. Specifically, 0.004 was selected as the step size parameter of PGD^{10}-ℓ_∞. The step size of 0.004 and the momentum factor of 0.5 were selected in the cases of MI-FGSM^{20} and DI^2-FGSM^{10}. In addition, 0.004 and 5 were respectively selected as the parameters of the step size and mean gradient estimation in the case of EOTPGD^{20}. The PGD-ℓ_2 attacks, specifically PGD^{10}-ℓ_2, PGD^{20}-ℓ_2, and PGD^{40}-ℓ_2, alongside the CW attack, were utilized in Fig. 3.6. The parameters were configured with the

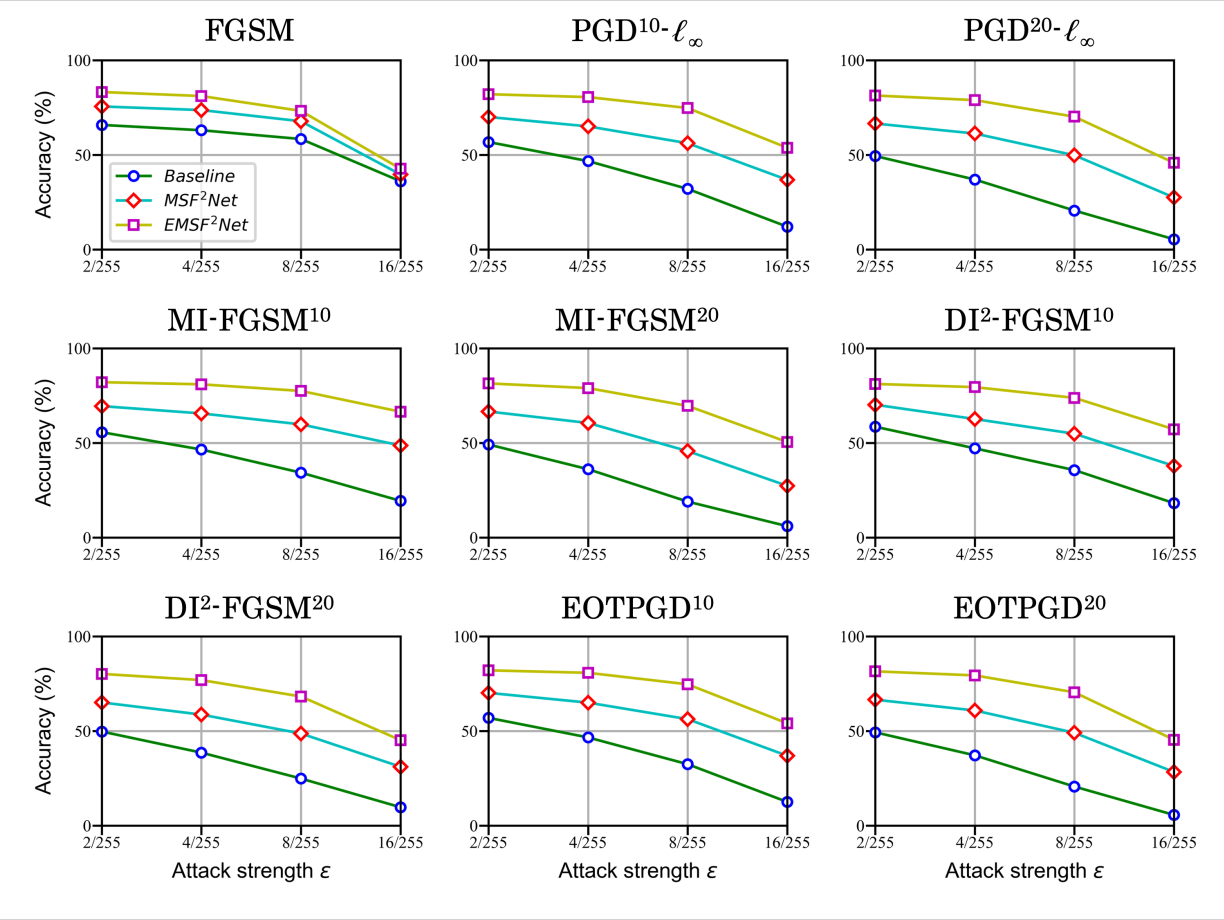


Figure 3.3: Results of the robust accuracy of the baseline, MSF^{2}Net, and EMSF^{2}Net. The adversarial attacks utilized were set as ℓ_∞-norm, with the evaluation on the dataset of CIFAR-10.

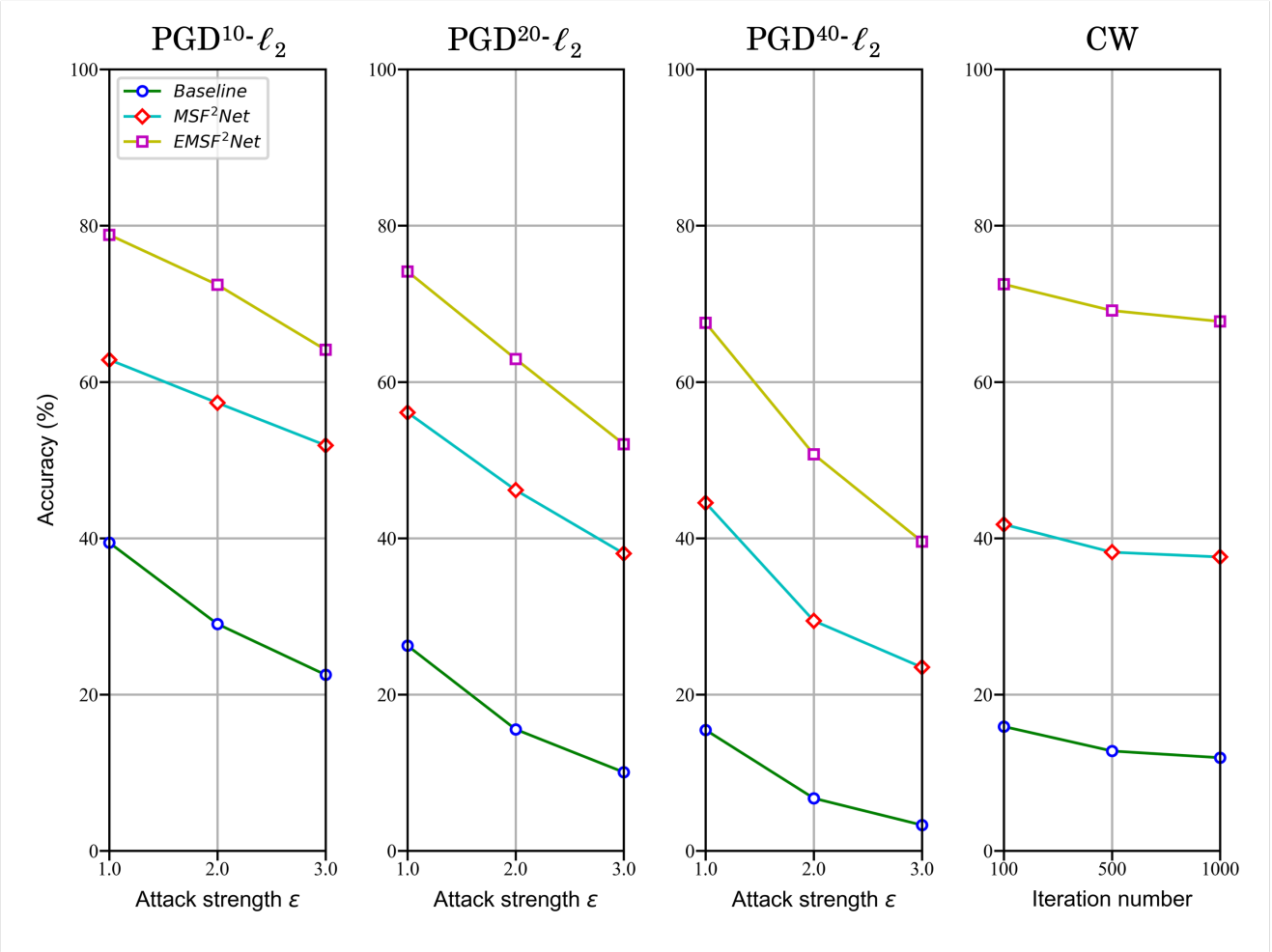


Figure 3.4: Results of the robust accuracy of the baseline, MSF^{2}Net, and EMSF^{2}Net. The adversarial attacks utilized were set as ℓ_2-norm, with the evaluation on the dataset of CIFAR-10.

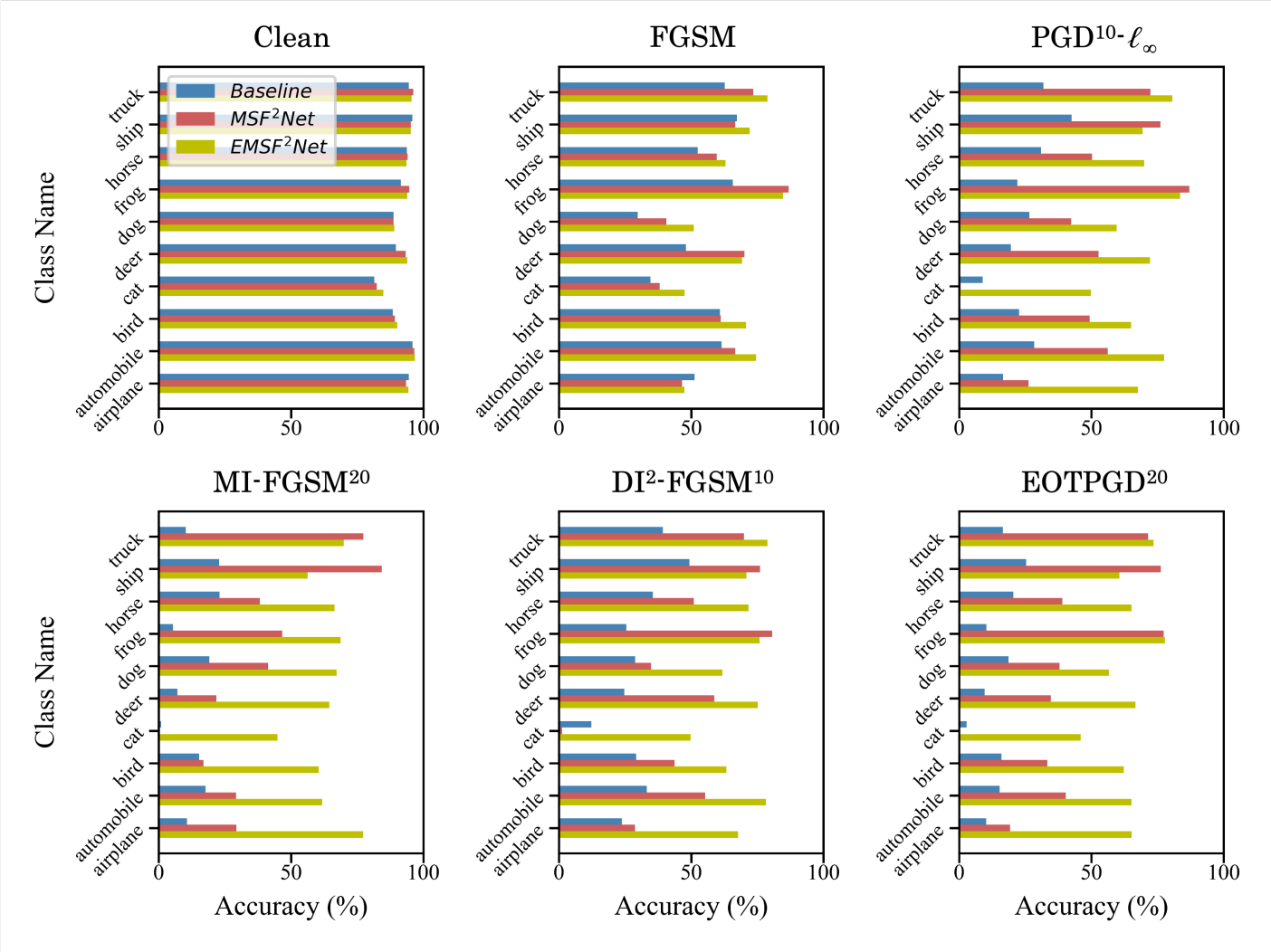


Figure 3.5: The accuracy results per class without any attacks and under the adversarial attacks with ℓ_∞-norm, with the evaluation of baseline, MSF^{2}Net, and EMSF^{2}Net on the dataset of CIFAR-10. In these sub-figures, the X-axis denotes the robust accuracy, while the Y-axis indicates the respective class name within the CIFAR-10 dataset.

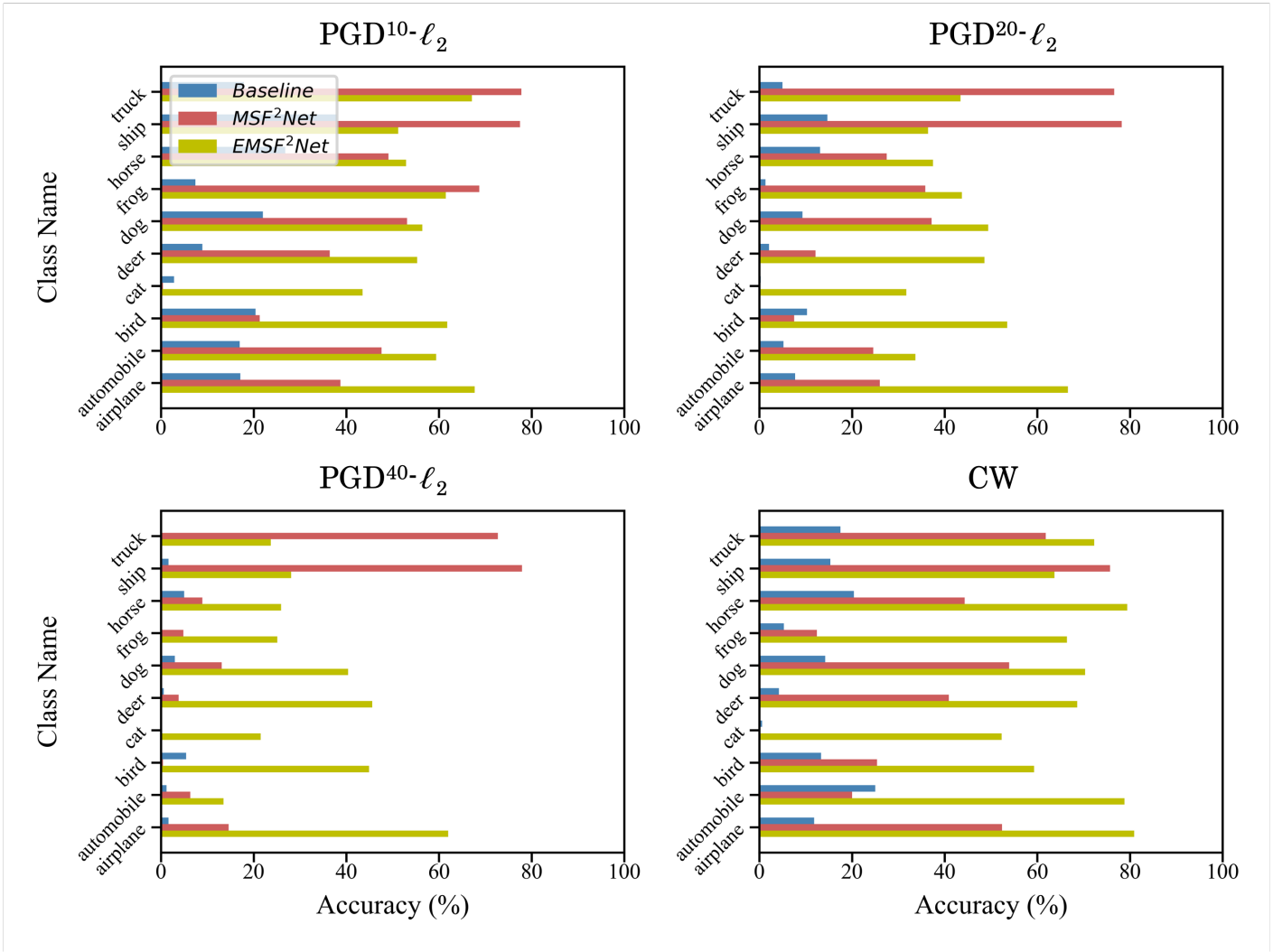


Figure 3.6: The accuracy results per class under the adversarial attacks with ℓ_2-norm, with the evaluation of baseline, MSF^{2}Net, and EMSF^{2}Net on the dataset of CIFAR-10. In these sub-figures, the X-axis denotes the robust accuracy, while the Y-axis indicates the respective class name within the CIFAR-10 dataset.

attack strength of 4.0 and the step size of 0.4 in the case of PGD-ℓ_2 attacks. Regarding the CW attack, the settings included the box-constraint parameter c at 1.0, the confidence κ set to 0, and the iteration number fixed at 400.

The accuracies per class without any attacks in the upper left part of Fig. 3.5 are depicted. This illustration reveals that, in the absence of any attacks, the accuracy per class remains virtually identical for all three methods. Nonetheless, the emergence of accuracy discrepancies among the three methodologies becomes evident when facing adversarial attacks. As depicted in Figs. 3.5 and 3.6, in the majority of cases, the proposed framework maintains comparable or superior robust performance. Particularly, while the performance of both baseline and MSF^{2}Net in the "cat" class is notably poor, the proposed framework consistently demonstrates high robust accuracy of classification. As for the cause of this phenomenon, we posit that images labeled as "cat" encompass more intricate structural details compared to those in other classes. Furthermore, as previously stated, the strong capability for information representation is exhibited in the features after FE Block of proposed framework. Consequently, these features are more adept at representing the complex information inherent in images labeled with "cat", in contrast to the baseline and MSF^{2}Net, which may not be as effective in representing such detailed information. Consequently, following the application of regularization, both the baseline and MSF^{2}Net exhibit poor performance in terms of adversarial robustness for the "cat" class.

3.5.3 Visualization by Grad-CAM

In this subsection, the utilization of grad-cam was employed to visualize the output features of STAGEs 1 to 4 across the trio of methods for comprehending the specific features that these methods focuses on under adversarial attacks. Moreover, we can reveal the mechanism about recognition of the trio of methods through this approach. In addition, the understanding can also be gained regarding the reasons behind the sustained high robustness of the proposed framework.

Visualization outcomes for the attacks with ℓ_∞-norm and attacks with ℓ_2-norm are respectively presented in Figs. 3.7 and 3.8, these attacks are both the white-box configuration. Additionally, in Figs. 3.7 and 3.8, "BS" signifies the baseline approach. Two specific attack methods named PGD^{10}-ℓ_∞ and EOTPGD^{20} were employed in the case of attacks with ℓ_∞-norm. In these attack approaches, 0.002 and 0.02 were respectively selected as the parameters of step size and attack

strength ε. Furthermore, in the case of EOTPGD^{20}, 5 was selected as the parameter of mean gradient estimation. In the case of adversarial attacks with ℓ_2-norm, the PGD-ℓ_2 attacks with various iterations (specifically PGD^{10}-ℓ_2, PGD^{20}-ℓ_2, and PGD^{40}-ℓ_2) and the CW attack renowned for its challenging nature were employed. The parameters were configured with the step size of 0.2 and attack strength of 2.0 in the case of PGD-ℓ_2 attacks. Regarding the CW attack, settings included the iteration number fixed at 500, confidence parameter with 0, and the box-constraint parameter at 1.0.

Analyzing Figs. 3.7 and 3.8, we observe that while the attention regions in STAGEs 1 to 3 are confusing accross the trio of approaches, in STAGE 4, the differences in the attention regions of these three methods can be readily observed. In particular, in STAGE 4, the regions of attention of both the baseline and MSF^{2}Net tend to be either misaligned from the target class or relatively large. Conversely, the crucial features of the target are always focused by the proposed framework. Our assertion is that in the case of larger attention regions in the non-denoising network, the likelihood of misclassification increases when an adversarial image is input. The contamination of texture features in adversarial images leads to the extraction of more incorrect features when attention is given to more regions. Conversely, the model may enhance the classification accuracy when the features with the highest importance in the adversarial example are always focused and extracted by the model. This is attributable to the fact that these important features are typically less contaminated by the adversarial noises.

Furthermore, in considering the significance of specific features within the CNN, observations from Figs. 3.7 and 3.8 reveal that the focus areas of STAGEs 1 to 3 (shallow layers) do not align with the target areas. However, the ultimate results of classification can be potentially influenced by the output features produced by STAGE 4 (deep layers). Consequently, it is inferred that these deeper layers hold greater importance and are more apt for regularization.

3.5.4 Visualization by t-SNE

In this subsection, visualization of the output features from the last layer of the three methods is conducted using the t-SNE technique. Therefore, the examination of the feature distributions within the baseline, MSF^{2}Net, and EMSF^{2}Net is facilitated. The results of visualization in various scenarios, which include absence of attacks, ℓ_∞-norm attacks with less complexity, ℓ_∞-norm

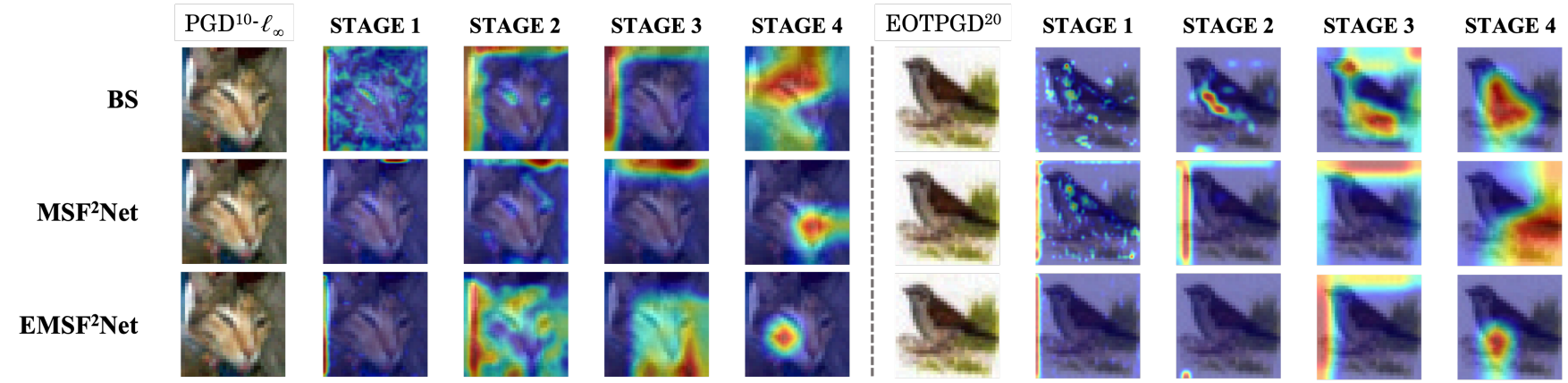


Figure 3.7: Visualization results using Grad-CAM, with the utilization of the ℓ_∞-norm adversarial attacks for baseline, MSF^{2}Net, and EMSF^{2}Net. Original data was selected from CIFAR-10. The darker red areas indicate that the models pay more attention to those parts.

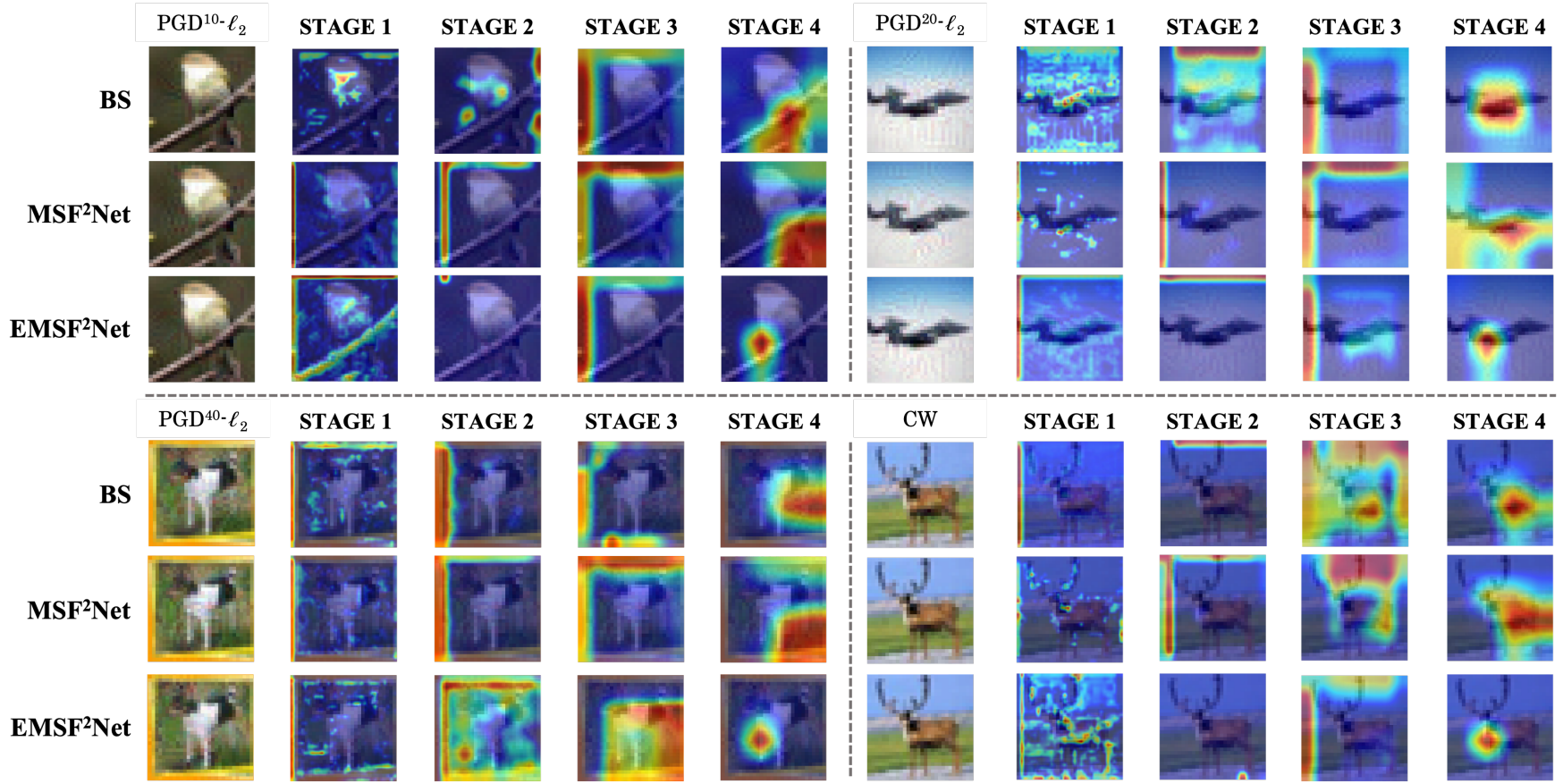


Figure 3.8: Visualization results using Grad-CAM, with the utilization of the ℓ_2-norm adversarial attacks for baseline, MSF^{2}Net, and EMSF^{2}Net. Original data was selected from CIFAR-10. The darker red areas indicate that the models pay more attention to those parts.

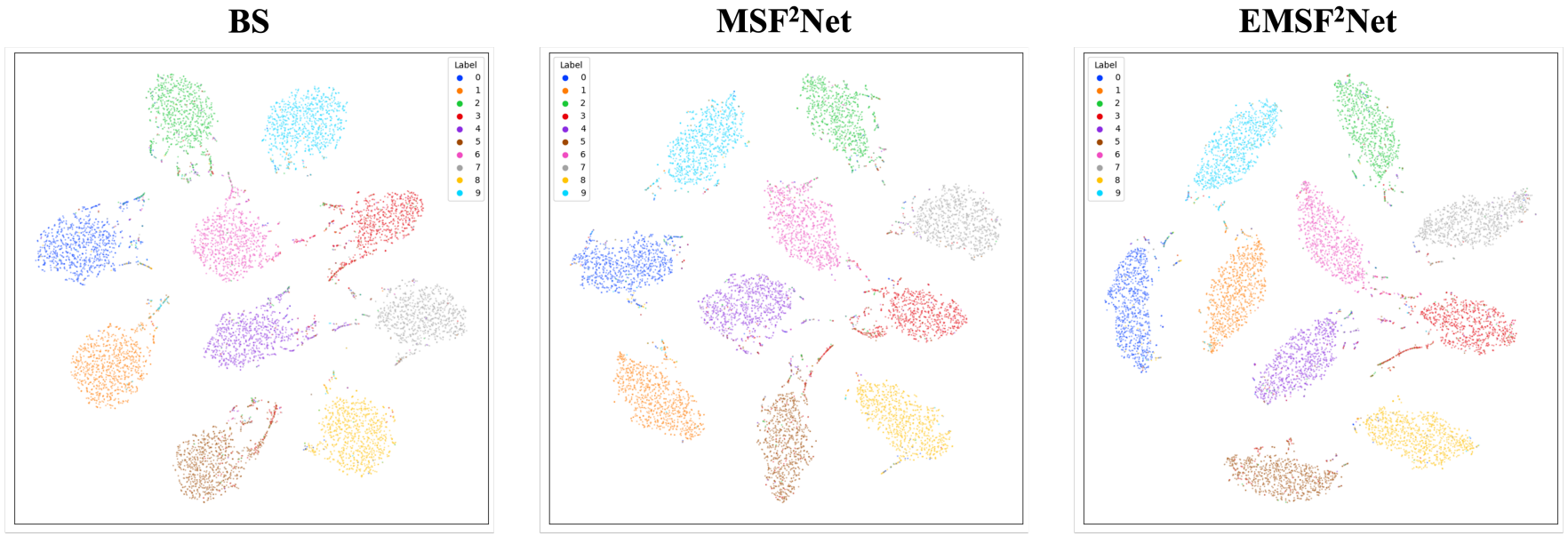


Figure 3.9: Visualization results using t-SNE, with the condition of no attacks for the baseline, MSF^{2}Net, and EMSF^{2}Net. Experiments were performed on CIFAR-10.

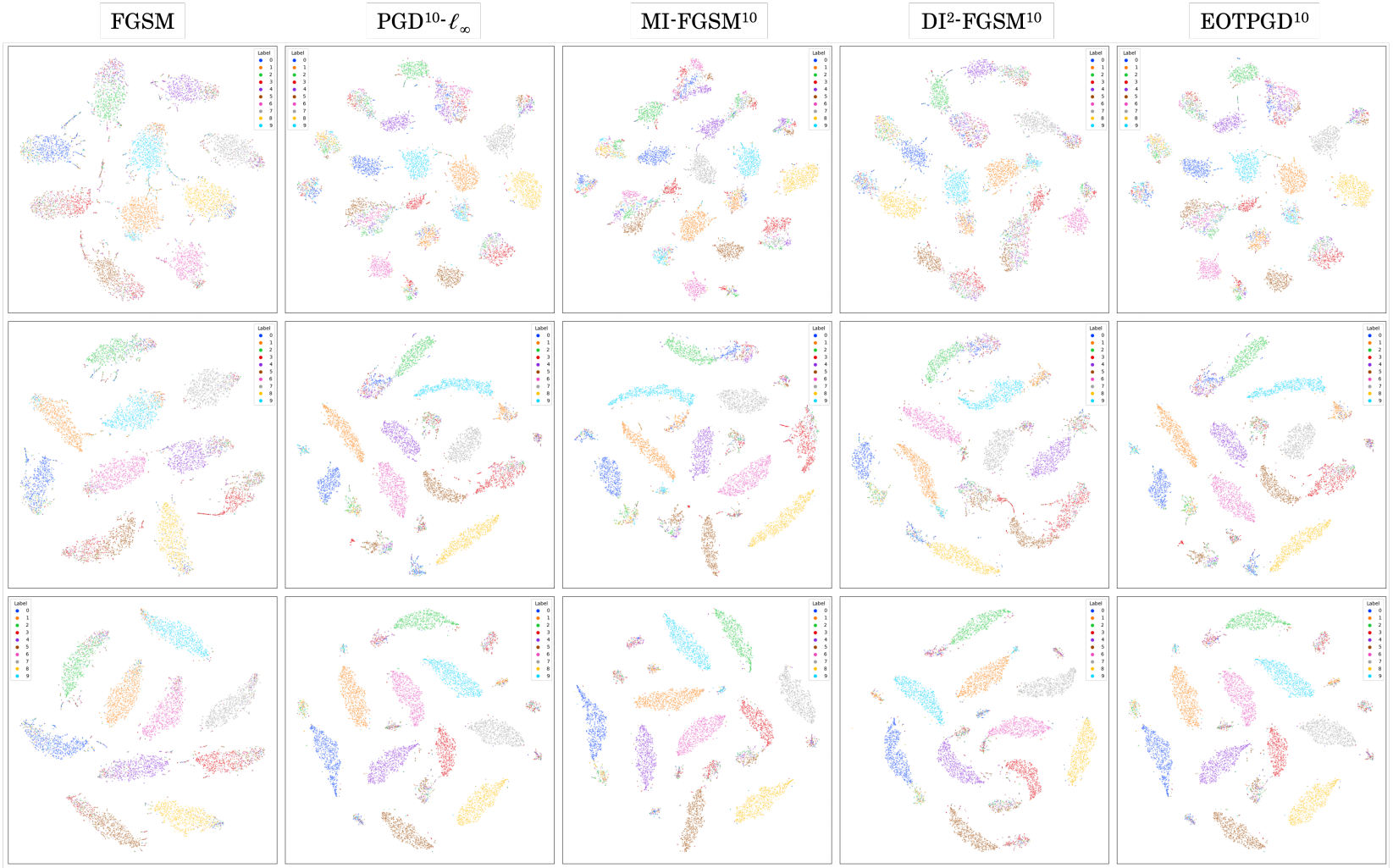


Figure 3.10: Visualization results using t-SNE, with the utilization of the less complex ℓ_∞-norm attacks for the baseline, MSF^{2}Net, and EMSF^{2}Net. Experiments were performed on CIFAR-10.

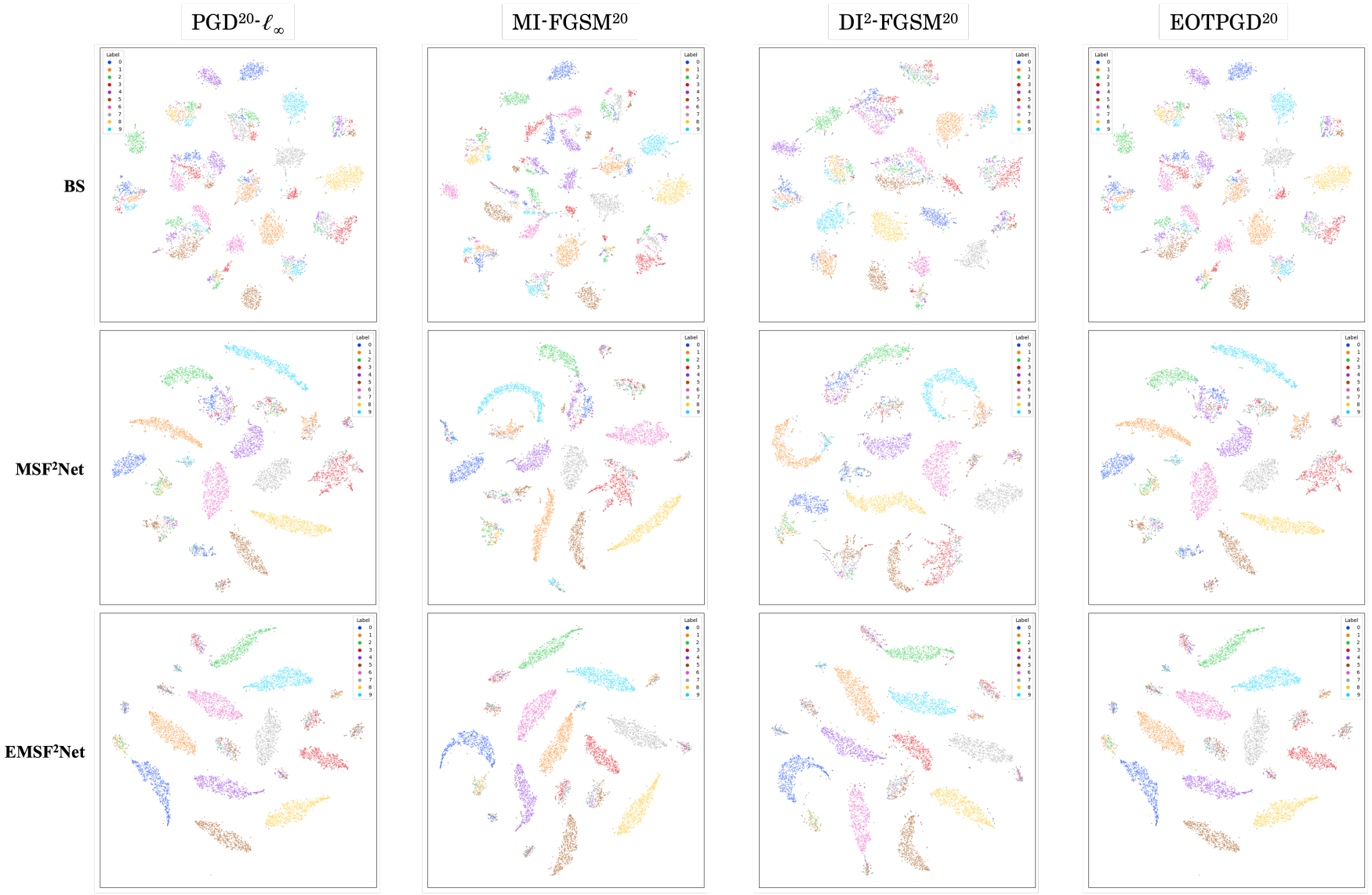


Figure 3.11: Visualization results using t-SNE, with the utilization of the more complex ℓ_∞-norm attacks for the baseline, MSF^{2}Net, and EMSF^{2}Net. Experiments were performed on CIFAR-10.

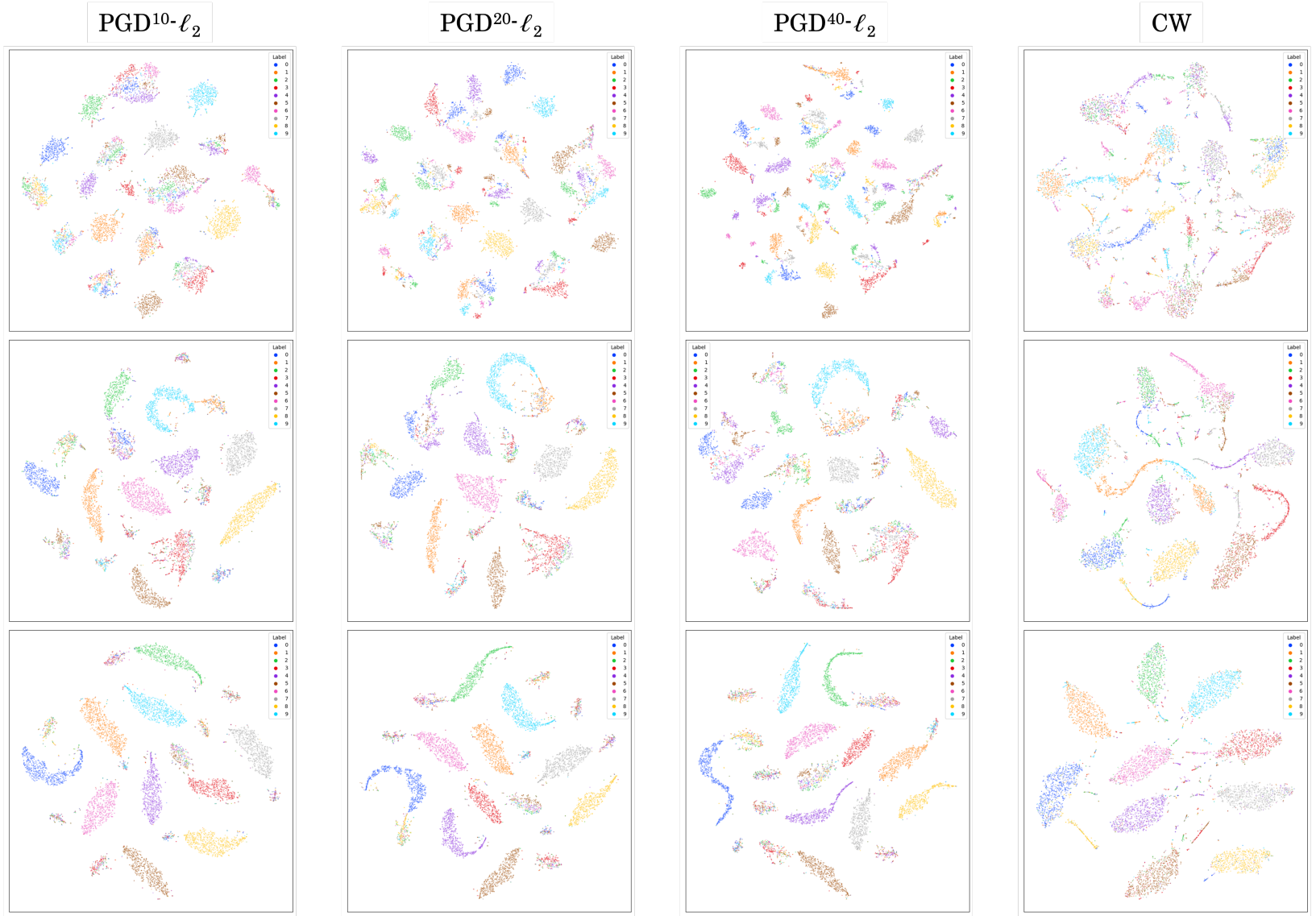


Figure 3.12: Visualization results using t-SNE, with the utilization of the ℓ_2-norm attacks for the baseline, MSF^{2}Net, and EMSF^{2}Net. Experiments were performed on CIFAR-10.

attacks with more complexity, adversarial attacks with ℓ_2-norm, are respectively presented in Figs. 3.9 – 3.12. In these figures, the baseline approach is signified as "BS", while the white-box configuration was applied to the employed adversarial attacks. The numbers 0 through 9 in these figures correspond to the CIFAR-10 dataset classes, sequentially representing "airplane", "automobile", "bird", "cat", "deer", "dog", "frog", "horse", "ship", and "truck".

In Fig. 3.10, various less complex attack methodologies, which include FGSM, PGD^{10}-ℓ_∞, MI-FGSM^{10}, DI^2-FGSM^{10}, and EOTPGD^{10}, were utilized. An identical attack strength with the value $\varepsilon = 4/255$ was applied to these attacks. 5, 0.5, and $\varepsilon/10$ were respectively configured as the parameters of mean gradient estimation number, momentum factor, and step size. In Fig. 3.11, PGD^{20}-ℓ_∞, MI-FGSM^{20}, DI^2-FGSM^{20}, and EOTPGD^{20} were selected as the ℓ_∞-norm attacks with more complexity. Additionally, all parameters were consistent with those selected in Fig. 3.10, with the exception of the number of iteration. Lastly, for the adversarial attacks with ℓ_2-norm in Fig. 3.12, we adopted the PGD-ℓ_2 attack with varying iteration numbers, specifically PGD^{10}-ℓ_2, PGD^{20}-ℓ_2, and PGD^{40}-ℓ_2, alongside the CW attack. In the case of PGD-ℓ_2 attacks, 0.1 and 1.0 were respectively selected as the parameters of step size and attack strength. In addition, regarding the CW attack, 50, 0, and 1.0 were respectively selected as its parameters of iteration number, confidence, and box-constraint.

As we can see from Fig. 3.9, three methods have relatively distinct clustering boundaries for each class of CIFAR-10. This suggests that in the absence of attacks, the CIFAR-10 dataset can be effectively classified by these approaches. Under diverse adversarial attacks, distinctions among the three methodologies become evident. Observations from Figs. 3.10 – 3.12 reveal that, when subjected to attacks, the classification results of baseline are considerably disordered. In contrast, the MSF^{2}Net exhibits a marginally improved performance under these conditions. Conversely, in a majority of scenarios, the clear classification boundaries can be consistently maintained by the proposed framework, effectively showcasing its effectiveness and adversarial robustness.

3.6 Conclusions

Our focus was on adversarial defense employing the regularization technique in this chapter. It was noted that the current adversarial defense approaches based on regularization lack com-

prehensive analysis regarding the specific features that would enhance the regularization effectiveness, thereby bolstering the defense performance. In other words, sufficiently robust features are not integrated into the image classification system in the current regularization-based defense methods. Consequently, a novel CNN-based framework, EMSF^{2}Net, has been proposed, characterized by three foundational components: MSFE, MSF, and regularization. The demonstrated effectiveness of the proposed framework suggests a notable enhancement in CNN robustness through the regularization of enhanced multi-stage fusion features. Moreover, it can introduce the robust regularization-based enhanced multi-stage fusion features. A comprehensive series of comparative experiments and ablation studies, focusing on white-box adversarial attacks under various configurations, have validated the adversarial robustness and effectiveness of EMSF^{2}Net. This is attributed to the similar mechanisms of visual information processing inherent in diverse CNN-based architectures. We held the view that the operations like convolution are utilized in the architectures based on CNNs to extract the local data correlations, thereby effectively learning the representation information per class. Consequently, it is our conviction that EMSF^{2}Net exhibits comparable performance in other structures based on CNNs.

Chapter 4

Integration of Robust Caption Features from Stable RNN into Image Captioning System

4.1 Introduction

This chapter introduces our work on constructing the adversarially robust retrieval-based image captioning system. The part of text feature processing within the retrieval-based image captioning system is our primary emphasis in this chapter. A kind of robust and novel structure of the recurrent neural network (RNN), termed Lipschitz RNN (LRNN) [105], is incorporated into the retrieval-based image captioning system. Specifically, LRNN is grounded in the continuous dynamic system theory. It is widely acknowledged that the vanilla RNN is plagued by issues like vanishing and exploding gradients, leading to a lack of robustness and stability in its performance. In contrast, LRNN distinguishes itself from the vanilla RNN through its unique hidden-to-hidden weight matrices, which endow it with an enhanced capacity for expression. Subsequently, the approach proposed by Shi et al. [106] is employed to produce the attacked image captions with noun type, which are then considered as the adversarial examples in our study. In the inference phase, we integrate the adversarial examples with the clean captions to form the new retrieval databases to validate the performance of the proposed method when facing adversarial attacks. The experimental results conclusively demonstrate that, in the case of noun-type adversarial attacks, the proposed approach exhibits greater robustness compared to the retrieval-based image captioning system with the vanilla RNN.

4.2 Proposed Approach

The implementation details of the proposed retrieval-based image captioning system are delved into in this section. This encompasses an overview of the entire framework of the image captioning network, followed by an in-depth discussion of the technical specifics of the LRNN.

4.2.1 Overall Framework

The framework of the proposed image captioning system based on retrieval, depicted in Fig. 4.1, consists of two principal components: the image feature extraction and the caption feature extraction parts. The process of obtaining the image embedding involves the utilization of the part of image feature extraction. Specifically, the features extracted from the image embedding part of VSE (Visual Semantic Embedding), proposed in the work [107], are used as the image embedding in this chapter.

Regarding the part of the caption feature extraction, the method of adversarial attack with noun type [106] is initially applied to perform the attack operation on the original captions from MS-COCO dataset [94], thereby generating the adversarial (attacked) captions. Following this, these attacked captions, along with the original embedding part that encompasses MS-COCO caption features with pre-computation from the work [107], are fed into the module of LRNN. Subsequently, the embeddings for both the attacked caption and the original caption are available. Finally, two types of hinge loss functions are calculated: one between the image embedding $\boldsymbol{I}_{emb}$ and the original caption embedding $\boldsymbol{C}_{emb}$, designated as $\mathcal{L}_1$, and the other encompassing the image embedding $\boldsymbol{I}_{emb}$, original caption embedding $\boldsymbol{C}_{emb}$, and attacked caption embedding $\widetilde{\boldsymbol{C}}_{emb}$, designated as $\mathcal{L}_2$. By integrating these hinge loss functions, the total loss function $\mathcal{L}_{total}$ for the proposed framework can be derived. Therefore, the equations of $\mathcal{L}_{total}$, $\mathcal{L}_1$, and $\mathcal{L}_2$ can be expressed as follows:

$$\mathcal{L}_{total} = \mathcal{L}_1(\boldsymbol{I}_{emb}, \boldsymbol{C}_{emb}) + \mathcal{L}_2(\boldsymbol{I}_{emb}, \boldsymbol{C}_{emb}, \widetilde{\boldsymbol{C}}_{emb}), \tag{4.1}$$

$$\begin{aligned}\mathcal{L}_1(\boldsymbol{I}_{emb}, \boldsymbol{C}_{emb}) = &\max_{\boldsymbol{C}'_{emb} \neq \boldsymbol{C}_{emb}} \left[\alpha + \mathrm{Sim}(\boldsymbol{I}_{emb}, \boldsymbol{C}'_{emb}) - \mathrm{Sim}(\boldsymbol{I}_{emb}, \boldsymbol{C}_{emb})\right]_+ \\ &+ \max_{\boldsymbol{I}'_{emb} \neq \boldsymbol{I}_{emb}} \left[\alpha + \mathrm{Sim}(\boldsymbol{I}'_{emb}, \boldsymbol{C}_{emb}) - \mathrm{Sim}(\boldsymbol{I}_{emb}, \boldsymbol{C}_{emb})\right]_+ ,\end{aligned} \tag{4.2}$$

$$\mathcal{L}_2(\boldsymbol{I}_{emb}, \boldsymbol{C}_{emb}, \widetilde{\boldsymbol{C}}_{emb}) = \max_{\widetilde{\boldsymbol{C}}_{emb}} \left[\alpha + \mathrm{Sim}(\boldsymbol{I}_{emb}, \widetilde{\boldsymbol{C}}_{emb}) - \mathrm{Sim}(\boldsymbol{I}_{emb}, \boldsymbol{C}_{emb})\right]_+, \tag{4.3}$$

where α is a hyperparameter that controls the margin, and $\mathrm{Sim}(\cdot,\cdot)$ represents the similarity measure function.

4.2.2 Details of LRNN

As previously mentioned, the LRNN module is developed from the theory of dynamic system. The expressions of LRNN can be analytically described as follows in line with dynamical system theory,

$$\begin{cases} \dot{\boldsymbol{h}} = \boldsymbol{A}_{\beta_1,\gamma_1}\boldsymbol{h} + \tanh\left(\boldsymbol{W}_{\beta_2,\gamma_2}\boldsymbol{h} + \boldsymbol{U}\boldsymbol{x} + \boldsymbol{b}\right) \\ \boldsymbol{y} = \boldsymbol{D}\boldsymbol{h} \end{cases}, \tag{4.4}$$

where the input and output of the LRNN module are represented by $\boldsymbol{x} = \boldsymbol{x}(t) \in \mathbb{R}^{n_1}$ and $\boldsymbol{y} = \boldsymbol{y}(t) \in \mathbb{R}^{n_2}$, respectively. In Eq. 4.4, the hidden state at time t is denoted by $\boldsymbol{h} = \boldsymbol{h}(t) \in \mathbb{R}^N$, while the temporal derivative of $\boldsymbol{h}(t)$ is represented by $\dot{\boldsymbol{h}} = \partial \boldsymbol{h}(t)/\partial t$. The encoder and decoder matrices are respectively denoted by $\boldsymbol{U} \in \mathbb{R}^{N\times n_1}$ and $\boldsymbol{D} \in \mathbb{R}^{N\times n_2}$, while the offset is represented by $\boldsymbol{b}$. The hidden-to-hidden weight matrices are represented by $\boldsymbol{A}_{\beta_1,\gamma_1} \in \mathbb{R}^{N\times N}$ and $\boldsymbol{W}_{\beta_2,\gamma_2} \in \mathbb{R}^{N\times N}$. Therefore, $\boldsymbol{A}$ and $\boldsymbol{W}$ are defined as follows:

$$\begin{cases} \boldsymbol{A} = (1-\beta_1)\left(\boldsymbol{M}_1 + \boldsymbol{M}_1^\top\right) + \beta_1\left(\boldsymbol{M}_1 - \boldsymbol{M}_1^\top\right) - \gamma_1\mathcal{I} \\ \boldsymbol{W} = (1-\beta_2)\left(\boldsymbol{M}_2 + \boldsymbol{M}_2^\top\right) + \beta_2\left(\boldsymbol{M}_2 - \boldsymbol{M}_2^\top\right) - \gamma_2\mathcal{I} \end{cases}, \tag{4.5}$$

where β_1 and β_2 are hyperparameters within the range $[0, 1]$, while γ_1 and γ_2 are positive hyperparameters exceeding zero. In addition, $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ denote the trainable matrices, and the identity matrix is denoted by $\mathcal{I}$.

Learning weights $\boldsymbol{A}_{\beta_1,\gamma_1}$, $\boldsymbol{W}_{\beta_2,\gamma_2}$, $\boldsymbol{U}$ and $\boldsymbol{b}$ constitutes the primary objective in LRNN training. It is necessary to discretize the analytical expression to attain this objective. Let $\boldsymbol{h}_t$ be the hidden state at time t, and Δt the interval of time, the analytical solution to the differential equation

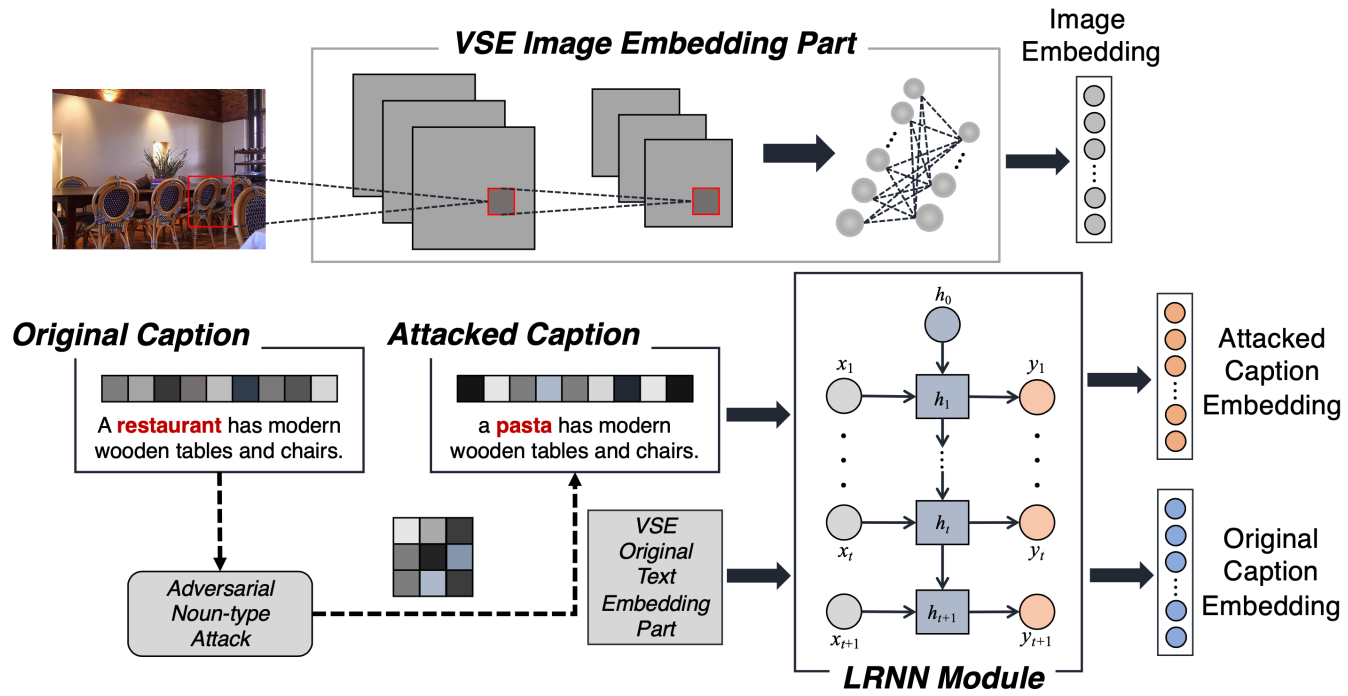


Figure 4.1: The proposed retrieval-based image captioning system, with the design of implementing the defense against the adversarial attack with noun type. In addition, the visual semantic embedding is represented by VSE in the figure.

presented in Eq. 4.4 can be expressed as follows:

$$\boldsymbol{h}_{t+1} = \boldsymbol{h}_t + \int_t^{t+\Delta t} \boldsymbol{A}\boldsymbol{h}(s) + \tanh\left(\boldsymbol{W}\boldsymbol{h}(s) + \boldsymbol{U}\boldsymbol{x}(s) + \boldsymbol{b}\right) \mathrm{d}s. \tag{4.6}$$

Utilizing the forward Euler method allows for the rephrasing of the aforementioned equation as follows:

$$\boldsymbol{h}_{t+1} = \boldsymbol{h}_t + \Delta t \cdot \boldsymbol{A}\boldsymbol{h}_t + \Delta t \cdot \tanh\left(\boldsymbol{W}\boldsymbol{h}_t + \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{b}\right). \tag{4.7}$$

The aforementioned equation can be further streamlined as follows for facilitate training:

$$\boldsymbol{h}_{t+1} = \boldsymbol{h}_t + \alpha \cdot \epsilon \cdot \boldsymbol{A}\boldsymbol{h}_t + \epsilon \cdot \tanh\left(\boldsymbol{W}\boldsymbol{h}_t + \mathrm{Linear}\left(\boldsymbol{x}_t\right)\right), \tag{4.8}$$

where the fully connected layer is denoted by Linear$(\cdot)$, while ϵ and α serve as hyperparameters.

4.2.3 Gate LRNN

In the part of LRNN module, the Gate LRNN, a member of the LRNN family, is utilized in this chapter. The hidden state $\widetilde{\boldsymbol{h}}_t$ in the Gate LRNN is characterized by the following recurrence relation:

$$\widetilde{\boldsymbol{h}}_{t+1} = \widetilde{\boldsymbol{h}}_t + \epsilon \cdot \left(\alpha \cdot \boldsymbol{A}\widetilde{\boldsymbol{h}}_t + \tanh\left(\boldsymbol{m}_t\right)\right) \cdot \mathrm{Sigmoid}\left(\boldsymbol{m}_t\right), \tag{4.9}$$

where $\boldsymbol{m}_t = \boldsymbol{W}\widetilde{\boldsymbol{h}}_t + \mathrm{Linear}\left(\boldsymbol{x}_t\right)$. In contrast to the classic LRNN, the Gate LRNN demonstrates an enhanced capability to perceive variations in the middle-layer subtle features, thereby improving the feature identification ability of the network.

4.3 Experiments

4.3.1 Dataset & Attack Method

In the experiments, the popular public dataset MS-COCO [94] was used to validate the performance of the proposed method. In addition, the adversarial example generation strategy used in this chapter was the noun-type attack [106]. Key nouns within the clean captions are attacked by the approach of noun-type attack. Table 4.1 presents the examples of clean captions and

their corresponding adversarial captions generated through the noun-type attack method. The clean captions were selected from the MS-COCO dataset. This table distinctly illustrates that, following the noun-type attack, the key nouns in the original (clean) captions have undergone replacement. In the experiments, nearly all the caption retrieval databases contained the generated adversarial examples, and the queries were the images.

4.3.2 Baseline

During the experiments, we found that training the retrieval-based image captioning system with the vanilla RNN was difficult to converge, so we made improvements to it. Specifically, the hidden state of the vanilla RNN was written as the residual form first. Subsequently, the weight matrix $\boldsymbol{W}$, as referenced in the LRNN, was introduced into the residual form of the vanilla RNN, supplanting the linear layer of the hidden state $\boldsymbol{h}_t$ in the vanilla RNN. The RNN with such hidden state was referred to as the modified vanilla RNN in this chapter. In addition, we regarded the retrieval-based image captioning system with the modified vanilla RNN as the baseline method in the experiments.

4.3.3 Evaluation Metrics

In the experiments, Recall@K (R@K), a measurement benchmark commonly adopted for retrieval tasks, was employed. R@K is defined as the proportion of the number of relevant results in the top K retrieved results to the total number of relevant results in the database. The recall rate of the system can be measured by this metric. Specifically, the metrics R@1, R@5, and R@10, along with the average of these three indicators, were calculated, with the computation of R@1, R@5, and R@10 adhering to the paper [106].

Table 4.1: Some original captions from the dataset of MS-COCO and some adversarial examples (captions) generated using the noun-type attack method. The words in bold represent the original key nouns in the original caption, while the words in both bold and italic indicate the adversarial key nouns resulting from the noun-type attack on the original caption.

Original caption	Black and white **photograph** of two **men** on motorcycles.
Adversarial captions	Black and white ***greens*** of two men on motorcycles. Black and white ***drink*** of two men on motorcycles. Black and white photograph of two ***dresser*** on motorcycles. Black and white ***cement*** of two men on motorcycles. Black and white photograph of two ***grass*** on motorcycles.

Original caption	A **herd** of **sheep** standing on top of snow covered field.
Adversarial captions	A ***sauce*** of sheep standing on top of snow covered field. A ***garage*** of sheep standing on top of snow covered field. A ***donut*** of sheep standing on top of snow covered field. A herd of ***steam*** standing on top of snow covered field. A herd of ***surfer*** standing on top of snow covered field.

Original caption	Two ducks floating together on a **body** of **water**.
Adversarial captions	Two ducks floating together on a body of ***rock***. Two ducks floating together on a ***sheep*** of water. Two ducks floating together on a ***sofa*** of water. Two ducks floating together on a body of ***pipe***. Two ducks floating together on a body of ***telephone***.

Original caption	The small **bathroom** has wooden cabinets around the **sink**.
Adversarial captions	The small ***machine*** has wooden cabinets around the sink. The small ***stadium*** has wooden cabinets around the sink. The small bathroom has wooden cabinets around the ***electronics***. The small ***pasta*** has wooden cabinets around the sink. The small ***sun*** has wooden cabinets around the sink.

Table 4.2: Experimental results of the proposed method (PM) and the baseline, where PM represents the retrieval-based image captioning system with the Gate LRNN, while baseline represents the retrieval-based image captioning system with the modified vanilla RNN.

Method	**Baseline**	**PM**
R@1	**25.0%**	23.6%
R@5	50.9%	**53.0%**
R@10	62.2%	**65.3%**
Average	46.0%	**47.3%**

4.3.4 Experimental Results

In the experiments, values of 0.01 and 1 were respectively selected as the parameters of ϵ and α. In the weight matrices $\boldsymbol{A}$ and $\boldsymbol{W}$, parameters β_1, β_2 were assigned the value of 0.8, while γ_1, γ_2 were set at 0.01. Table 4.2 shows the experimental results of the proposed approach, as well as the baseline method. This table distinctly demonstrates that the proposed method can achieve a broader range of retrieval results, which also shows the proposed method outperforms the retrieval-based image captioning system with the modified vanilla RNN, thereby confirming its effectiveness and robustness.

4.4 Conclusions

A novel method was proposed to realize the defense of retrieval-based image captioning system against adversarial attacks in this chapter. Specifically, a kind of novel RNN with stability and robustness, grounded in the theory of dynamic system, was integrated into the retrieval-based image captioning system. The adversarially robust caption features generated from a kind of stable RNN structure can be introduced into the retrieval-based image captioning system, thereby enhancing its robustness. The robustness and effectiveness of the proposed approach were conclusively affirmed by the results from the noun-type attack experiments on the image captions.

Chapter 5

Integration of Robust Caption Features from Pre-trained Models into Image Classification System

5.1 Introduction

In most existing adversarial defense methods for image classification, only information within the input images is utilized, meaning that no additional information is being employed. Conversely, the human visual system is inherently multimodal, meaning that prior heterogeneous multimodal features are always employed by humans to facilitate object perception. This mechanism may contribute to preventing the human visual system from misclassifying adversarial examples. Image captioning features, encompassing the information with detailed descriptions of an image, contain more details than classic image features. Hence, our hypothesis posits that caption features from pre-trained models exhibit greater robustness compared to conventional image features when facing adversarial examples. In order to substantiate this hypothesis, our initial step involves subjecting VGG-13 [3] to the Projected Gradient Descent (PGD) attack, thereby producing adversarial examples. Subsequently, these examples are input into ResNet-50 [1] and pre-trained ClipCap [108] for the extraction of image and caption features, respectively. We chose the images from the CIFAR-10 [79] test dataset for performing the attack. Fig. 5.1 shows some examples. As illustrated in this figure, ClipCap demonstrates the capability to the accurate object perception of the adversarial example, which proves that caption features from the pre-trained models are adversarially robust. Furthermore, we also have confirmed that the ro-

bust caption features can improve the robustness of the retrieval-based image captioning system. Hence, one can anticipate the enhancement in the adversarial robustness of the image classification system with the integration of robust caption features. Furthermore, to the best of our understanding, the integration of caption features into adversarial defenses within the realm of image classification has not yet been realized by others.

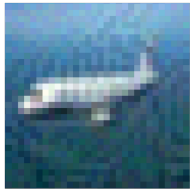
In this chapter, a novel method, termed heterogeneous feature fusion network (HFFN), is proposed. Specifically, HFFN is composed of a pair of sub-networks dedicated to image feature extraction and caption feature extraction. Then the hetero-center loss (HC loss) [109] is introduced, which plays the role in the "model gap" reduction between these image and caption features, facilitating the acquisition of superior fused features. In addition, the class center distances between image features and caption features can be limited by the HC loss, thereby enabling the learning of modal alignment information between these heterogeneous features. Additionally, for the purpose of adversarial defense, an effective loss function, which is called guided complement entropy loss (GCE loss) [110], is utilized to enhance the robustness of the fusion heterogeneous feature. The GCE loss method effectively defends against adversarial attacks, constraining the model prediction probability of incorrect classes, without any computational overhead increase. There is a more natural integration of caption features into the classic image features, thereby enhancing the ability of CNNs to effectively "assimilate" their robustness. Moreover, the GCE loss also exhibits the capability of suppressing the prediction probability for the classes that do not align with the ground truth to a certain extent, thereby enhancing the robustness. The contributions of this chapter can be categorized into the following three distinct aspects,

- Drawing inspiration from the human object perception process, we pioneer the integration of robust heterogeneous features, such as robust caption features, into the image classification system for adversarial defense. The name of the proposed method is HFFN.

- A kind of novel strategy for training termed multi-purpose loss calculation, which is designed to more effectively integrate caption features into CNNs built from visual data and fully leverage the robustness against adversarial attacks, is proposed.

- The effectiveness and robustness of the proposed HFFN are demonstrated by the comprehensive experimental results conducted across the attacks of black-box scenarios.

Label: airplane
ResNet-50: deer
ClipCap: The plane was flying at an altitude of metres above sea level.



Label: airplane
ResNet-50: ship
ClipCap: Aerial view of a passenger plane flying above the clouds.

Label: bird
ResNet-50: frog
ClipCap: A bird in the garden.

Label: cat
ResNet-50: dog
ClipCap: Cute cats playing on the floor.

Figure 5.1: The outputs of ResNet-50 and ClipCap when the inputs are adversarial examples.

5.2 Proposed Method

The domain of adversarial defense and the task of multimodality are concentrated by the proposed HFFN. Fig. 5.2 shows the framework of the proposed HFFN. In this framework, the heterogeneous features are extracted and integrated smoothly and naturally. The specifics of implementation of the proposed framework are delineated in two key parts: extraction of heterogeneous features and calculation of multi-purpose loss functions.

5.2.1 Extraction of Heterogeneous Features

Given that the previous study [111] has demonstrated the significant impact of highly expressive visual features on the adversarial robustness enhancement for the CNNs, the extraction and utilization of these features become crucial. Specifically, the ResNet-50 model is selected as the image feature extraction sub-network, owing to its proficiency in extracting visual features with expressive richness, as well as its widespread use as the backbone in the field of adversarial defense. Consequently, the calculation of the extracted feature with image modality is as follows:

$$\boldsymbol{F}^{\mathrm{i}} = \Im\left(\boldsymbol{X}\right), \tag{5.1}$$

where $\boldsymbol{X}$ represents the input image, while $\boldsymbol{F}^{\mathrm{i}}$ and $\Im\left(\cdot\right)$ correspond to the extracted feature of image modality and the image feature extraction sub-network.

Furthermore, the majority of current prevalent strategies for adversarial defense predominantly concentrate on the image domain, developing solutions with customization. In addition, it differs from the human visual system, which effectively utilizes information with multimodality. Consequently, a kind of novel robust caption feature is introduced, extracted by the sub-network of caption feature extraction. This sub-network, a portion of ClipCap, is depicted in detail in Fig. 5.2. This figure illustrates that two distinct components, which are the pre-trained CLIP [112] image encoder and the pre-trained multi-layer perceptron (MLP), constitute the caption feature extraction sub-network.

Utilizing the contrastive loss, the CLIP model, through training on a vast corpus of image-text pairs, aids the proposed HFFN in learning the common representation features that are not only aligned with the perception of visual language but also rich in semantics. Additionally, the

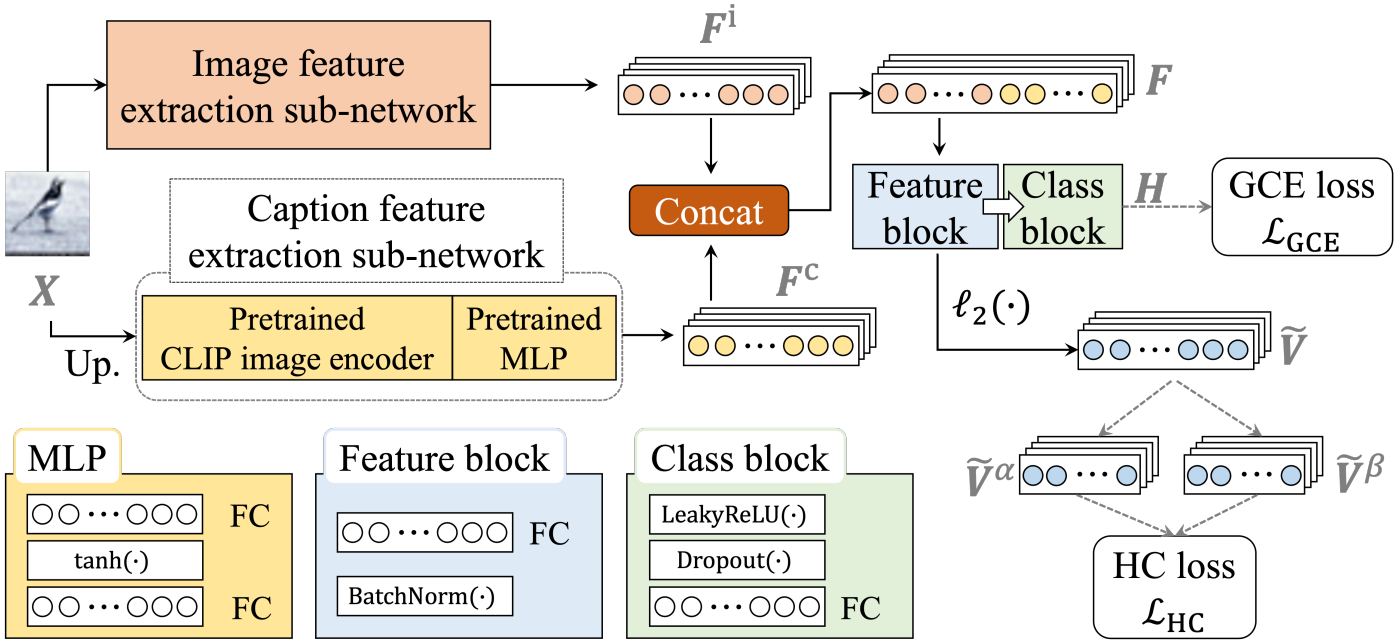


Figure 5.2: Framework of HFFN, where Up. signifies the upsampling operation, Concat denotes the operation of concatenation, and $\ell_2(\cdot)$ corresponds to ℓ_2 normalization.

common representation features are mapped into the domain of text through the application of the MLP model to acquire the robust caption features. A pair of fully-connected (FC) layers and an activation function layer $\tanh(\cdot)$ form this MLP model, and this MLP model has been trained the dataset of Conceptual Captions [113], which is famous for its complexity. Our conviction is that the utilization of the pre-trained model obtained by the training on the datasets with complexity, can enhance the capability of HFFN to extract caption features that are not only more robust but also demonstrate the superior performance of generalization. Consequently, the caption feature after mapping, which is represented by $\boldsymbol{F}^{\mathrm{c}}$, is formulated below:

$$\boldsymbol{F}^{\mathrm{c}} = \mathcal{M}\left(C\left(\mathcal{U}\left(\boldsymbol{X}\right)\right)\right), \tag{5.2}$$

where the image encoder of CLIP model is denoted by $C(\cdot)$, while the MLP model is signfied by $\mathcal{M}(\cdot)$. Furthermore, due to the requirement of input size for the CLIP model, the operation of upsampling $\mathcal{U}(\cdot)$ is utilized.

The concatenation operation is performed on the $\boldsymbol{F}^{\mathrm{i}}$ and $\boldsymbol{F}^{\mathrm{c}}$ to acquire the fused heterogeneous features, which can be utilized for the loss calculation. Subsequently, a feature block, which includes a single FC layer and a batch-norm layer, is employed for the fusion of the feature after concatenation. Consequently, the calculation of the fused feature $\boldsymbol{V} \in \mathbb{R}^{b \times d}$ can be formulated below:

$$\boldsymbol{V} = BN\left(FC\left(\boldsymbol{F}\right)\right), \tag{5.3}$$

where the feature after concatenation is represented by $\boldsymbol{F} = \left[\boldsymbol{F}^{\mathrm{i}}; \boldsymbol{F}^{\mathrm{c}}\right]$, while the batch-norm layer and FC layer are respectively denoted by $BN(\cdot)$ and $FC(\cdot)$. In addition, the sample number within a mini-batch and the dimensions of feature are respectively represented by b and d.

5.2.2 Calculation of Multi-purpose Loss Functions

Generally, the ultimate performance of the model is also greatly influenced by the selection of the suitable strategy of training. Therefore, a novel strategy of training, which is termed the calculation of multi-purpose loss functions, is introduced for enhancing the adversarial robustness of the proposed HFFN.

Despite the initial acquisition of the heterogeneous fused feature $\boldsymbol{V}$, the full potential of complementarity between different modalities remains underutilized, as the image features and caption features are fused simply through the operation of concatenation. Furthermore, the model exhibits limited adversarial robustness and generalization due to the inability of $\boldsymbol{V}$ to eliminate the gap between different modalities. Through the tight common representation learning of different modalities, the HC loss facilitates a more natural fusion of heterogeneous features and eliminates the information with the redundancy between them. Consequently, the integration of the HC loss into the HFFN framework is implemented to enhance the defense capabilities against adversarial attacks. As illustrated in Fig. 5.2, to address the issue of possible degradation in model performance, ℓ_2 normalization is applied to $\boldsymbol{V}$, resulting in the feature $\tilde{\boldsymbol{V}} \in \mathbb{R}^{b\times d}$. Through the limitation of the class center distances between the features, namely $\tilde{\boldsymbol{V}}^{\alpha}$ and $\tilde{\boldsymbol{V}}^{\beta} \in \mathbb{R}^{b\times \frac{d}{2}}$, which are divided from $\tilde{\boldsymbol{V}}$, the HC loss $\mathcal{L}_{\mathrm{HC}}$ effectively enhances the joint expressiveness and generalization of $\boldsymbol{V}$. Consequently, the calculation of $\mathcal{L}_{\mathrm{HC}}$ is expressed below:

$$\mathcal{L}_{\mathrm{HC}} = \sum_{k=1}^{N} \left[\max\left\{\mathrm{Dis}\left(\hat{\boldsymbol{v}}^{\alpha_k}, \hat{\boldsymbol{v}}^{\beta_k}\right) - \varphi, 0\right\}\right], \tag{5.4}$$

where the class centers of $\tilde{\boldsymbol{V}}^{\alpha}$ and $\tilde{\boldsymbol{V}}^{\beta}$ on class k are represented by $\hat{\boldsymbol{v}}^{\alpha_k}$ and $\hat{\boldsymbol{v}}^{\beta_k}$. Through the operation of average performed on the $\tilde{\boldsymbol{V}}^{\alpha_k} \in \mathbb{R}^{b_k\times \frac{d}{2}}$ $(b_k < b)$ in $\tilde{\boldsymbol{V}}^{\alpha}$, $\hat{\boldsymbol{v}}^{\alpha_k}$ can be acquired, and $\hat{\boldsymbol{v}}^{\beta_k}$ is acquired in the same procedure. The margin of $\mathcal{L}_{\mathrm{HC}}$ is denoted by φ, aiming to mitigate the overfitting problem. The function of cosine similarity is denoted by $\mathrm{Dis}(\cdot,\cdot)$. In addition, throughout the training phase, there are continual updates of the weights of the class centers $\hat{\boldsymbol{v}}^{\alpha_k}$ and $\hat{\boldsymbol{v}}^{\beta_k}$.

As the model tends to converge, the distances between class centers in $\mathcal{L}_{\mathrm{HC}}$ can narrow smoothly due to ongoing updates to them. Consequently, information with meaningfulness can be "distilled" by the model from the heterogeneous features, and the representation with common generalization can also be learned. Therefore, through the optimization of $\mathcal{L}_{\mathrm{HC}}$, enhancing the defense capabilities of $\boldsymbol{V}$ can be achieved.

The strong correlation between the model output distribution and the model robustness has been demonstrated by the previous works [63, 114]. Nevertheless, only the relationship between features of different modalities is focused by $\mathcal{L}_{\mathrm{HC}}$, rather than the optimization of the output distribution of heterogeneous features. Consequently, the GCE loss, denoted as $\mathcal{L}_{\mathrm{GCE}}$, is incor-

porated into the HFFN framework. Through the suppression of the model output probability for the classes of non-ground truth (non-GT), the distribution, which exhibits a significant disparity between the confidence levels of the classes of GT and non-GT, can be output. Consequently, the proposed model demonstrates robustness against adversarial attacks.

To perform the $\mathcal{L}_{\mathrm{GCE}}$ calculation, $\boldsymbol{V}$ is initially fed into the class block depicted in Fig. 5.2, resulting in $\boldsymbol{H} \in \mathbb{R}^{b\times N}$. Subsequently, $\boldsymbol{H}$ is processed through a softmax layer, resulting in the output probability matrix $\hat{\boldsymbol{H}} = \left[\hat{\boldsymbol{h}}_1; \cdots; \hat{\boldsymbol{h}}_b\right] \in \mathbb{R}^{b\times N}$, where the probability vector is represented by $\hat{\boldsymbol{h}}_r = \left[\hat{h}_{r1}, \cdots, \hat{h}_{rN}\right] \in \mathbb{R}^N$ $(1 \le r \le b)$, which is corresponding to the input data. Therefore, the calculation of $\mathcal{L}_{\mathrm{GCE}}$ is expressed below:

$$\mathcal{L}_{\mathrm{GCE}} = \frac{1}{b}\sum_{r=1}^{b}\left[\hat{h}_{rt}^{\theta} \times \frac{1}{\log N} \times \sum_{s=1, s\neq t}^{N}\left(\frac{\hat{h}_{rs}}{1-\hat{h}_{rt}} \times \log\left(\frac{\hat{h}_{rs}}{1-\hat{h}_{rt}}\right)\right)\right], \tag{5.5}$$

where the prediction probability of $\hat{\boldsymbol{h}}_r$ about the GT class is represented by $\hat{h}_{rt}$, while θ serves as a hyperparameter. Incorporating $\mathcal{L}_{\mathrm{GCE}}$ into the proposed HFFN enables the model to apply penalties to the classes with non-GT, fostering the generation of distribution with higher distinguishability. This operation bolsters the model robustness against adversarial attacks without incurring extra computational costs.

Through the combination of the aforementioned two loss functions, we can define the expression of the total loss $\mathcal{L}_{\mathrm{total}}$ within the HFFN framework:

$$\mathcal{L}_{\mathrm{total}} = \lambda \cdot \mathcal{L}_{\mathrm{HC}} + \mathcal{L}_{\mathrm{GCE}}, \tag{5.6}$$

where the weight of the HC loss function is denoted by λ.

Through the joint optimization of $\mathcal{L}_{HC}$ and $\mathcal{L}_{GCE}$, the proposed framework is capable of not only yielding the representation with efficiency and robustness between the different modality information but also enhancing the probability distribution disparity between the classes of GT and non-GT. Consequently, both the robustness and effectiveness of CNNs can be simultaneously improved in two aspects by the proposed HFFN.

5.3 Experiments

5.3.1 Experimental Settings

5.3.1.1 Datasets

In this chapter, extensive experiments were conducted by two widely used public datasets: CIFAR-10 and CIFAR-100 [79]. Color images, which are categorized into 10 distinct classes, constitute the CIFAR-10 dataset. In addition, the CIFAR-100 dataset encompasses 100 distinct classes, which can be additionally aggregated into 20 superclasses.

5.3.1.2 Attack Settings

In the CIFAR-10 experiments, a range of attack methods was employed, including FGSM [81], PGD-ℓ_∞ [82] with iteration 10 (PGD^{10}-ℓ_∞), PGD-ℓ_∞ with iteration 20 (PGD^{20}-ℓ_∞), MI-FGSM [83] with iteration 20 (MI-FGSM^{20}), Jitter [115] with iteration 40 (Jitter^{40}), PGD-ℓ_2 [82] with iteration 20 (PGD^{20}-ℓ_2), and PGD-ℓ_2 with iteration 40 (PGD^{40}-ℓ_2). For the FGSM, PGD-ℓ_∞, and MI-FGSM^{20} attacks, values of 2/255, 4/255, 8/255, 16/255, and 32/255 were selected as their parameter of attack strength ε. For the Jitter^{40} attack, values of 0.1, 0.2, 0.3, 0.4, and 0.5 were set as its parameter of attack strength ε. For the PGD-ℓ_2 attacks, values of 1.0, 2.0, 3.0, 4.0, and 5.0 were adopted as their parameter of attack strength ε. In the CIFAR-100 experiments, the attacks of FGSM, PGD^{10}-ℓ_∞, PGD^{20}-ℓ_∞, MI-FGSM^{20}, Jitter^{40}, PGD^{20}-ℓ_2, and PGD^{40}-ℓ_2 were adopted. Specifically, for the FGSM, PGD-ℓ_∞, and MI-FGSM^{20} attacks, the values of 16/255 and 32/255 were selected as the attack strength parameter ε, while the values of 0.1 and 0.2 were set as the ε parameter for the Jitter^{40} attack. For the PGD-ℓ_2 attacks, the values of 3.0, 4.0, and 5.0 were adopted as their ε parameter. Additionally, in the PGD-ℓ_2 attacks used in this chapter, the slight perturbations also were not added to the initial value, the same as in Chapter 3.

In this chapter, the setting of the black box was applied to the attack methods employed, as we held the view that such a setting better reflects real-world attack scenarios. Specifically, the adversarial examples for the CIFAR-10 and CIFAR-100 datasets were produced by conducting adversarial attacks on a trained substitute VGG-13 model. In addition, the adversarial examples were generated utilizing the toolbox of Torchattacks [101].

5.3.1.3 Comparison Methods

For performing the comparison analysis with HFFN, the following state-of-the-art adversarial defense approaches and robust framework were selected.

Guided complement entropy (GCE) [110]

While our proposed HFFN incorporates this loss, in this case, we utilized a classic CNN and combined this loss for training, and we viewed this trained CNN model as the comparison method.

Probabilistically compact loss with logit constraints (PCLC) [116]

PCLC, a novel defense approach characterized by its low computation complexity, enhances the separation between correct and incorrect classes. Owing to the logit constraint inherent in this approach, the subtle adversarial perturbations fail to negate the separation.

Hebbian/anti-Hebbian (HaH) [117, 118]

HaH, a robust framework inspired by neural mechanisms, focuses on searching a small number of neurons with large activations, instead of the multitude of neurons with small activations produced by the traditional CNN frameworks.

5.3.1.4 Ablation Methods

Additionally, an array of ablation approaches was also designed to comprehensively validate the effectiveness of the components within the proposed HFFN.

Specifically, the method, which is the combination of the ResNet-50 backbone and the cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$, was selected as Baseline. To investigate the impact of $\mathcal{L}_{\mathrm{HC}}$, the loss $\mathcal{L}_{\mathrm{HC}}$ in the HFFN framework was removed and only one loss $\mathcal{L}_{\mathrm{GCE}}$ was remained, and this method was denoted as Ablation-I. Furthermore, the loss $\mathcal{L}_{\mathrm{GCE}}$ in the method Ablation-I was replaced by $\mathcal{L}_{\mathrm{CE}}$, and this method was denoted as II. Lastly, the loss $\mathcal{L}_{\mathrm{GCE}}$ in HFFN was substituted with the loss $\mathcal{L}_{\mathrm{CE}}$, and this approach was denoted as Ablation-III, λ and 1 were also respectively set to the weight of $\mathcal{L}_{\mathrm{HC}}$ and $\mathcal{L}_{\mathrm{CE}}$ for comparing fairly with HFFN.

5.3.1.5 Parameter Settings

In the experiments, value 0.1 was selected as the margin parameter φ in $\mathcal{L}_{\text{HC}}$. For the CIFAR-10 dataset, value 10 was selected as the class number parameter N in $\mathcal{L}_{\text{HC}}$. In the case of CIFAR-100, the value 20, which is the superclass number of CIFAR-100, was selected for the class number parameter N in $\mathcal{L}_{\text{HC}}$ due to the batch size limitation. Regarding the loss $\mathcal{L}_{\text{GCE}}$, the class number parameter was assigned the value of 10 in the case of CIFAR-100 and 100 in the case of CIFAR-100. Additionally, the hyperparameter of θ within $\mathcal{L}$GCE was established at a value of 0.333, while the parameter of weight λ within the loss $\mathcal{L}_{\text{total}}$ was determined to be 0.5.

5.3.2 Results and Discussion

The adversarial attack experiment results about CIFAR-10 and CIFAR-100 of HFFN, alongside three comparison approaches, and four ablation methods are presented in this subsection.

Table 5.1 presents the clean results and robust accuracy results under the FGSM attack of the proposed framework and the comparison approaches for the CIFAR-10 dataset. Tables 5.2 – 5.4 show the more detailed robust accuracy results of HFFN and the comparison approaches for the CIFAR-10 dataset. These tables reveal that, despite the lower clean accuracy of HFFN compared to the comparison approaches, it consistently maintains a relatively high level of accuracy against adversarial examples. This performance stands in stark contrast to the significantly reduced robust accuracy in the three comparison methods. Particularly, in the case of ε with 32/255, the robust accuracy results under the PGD^{10}-ℓ_∞, PGD^{20}-ℓ_∞, and MI-FGSM^{20} attacks are all only around 1%. In contrast, our proposed HFFN method maintains exceptional robustness. Subsequently, Fig. 5.3 depicts the ablation study results of the proposed HFFN and the ablation approaches. This figure distinctly illustrates that, after the removal of different components from HFFN, the reduction in the robustness of various ablation methods to differing extents compared to HFFN. Consequently, it is demonstrable that each component of HFFN plays a role in enhancing the adversarial robustness of HFFN.

Additionally, the effectiveness of HFFN was also assessed using the more complex and challenging CIFAR-100 dataset. Specifically, Tables 5.5 – 5.7 show the results of comparison and ablation studies. From these tables, it is evident that HFFN surpasses both the comparison and

ablation approaches, indicating the suitability of HFFN even in the case of the dataset with increased complexity.

However, as we can see from the tables, the clean accuracy of our approach does not demonstrate superiority when compared to other methods. Concerning the cause of this occurrence, our argument is that $\mathcal{L}_{\text{HC}}$ is effective in diminishing the modality gap, the gap is not entirely eliminated.

5.4 Conclusions

An effective adversarial defense method has been proposed in this chapter. In the proposed approach, the robust caption features were integrated with the classic image features, thereby enhancing the robustness of the image classification system. Moreover, to the best of our knowledge, this kind of operation has not been focused by existing adversarial defense methods. Additionally, the hetero-center loss, which aims to diminish the modal disparity inherent in heterogeneous features, and the guided complement entropy loss, whose purpose is ensuring that the fused heterogeneous features focus more intently on the true class, were introduced into the proposed method. The adversarial robustness of HFFN and the effectiveness of the components in HFFN were revealed by a comprehensive series of black-box setting adversarial attack experiments among HFFN, comparison approaches, and ablation methods.

Table 5.1: Accuracy results of HFFN and three comparison methods in clean and FGSM attack scenarios on the CIFAR-10 dataset.

Attack Method	Parameter (ε)	GCE	PCLC	HaH	HFFN (Ours)
No Attack	-	93.43%	**93.94%**	88.65%	85.14%
FGSM	2/255	82.68%	**82.82%**	77.13%	81.52%
	4/255	67.95%	66.82%	62.90%	**76.25%**
	8/255	47.30%	46.08%	44.42%	**64.36%**
	16/255	31.72%	30.49%	29.46%	**46.50%**
	32/255	17.03%	15.50%	20.11%	**30.44%**

Table 5.2: Accuracy results among HFFN and comparison methods in the scenarios of PGD^{10}-ℓ_∞ and PGD^{20}-ℓ_∞ attacks on the CIFAR-10 dataset.

	PGD^{10}-ℓ_∞				
Method	$\varepsilon = 2/255$	$\varepsilon = 4/255$	$\varepsilon = 8/255$	$\varepsilon = 16/255$	$\varepsilon = 32/255$
GCE	**85.25%**	69.39%	37.21%	10.54%	1.47%
PCLC	85.17%	68.86%	36.34%	10.86%	1.79%
HaH	79.40%	65.28%	38.11%	11.81%	1.56%
HFFN (Ours)	82.46%	**78.94%**	**70.12%**	**53.85%**	**34.63%**
	PGD^{20}-ℓ_∞				
Method	$\varepsilon = 2/255$	$\varepsilon = 4/255$	$\varepsilon = 8/255$	$\varepsilon = 16/255$	$\varepsilon = 32/255$
GCE	80.51%	55.90%	20.74%	3.83%	0.29%
PCLC	80.74%	55.79%	20.29%	4.10%	0.54%
HaH	74.77%	54.88%	23.77%	4.83%	0.46%
HFFN (Ours)	**81.45%**	**76.62%**	**66.09%**	**46.61%**	**26.66%**

Table 5.3: Accuracy results among HFFN and comparison methods in the scenarios of MI-FGSM^{20} and Jitter^{40} attacks on the dataset CIFAR-10.

	MI-FGSM^{20}				
Method	$\varepsilon = 2/255$	$\varepsilon = 4/255$	$\varepsilon = 8/255$	$\varepsilon = 16/255$	$\varepsilon = 32/255$
GCE	80.12%	54.65%	19.38%	3.60%	0.97%
PCLC	80.21%	54.38%	18.74%	3.88%	1.07%
HaH	74.24%	53.29%	21.87%	4.24%	1.24%
HFFN (Ours)	**81.18%**	**76.40%**	**65.33%**	**44.56%**	**23.07%**
	Jitter^{40}				
Method	$\varepsilon = 0.1$	$\varepsilon = 0.2$	$\varepsilon = 0.3$	$\varepsilon = 0.4$	$\varepsilon = 0.5$
GCE	42.84%	14.81%	10.54%	10.08%	10.00%
PCLC	36.65%	12.81%	11.46%	11.06%	10.83%
HaH	51.42%	27.19%	16.20%	10.90%	8.31%
HFFN (Ours)	**60.76%**	**45.16%**	**29.39%**	**18.74%**	**14.69%**

Table 5.4: Accuracy results among HFFN and comparison methods in the scenarios of PGD^{20}-ℓ_2 and PGD^{40}-ℓ_2 attacks on the dataset CIFAR-10.

	PGD^{20}-ℓ_2				
Method	$\varepsilon = 1.0$	$\varepsilon = 2.0$	$\varepsilon = 3.0$	$\varepsilon = 4.0$	$\varepsilon = 5.0$
GCE	26.33%	14.04%	12.33%	11.63%	11.31%
PCLC	27.02%	14.20%	12.37%	11.67%	11.35%
HaH	30.05%	15.20%	12.71%	11.76%	11.46%
HFFN (Ours)	**68.60%**	**50.07%**	**39.24%**	**33.30%**	**29.69%**
	PGD^{40}-ℓ_2				
Method	$\varepsilon = 1.0$	$\varepsilon = 2.0$	$\varepsilon = 3.0$	$\varepsilon = 4.0$	$\varepsilon = 5.0$
GCE	24.30%	12.16%	10.81%	10.19%	9.88%
PCLC	24.90%	12.33%	10.78%	10.20%	9.90%
HaH	28.77%	13.32%	11.17%	10.29%	9.97%
HFFN (Ours)	**68.94%**	**49.12%**	**35.46%**	**28.07%**	**23.85%**

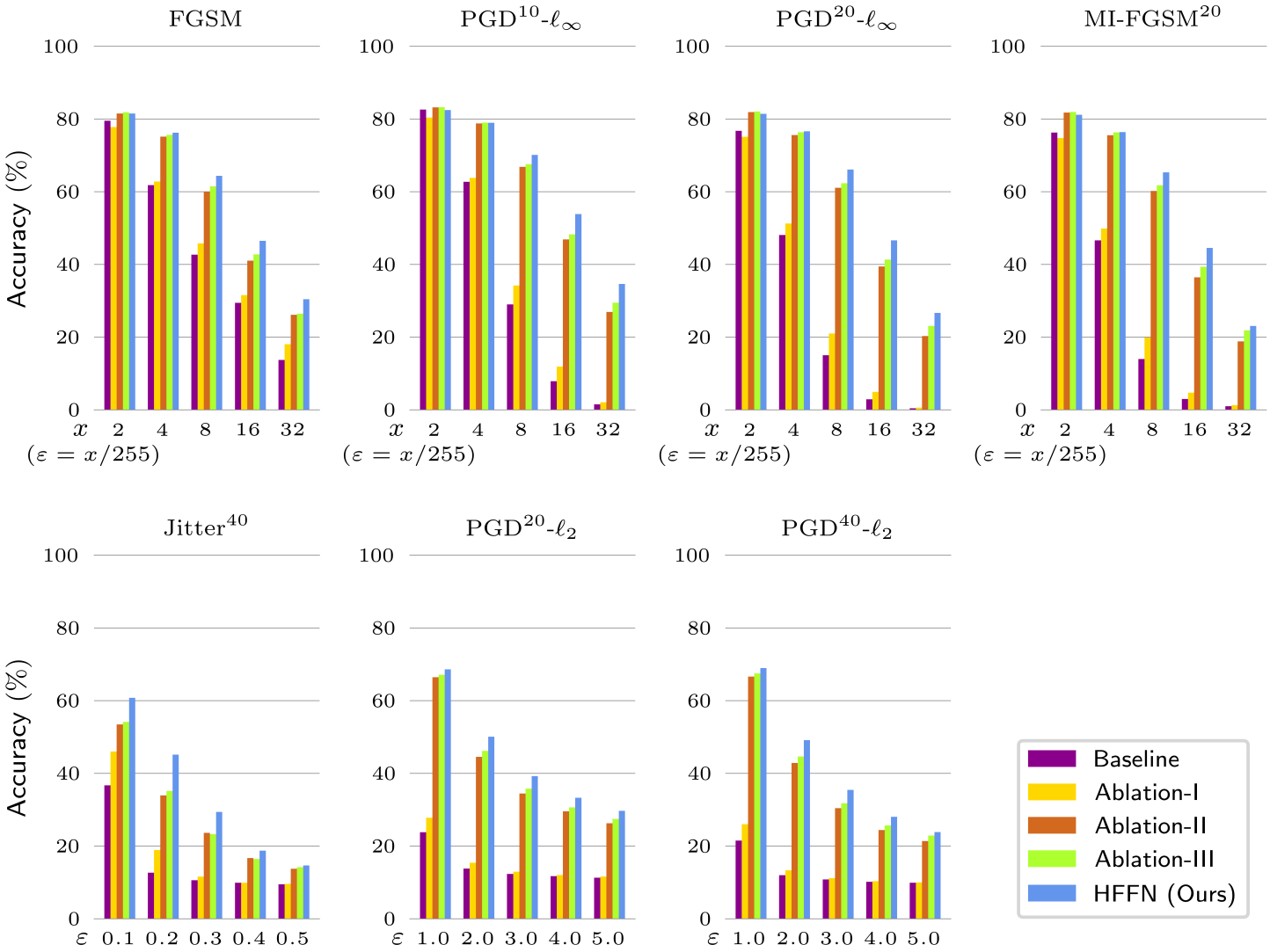


Figure 5.3: Accuracy results among HFFN and four ablation methods on the CIFAR-10 dataset.

Table 5.5: Accuracy results of HFFN, comparison methods, and ablation approaches in clean and FGSM attack scenarios on the CIFAR-100 dataset.

Method	**Clean**	**FGSM** $\varepsilon = 16/255$	**FGSM** $\varepsilon = 32/255$
GCE	**75.15%**	19.93%	6.32%
PCLC	71.11%	18.57%	4.79%
HaH	50.20%	21.38%	9.19%
Baseline	69.41%	17.21%	4.92%
Ablation-I	64.66%	22.05%	7.83%
Ablation-II	64.84%	23.96%	10.91%
Ablation-III	66.24%	21.06%	10.09%
HFFN (Ours)	57.28%	**30.66%**	**14.09%**

Table 5.6: Accuracy results of HFFN, comparison methods, and ablation approaches in the scenarios of the PGD^{10}-ℓ_∞, PGD^{20}-ℓ_∞, MI-FGSM^{20}, and Jitter^{40} attacks on the CIFAR-100 dataset.

	PGD^{10}-ℓ_∞		**PGD^{20}-ℓ_∞**		**MI-FGSM^{20}**		**Jitter^{40}**	
Method	$\varepsilon = 16/255$	$\varepsilon = 32/255$	$\varepsilon = 16/255$	$\varepsilon = 32/255$	$\varepsilon = 16/255$	$\varepsilon = 32/255$	$\varepsilon = 0.1$	$\varepsilon = 0.2$
GCE	16.01%	5.18%	9.75%	3.20%	10.41%	4.91%	20.41%	5.74%
PCLC	17.96%	6.80%	12.24%	4.55%	12.31%	6.14%	21.69%	7.55%
HaH	29.04%	14.36%	24.20%	10.50%	23.97%	12.11%	31.02%	16.93%
Baseline	18.70%	6.92%	12.85%	4.57%	12.79%	6.15%	20.92%	7.07%
Ablation-I	29.04%	12.19%	22.73%	8.47%	22.29%	10.02%	30.73%	11.50%
Ablation-II	32.55%	16.88%	27.63%	13.65%	27.01%	13.88%	27.22%	13.94%
Ablation-III	29.31%	14.12%	24.74%	11.60%	24.71%	12.12%	23.79%	10.51%
HFFN (Ours)	**39.24%**	**26.01%**	**34.53%**	**20.64%**	**33.07%**	**18.53%**	**39.26%**	**26.46%**

Table 5.7: Accuracy results of HFFN, comparison methods, and ablation approaches in the scenarios of the PGD^{20}-ℓ_2 and PGD^{40}-ℓ_2 attacks on the CIFAR-100 dataset.

	PGD^{20}-ℓ_2			**PGD^{40}-ℓ_2**		
Method	$\varepsilon = 3.0$	$\varepsilon = 4.0$	$\varepsilon = 5.0$	$\varepsilon = 3.0$	$\varepsilon = 4.0$	$\varepsilon = 5.0$
GCE	14.85%	14.13%	13.74%	13.56%	12.84%	12.51%
PCLC	16.69%	15.31%	14.64%	14.97%	13.81%	13.07%
HaH	23.76%	20.96%	19.38%	21.88%	18.48%	16.28%
Baseline	16.88%	15.48%	14.73%	15.32%	13.96%	13.28%
Ablation-I	22.31%	19.82%	18.35%	20.43%	17.46%	15.47%
Ablation-II	26.93%	23.23%	21.30%	23.74%	19.63%	17.49%
Ablation-III	25.50%	22.03%	19.99%	22.34%	18.76%	17.03%
HFFN (Ours)	**29.77%**	**26.44%**	**24.03%**	**27.59%**	**22.56%**	**19.97%**

Chapter 6

Conclusions

As conclusions of this thesis, the summary of this thesis and the future research directions are presented in this chapter.

6.1 Summary of This Thesis

In this thesis, adversarial defense research for deep learning-based architectures, as exemplified by the image classification and image captioning systems, was focused upon. To achieve this goal, a novel method for the introduction of robust features has been proposed, encompassing the following three directions: integrating the robust regularization-based enhanced multi-stage features for the image classification system, incorporating the robust caption features generated from the stable RNN structure for the retrieval-based image captioning system, and introducing the robust caption features generated by the pre-trained models for the image classification system.

In Chapter 3, to address the issue in existing regularization-based adversarial defense approaches where there is no further discussion on which features are suitable for regularization to enhance the adversarial robustness, we introduced the robust regularization-based enhanced multi-stage features into the image classification system and proposed a novel framework comprising the following three core components: multi-stage feature enhancement, multi-stage feature fusion, and regularization.

In Chapter 4, to address the issue of vulnerability to adversarial attacks in the existing image captioning system, which stems from the lack of robust feature incorporation, we introduced a

kind of robust RNN features into the retrieval-based image captioning system to achieve the goal of the adversarial defense. Specifically, the introduced stable RNN architecture was designed based on the theory of dynamical systems.

In Chapter 5, to address the issue that most existing defense methods for image classification rely solely on the information from input images and do not incorporate the robust multimodal features, thereby limiting further robustness enhancements of the image classification system, we drew inspiration from the human ability to adeptly utilize multimodal information for the object perception. Consequently, we introduced the robust caption features from the pre-trained models into the image classification system and fused them with the image features. In addition, to maximize the robust performance of the introduced features, we also incorporated two novel loss functions: the hetero-center loss, which was introduced to minimize the distances between different modalities, and the guided complement entropy loss, which was introduced to further enhance the robustness of the fused heterogeneous features.

In summary, the contribution of this thesis is the proposition of the adversarial defense methods for deep learning models represented by image classification and image captioning systems, from a unified perspective. Specifically, we have introduced robust features into the image classification and image captioning systems. Additionally, extensive experiments on various datasets have fully demonstrated the robustness and effectiveness of the proposed method.

6.2 Future Research Directions

In this section, the future research directions are shown. In this thesis, we have proposed an adversarial defense approach by incorporating robust features, which has enhanced the adversarial robustness of the deep learning models, as exemplified by the image captioning and image classification systems. We contend that there exists potential for refinement in the proposed method in the following aspects:

- In the pursuit of constructing the adversarially robust image captioning system, this thesis has specifically concentrated on the incorporation of the robust features into the retrieval-based image captioning framework. Moving forward, we would like to contemplate extending our methodology to the popular generation-based image captioning system to bolster its robustness.

- Moreover, from a macroscopic perspective, this thesis concentrated on the adversarial robustness of two representative systems based on deep learning and the proposed adversarial defense method that introduces robust features for these two systems. In the future, we would like to extend our approach to other deep learning-based systems, such as image segmentation and object detection.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Miki Haseyama. During my doctoral course at the Laboratory of Media Dynamics, Graduate School of Information Science and Technology, Hokkaido University, she provided me with invaluable advice and encouragement for my research. Without her precious guidance, this thesis would have not been possible.

Secondly, I would like to express my thankfulness to Specially Appointed Prof. Kenji Araki, Specially Appointed Prof. Yuji Sakamoto, Prof. Yoshinori Dobashi, and Prof. Takahiro Ogawa for their constructive comments and suggestions on my research conducted at the Graduate School of Information Science and Technology, Hokkaido University.

Thirdly, I would also like to show my sincere appreciation to Prof. Takahiro Ogawa for his patient guidance in my research and for providing invaluable insights about my research during the discussions.

Moreover, I would like to offer my heartfelt thanks to Specially Appointed Associate Prof. Keisuke Maeda for the support and inspiration provided throughout my research life. In addition, I would also like to voice my sincere thanks to Specially Appointed Assistant Prof. Ren Togo, Assistant Prof. Naoki Saito, and Specially Appointed Assistant Prof. Guang Li for their wonderful insights generously offered throughout my academic life.

Finally, I would like to express my sincere thanks to all the members of the Laboratory of Media Dynamics, Graduate School of Information Science and Technology, Hokkaido University. The moments of studying together are memories I will cherish forever. These memories will forever be engraved in my heart and serve as a source of motivation as I move forward in my life.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[6] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1492–1500.

[7] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[8] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in

Proceedings of the International Conference on Machine Learning, 2015, pp. 2048–2057.

[9] J. Gao, S. Wang, S. Wang, S. Ma, and W. Gao, "Self-critical n-step training for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6300–6308.

[10] Y. Qin, J. Du, Y. Zhang, and H. Lu, "Look back and predict forward in image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8367–8375.

[11] C. Chen, S. Mu, W. Xiao, Z. Ye, L. Wu, and Q. Ju, "Improving image captioning with conditional generative adversarial nets," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 8142–8150.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2015.

[13] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[15] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154–6162.

[16] K. Rao, H. Sak, and R. Prabhavalkar, "Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2017, pp. 193–199.

[17] E. Battenberg, J. Chen, R. Child, A. Coates, Y. G. Y. Li, H. Liu, S. Satheesh, A. Sriram, and Z. Zhu, "Exploring neural transducers for end-to-end speech recognition," in *Proceedings*

of the IEEE Automatic Speech Recognition and Understanding Workshop, 2017, pp. 206–213.

[18] J. Chen, M. I. Jordan, and M. J. Wainwright, "HopSkipJumpAttack: A query-efficient decision-based attack," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2020, pp. 1277–1294.

[19] A. Modas, S.-M. Moosavi-Dezfooli, and P. Frossard, "SparseFool: A few pixels make a big difference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9087–9096.

[20] L. Gao, Q. Zhang, J. Song, X. Liu, and H. T. Shen, "Patch-wise attack for fooling deep neural network," in *Proceedings of the European Conference on Computer Vision*. Springer, 2020, pp. 307–322.

[21] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *Proceedings of the International Conference on Machine Learning*, 2018, pp. 2137–2146.

[22] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.-J. Hsieh, "EAD: Elastic-net attacks to deep neural networks via adversarial examples," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[23] M. Cheng, T. Le, P.-Y. Chen, J. Yi, H. Zhang, and C.-J. Hsieh, "Query-efficient hard-label black-box attack: An optimization-based approach," *arXiv preprint arXiv:1807.04457*, 2018.

[24] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2574–2582.

[25] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1924–1933.

[26] C.-C. Tu, P. Ting, P.-Y. Chen, S. Liu, H. Zhang, J. Yi, C.-J. Hsieh, and S.-M. Cheng, "AutoZOOM: Autoencoder-based zeroth order optimization method for attacking black-box neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 742–749.

[27] L. Schwinn, R. Raab, A. Nguyen, D. Zanca, and B. Eskofier, "Exploring misclassifications of robust neural networks to enhance adversarial attacks," *Applied Intelligence*, pp. 1–17, 2023.

[28] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," *arXiv preprint arXiv:1908.06281*, 2019.

[29] J. Pomponi, S. Scardapane, and A. Uncini, "Pixle: A fast and effective black-box attack based on rearranging pixels," in *Proceedings of the International Joint Conference on Neural Networks*, 2022, pp. 1–7.

[30] F. Croce and M. Hein, "Minimally distorted adversarial examples with a fast adaptive boundary attack," in *Proceedings of the International Conference on Machine Learning*, 2020, pp. 2196–2205.

[31] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 484–501.

[32] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proceedings of the IEEE European Symposium on Security and Privacy*, 2016, pp. 372–387.

[33] H. Li, X. Xu, X. Zhang, S. Yang, and B. Li, "QEBA: Query-efficient boundary-based blackbox attack," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1221–1230.

[34] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4312–4321.

[35] P. N. Williams and K. Li, "Black-box sparse adversarial attack via multi-objective optimisation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12291–12301.

[36] J. Zhang, B. Li, J. Xu, S. Wu, S. Ding, L. Zhang, and C. Wu, "Towards efficient data free black-box adversarial attack," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15115–15125.

[37] A. Ghiasi, A. Shafahi, and T. Goldstein, "Breaking certified defenses: Semantic adversarial examples with spoofed robustness certificates," in *Proceedings of the International Conference on Learning Representations*, 2020, pp. 1–16.

[38] A. S. Shamsabadi, R. Sanchez-Matilla, and A. Cavallaro, "ColorFool: Semantic adversarial colorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1151–1160.

[39] Y. Li, L. Li, L. Wang, T. Zhang, and B. Gong, "NAttack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks," in *Proceedings of the International Conference on Machine Learning*, 2019, pp. 3866–3876.

[40] Y. Shi, Y. Han, and Q. Tian, "Polishing decision-based adversarial noise with a customized sampling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1030–1038.

[41] H. Chen, H. Zhang, P.-Y. Chen, J. Yi, and C.-J. Hsieh, "Attacking visual language grounding with adversarial examples: A case study on neural image captioning," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 2587–2597.

[42] Y. Xu, B. Wu, F. Shen, Y. Fan, Y. Zhang, H. T. Shen, and W. Liu, "Exact adversarial attack to image captioning via structured output learning with latent variables," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4135–4144.

[43] C. Xie, M. Tan, B. Gong, A. Yuille, and Q. V. Le, "Smooth adversarial training," *arXiv preprint arXiv:2006.14536*, 2020.

[44] J. Zhang, J. Zhu, G. Niu, B. Han, M. Sugiyama, and M. Kankanhalli, "Geometry-aware instance-reweighted adversarial training," in *Proceedings of the International Conference on Learning Representations*, 2021, pp. 1–29.

[45] T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu, "Bag of tricks for adversarial training," in *Proceedings of the International Conference on Learning Representations*, 2021, pp. 1–21.

[46] B. Vivek and R. V. Babu, "Single-step adversarial training with dropout scheduling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 947–956.

[47] X. Jia, Y. Zhang, B. Wu, K. Ma, J. Wang, and X. Cao, "LAS-AT: Adversarial training with learnable attack strategy," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13398–13408.

[48] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[49] H. Zheng, Z. Zhang, J. Gu, H. Lee, and A. Prakash, "Efficient adversarial training with transferable adversarial examples," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1181–1190.

[50] G. W. Ding, Y. Sharma, K. Y. C. Lui, and R. Huang, "MMA training: Direct input space margin maximization through adversarial training," in *Proceedings of the International Conference on Learning Representations*, 2020, pp. 1–28.

[51] C. Song, K. He, J. Lin, L. Wang, and J. E. Hopcroft, "Robust local features for improving the generalization of adversarial training," pp. 1–12, 2020.

[52] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," pp. 1–12, 2017.

[53] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner, "Detecting adversarial samples from artifacts," *arXiv preprint arXiv:1703.00410*, 2017.

[54] D. Meng and H. Chen, "Magnet: A two-pronged defense against adversarial examples," in *Proceedings of the ACM Conference on Computer and Communications Security*, 2017, pp. 135–147.

[55] T. Pang, C. Du, Y. Dong, and J. Zhu, "Towards robust detection of adversarial examples," *Advances in neural information processing systems*, vol. 31, pp. 1–11, 2018.

[56] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in Neural Information Processing Systems*, vol. 31, pp. 1–11, 2018.

[57] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey, "Characterizing adversarial subspaces using local intrinsic dimensionality," in *Proceedings of the International Conference on Learning Representations*, 2018, pp. 1–15.

[58] P. Yang, J. Chen, C.-J. Hsieh, J.-L. Wang, and M. Jordan, "Ml-loo: Detecting adversarial examples with feature attribution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 6639–6647.

[59] B. Liang, H. Li, M. Su, X. Li, W. Shi, and X. Wang, "Detecting adversarial image examples in deep neural networks with adaptive noise reduction," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 1, pp. 72–85, 2021.

[60] S. Guo, X. Li, P. Zhu, and Z. Mu, "Ads-detector: An attention-based dual stream adversarial example detection method," *Knowledge-Based Systems*, vol. 265, pp. 1–14, 2023.

[61] X. Jia, X. Wei, X. Cao, and H. Foroosh, "ComDefend: An efficient image compression model to defend adversarial examples," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6084–6092.

[62] Z. Liu, Q. Liu, T. Liu, N. Xu, X. Lin, Y. Wang, and W. Wen, "Feature distillation: DNN-oriented JPEG compression against adversarial examples," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 860–868.

[63] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1778–1787.

[64] B. Sun, N.-h. Tsai, F. Liu, R. Yu, and H. Su, "Adversarial defense by stratified convolutional sparse coding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11447–11456.

[65] Y. Bakhti, S. A. Fezza, W. Hamidouche, and O. Déforges, "DDSA: A defense against adversarial attacks using deep denoising sparse autoencoder," *IEEE Access*, vol. 7, pp. 160397–160407, 2019.

[66] C. Xie, Y. Wu, L. v. d. Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 501–509.

[67] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "PixelDefend: Leveraging generative models to understand and defend against adversarial examples," in *Proceedings of the International Conference on Learning Representations*, 2018, pp. 1–20.

[68] C. Kou, H. K. Lee, T. K. Ng, and E.-C. Chang, "Enhancing transformation-based defenses against adversarial attacks with a distribution classifier," in *Proceedings of the International Conference on Learning Representations*, 2020, pp. 1–19.

[69] Y. Yang, G. Zhang, D. Katabi, and Z. Xu, "ME-Net: Towards effective adversarial robustness with matrix estimation," in *Proceedings of the International Conference on Machine Learning*, 2019.

[70] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten, "Countering adversarial images using input transformations," in *Proceedings of the International Conference on Learning Representations*, 2018, pp. 1–12.

[71] S.-M. Moosavi-Dezfooli, A. Fawzi, J. Uesato, and P. Frossard, "Robustness via curvature regularization, and vice versa," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9078–9086.

[72] A. Mustafa, S. Khan, M. Hayat, R. Goecke, J. Shen, and L. Shao, "Adversarial defense by restricting the hidden space of deep neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3385–3394.

[73] C. Agarwal, A. Nguyen, and D. Schonfeld, "Improving robustness to adversarial examples by encouraging discriminative features," in *Proceedings of the IEEE International Conference on Image Processing*, 2019, pp. 3801–3505.

[74] J. Xu, Y. Li, Y. Jiang, and S.-T. Xia, "Adversarial defense via local flatness regularization," in *Proceedings of the IEEE International Conference on Image Processing*, 2020, pp. 2196–2200.

[75] Q. Li, Q. Hu, C. Lin, D. Wu, and C. Shen, "Revisiting gradient regularization: Inject robust saliency-aware weight bias for adversarial defense," *IEEE Transactions on Information Forensics and Security*, pp. 1–14, 2023.

[76] M. Picot, F. Messina, M. Boudiaf, F. Labeau, I. B. Ayed, and P. Piantanida, "Adversarial robustness via Fisher-Rao regularization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 2698–2710, 2022.

[77] J. Tack, S. Yu, J. Jeong, M. Kim, S. J. Hwang, and J. Shin, "Consistency regularization for adversarial robustness," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 8414–8422.

[78] C. Summers and M. J. Dinneen, "Improved adversarial robustness via logit regularization methods," *arXiv preprint arXiv:1906.03749*, 2019.

[79] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[80] J. Zhang, K. Maeda, T. Ogawa, and M. Haseyama, "A note on realizing adversarial defense based on regularization of multi-stage squeeze-and-excitation features," in *ITE Technical Report*, 2022, vol. 46, pp. 87–90.

[81] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of the International Conference on Learning Representations*, 2015,

pp. 1–11.

[82] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of the International Conference on Learning Representations*, 2018.

[83] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9185–9193.

[84] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2017.

[85] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2730–2739.

[86] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 630–645.

[87] R. S. Zimmermann, "Comment on 'Adv-BNN: Improved adversarial defense through robust bayesian neural network'," *arXiv preprint arXiv:1907.00895*, 2019.

[88] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proceedings of the International Conference on Machine Learning*, 2018, pp. 284–293.

[89] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2017, pp. 39–57.

[90] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[91] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1765–1773.

[92] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 15–26.

[93] R. Shekhar, S. Pezzelle, Y. Klimovich, A. Herbelot, M. Nabi, E. Sangineto, and R. Bernardi, "FOIL it! Find one mismatch between image and language caption," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 255–265.

[94] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 740–755.

[95] J. Ji, X. Sun, Y. Zhou, R. Ji, F. Chen, J. Liu, and Q. Tian, "Attacking image captioning towards accuracy-preserving target words removal," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 4226–4234.

[96] S. Zhang, Z. Wang, X. Xu, X. Guan, and Y. Yang, "Fooled by imagination: Adversarial attack to image captioning via perturbation in complex domain," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2020, pp. 1–6.

[97] H. Wu, Y. Liu, H. Cai, and S. He, "Learning transferable perturbations for image captioning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 18, no. 2, pp. 1–18, 2022.

[98] S. Chen, Z. Song, M. Haque, C. Liu, and W. Yang, "NICGSlowDown: Evaluating the efficiency robustness of neural image caption generation models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15365–15374.

[99] S. Zhang, Y. Cheng, W. Zhu, X. Ji, and W. Xu, "CAPatch: Physical adversarial patch against image captioning systems," in *Proceedings of the USENIX Security Symposium*, 2023, pp. 679–696.

[100] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[101] H. Kim, "Torchattacks: A pytorch repository for adversarial attacks," *arXiv preprint arXiv:2010.01950,*, 2020.

[102] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu, "Improving adversarial robustness requires revisiting misclassified examples," in *Proceedings of the International Conference on Learning Representations*, 2020, pp. 1–14.

[103] M. Guo, Y. Yang, R. Xu, Z. Liu, and D. Lin, "When NAS meets robustness: In search of robust architectures against adversarial attacks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 631–640.

[104] S. Addepalli, V. BS, A. Baburaj, G. Sriramanan, and R. V. Babu, "Towards achieving adversarial robustness by enforcing feature consistency across bit planes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1020–1029.

[105] N. B. Erichson, O. Azencot, A. Queiruga, L. Hodgkinson, and M. W. Mahoney, "Lipschitz recurrent neural networks," in *Proceedings of the International Conference on Learning Representations*, 2021, pp. 1–21.

[106] H. Shi, J. Mao, T. Xiao, Y. Jiang, and J. Sun, "Learning visually-grounded semantics from contrastive adversarial sample," in *Proceedings of the International Conference on Computational Linguistics*, 2018, pp. 3715–3727.

[107] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," in *Proceedings of the British Machine Vision Conference*, 2018, pp. 1–13.

[108] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," *arXiv preprint arXiv:2111.09734*, 2021.

[109] Y. Zhu, Z. Yang, L. Wang, S. Zhao, X. Hu, and D. Tao, "Hetero-center loss for cross-modality person re-identification," *Neurocomputing*, vol. 386, pp. 97–109, 2020.

[110] H.-Y. Chen, J.-H. Liang, S.-C. Chang, J.-Y. Pan, Y.-T. Chen, W. Wei, and D.-C. Juan, "Improving adversarial robustness via guided complement entropy," in *Proceedings of the*

IEEE International Conference on Computer Vision, 2019, pp. 4881–4889.

[111] J. Zhang, K. Maeda, T. Ogawa, and M. Haseyama, "Regularization meets enhanced multi-stage fusion features: Making CNN more robust against white-box adversarial attacks," *Sensors*, vol. 22, no. 14, pp. 1–20, 2022.

[112] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 8748–8763.

[113] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 2556–2565.

[114] T. Pang, C. Du, and J. Zhu, "Max-mahalanobis linear discriminant analysis networks," in *Proceedings of the International Conference on Machine Learning*, 2018, pp. 4016–4025.

[115] L. Schwinn, R. Raab, A. Nguyen, D. Zanca, and B. Eskofier, "Exploring misclassifications of robust neural networks to enhance adversarial attacks," *arXiv preprint arXiv:2105.10304*, 2021.

[116] X. Li, X. Li, D. Pan, and D. Zhu, "Improving adversarial robustness via probabilistically compact loss with logit constraints," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 8482–8490.

[117] M. Cekic, C. Bakiskan, and U. Madhow, "Towards robust, interpretable neural networks via Hebbian/anti-Hebbian learning: A software framework for training with feature-based costs," *Software Impacts*, vol. 13, pp. 1–3, 2022.

[118] M. Cekic, C. Bakiskan, and U. Madhow, "Neuro-inspired deep neural networks with sparse, strong activations," in *Proceedings of the IEEE International Conference on Image Processing*, 2022, pp. 3843–3847.

Achievements of the Author

(A) **Journal**

[**A-1**] J. Zhang, K. Maeda, T. Ogawa, and M. Haseyama, "Regularization meets enhanced multi-stage fusion features: Making CNN more robust against white-box adversarial attacks," *Sensors* (2022 IF=3.9), vol. 22, no. 14: 5431, 2022.

(B) **International Conference**

[**B-1**] J. Zhang, K. Maeda, T. Ogawa, and M. Haseyama, "Defense against image captioning attacks via a robust and stable recurrent neural network," in *Proceedings of the IEEE Global Conference on Consumer Electronics (GCCE)*, pp. 785–786, 2021.

[**B-2**] J. Zhang, K. Maeda, T. Ogawa, and M. Haseyama, "Defense against black-box adversarial attacks via heterogeneous fusion features," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

(C) **Domestic Conference**

[**C-1**] J. Zhang, K. Maeda, T. Ogawa, and M. Haseyama, "A note on realizing adversarial defense based on regularization of multi-stage squeeze-and-excitation features," ITE Technical Report, vol. 46, no. 6, pp. 87–90, 2022.

[**C-2**] J. Zhang, R. Togo, K. Maeda, T. Ogawa, and M. Haseyama, "A note on specific object removal in urban scene using video inpainting approach," ITE Technical Report, vol. 47, no. 6, pp. 277–280, 2023.

(D) **Scholarship**

[**D-1**] Hokkaido University Ambitious Doctoral Fellowship (Information Science and AI), 2021.