Enhancing Intrinsic Adversarial Robustness via Feature Pyramid Decoder[‡]

Guanlin Li[1,*] Shuya Ding[2,*] Jun Luo[2] Chang Liu[2]

[1]Shandong Provincial Key Laboratory of Computer Networks, Shandong Computer Science Center (National Supercomputer Center in Jinan)

[2]School of Computer Science and Engineering, Nanyang Technological University

leegl@sdas.org {di0002ya,junluo,chang015}@ntu.edu.sg

Abstract

Whereas adversarial training is employed as the main defence strategy against specific adversarial samples, it has limited generalization capability and incurs excessive time complexity. In this paper, we propose an attack-agnostic defence framework to enhance the intrinsic robustness of neural networks, without jeopardizing the ability of generalizing clean samples. Our Feature Pyramid Decoder (FPD) framework applies to all block-based convolutional neural networks (CNNs). It implants denoising and image restoration modules into a targeted CNN, and it also constraints the Lipschitz constant of the classification layer. Moreover, we propose a two-phase strategy to train the FPD-enhanced CNN, utilizing ϵ-neighbourhood noisy images with multi-task and self-supervised learning. Evaluated against a variety of white-box and black-box attacks, we demonstrate that FPD-enhanced CNNs gain sufficient robustness against general adversarial samples on MNIST, SVHN and CALTECH. In addition, if we further conduct adversarial training, the FPD-enhanced CNNs perform better than their non-enhanced versions.

1. Introduction

The ever-growing ability of deep learning has found numerous applications mainly in image classification, object detection, and natural language processing [12, 17, 28]. While deep learning has brought great convenience to our lives, its weakness is also catching researchers' attention. Recently, researchers have started to pay more attention to investigating the weakness of neural networks, especially in its application to image classification. Since the seminal work by [26, 21], many follow-up works have demonstrated a great variety of methods in generating *adversarial samples*: though easily distinguishable by human eyes, they are often misclassified by neural networks. More specifically, most convolutional layers are very sensitive to perturbations brought by adversarial samples (*e.g.*, [6, 31]), resulting in misclassifications. These so-called *adversarial attacks* may adopt either white-box or black-box approaches, depending on the knowledge of the target network, and they mostly use gradient-based methods [10, 18, 27] or score-based methods [4] to generate adversarial samples.

To thwart these attacks, many defence methods have been proposed. Most of them use *adversarial training* to increase the network robustness, *e.g.*, [1, 14]. However, as training often targets a specific attack, the resulting defense method can hardly be generalized, as hinted in [27]. In order to defend against various attacks, a large amount and variety of adversarial samples are required to retrain the classifier, leading to a high time-complexity. In the meantime, little attention has been given to the direct design of robust frameworks in an *attack-agnostic* manner, except a few touches on denoising [25, 29] and obfuscating gradients [11, 25] that aim to directly enhance a target network in order to cope with any potential attacks.

To enhance the intrinsic robustness of neural networks, we propose an attack-agnostic defence framework, applicable to enhance all types of block-based CNNs. We aim to thwart both white-box and black-box attacks without crafting any specific adversarial attacks. Our Feature Pyramid Decoder (FPD) framework implants a target CNN with both denoising and image restoration modules to filter an input image at multiple levels; it also deploys a Lipschitz Constant Constraint at the classification layer to limit the output variation in the face of attack perturbation. In order to train an FPD-enhanced CNN, we propose a two-phase strategy; it utilizes ϵ-neighbourhood noisy images to drive multi-task and self-supervised learning.

As shown in Figure 1, FPD employs a front denoising module, an image restoration module, a middle denoising layer, and a back denoising module. Both front and back denoising modules consist of the original CNN blocks interleaved with inner denoising layers, and the inner denoising

*These authors contributed equally to this work.

†The author is also affiliated with Shandong Computer Science Center, Shandong Academy of Sciences, School of Cyber Security, Qilu University of Technology, China

‡https://github.com/GuanlinLee/FPD-for-Adversarial-Robustness

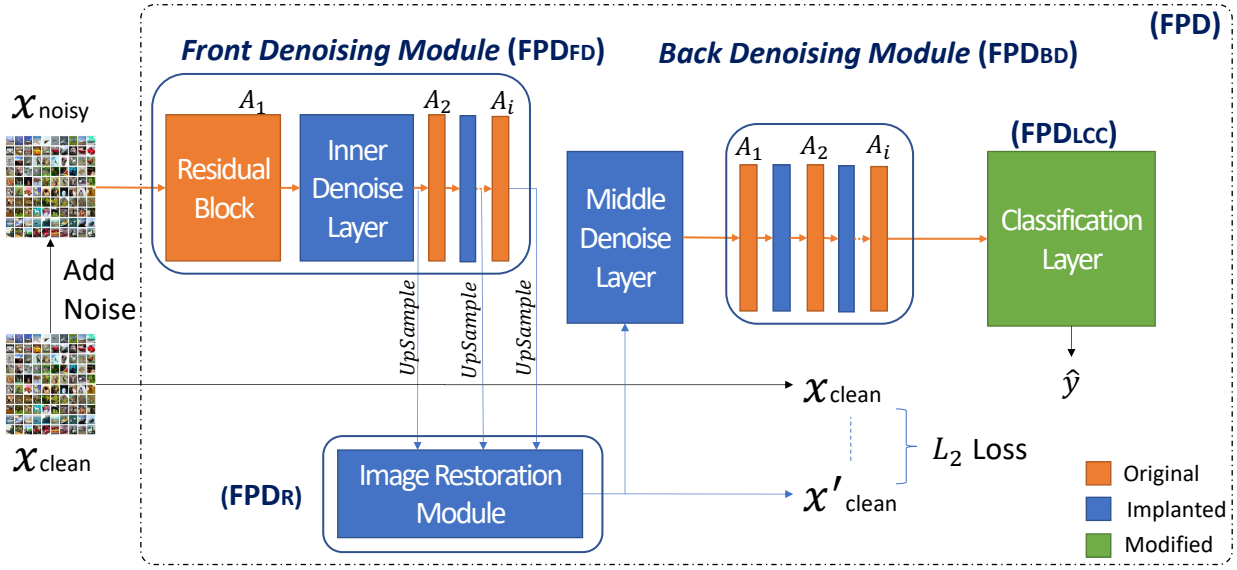


Figure 1: The structure of the block-based CNN, enhanced with the proposed framework named FPD: it consists of the Lipschitz constant constrained classification layer $\mathrm{FPD_{LCC}}$; the front denoising module $\mathrm{FPD_{FD}}$, the image restoration module $\mathrm{FPD_{R}}$, a middle denoising layer and the back denoising module $\mathrm{FPD_{BD}}$. ϵ-neighbourhood noisy samples x_{nosiy} and original samples x_{clean} are used to train the FPD. Orange, blue and green blocks represent the original components of the CNN, the proposed components that implanted to the CNN, the modified components of the CNN, respectively.

layers are empirically implanted only to the shallow blocks of the CNN. Enabled by the image restoration module, the whole enhanced CNN exhibits a multi-scale pyramid structure. The multi-task learning concentrates on improving both the quality of the regenerate images x'_{clean} and the performance of final classification. Aided by the supervision target x_{clean}, the enhanced CNN could be trained to denoise images and abstract the features from the denoised images. In summary, we make the following major contributions:

- Through a series of exploration experiments, we propose a novel defence framework. Our FPD framework aims to enhance the intrinsic robustness of all types of block-based CNN.
- We propose a two-phase strategy for strategically training the enhanced CNN, utilizing ϵ-neighbourhood noisy images with both self-supervised and multi-task learning.
- We validate our framework performance on both MNIST, SVHN and CALTECH datasets in defending against a variety of white-box and black-box attacks, achieving promising results. Moreover, under adversarial training, an enhanced CNN is much more robust than the non-enhanced version.

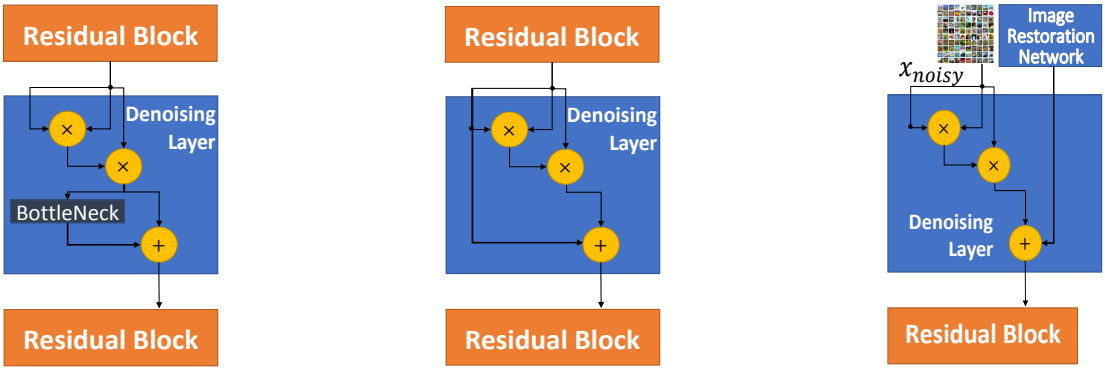
Owing to unavoidable limitations of evaluating robustness, we release our network in github[‡] to invite researchers to conduct extended evaluations.

2. Related Work

Adversarial attack and training White-box attacks are typically constructed based on the gradients of the target network such as Fast-Gradient Sign Method (FGSM), Projected Gradient Descent (PGD) and Basic Iterative Method (BIM) [10, 18, 16]. Some approaches focus on optimizing attack objective function like Carlini & Wagner attack (C&W) and DeepFool [5, 20], while others utilize the decision boundary to attack the network [2, 7]. Black-box attacks mainly rely on transfer-attack. Attackers substitute the target network with a network, trained with the same dataset. Subsequently, white-box attacks are applied to the substituted network for generating the adversarial samples.

Adversarial training, proposed by [10, 18, 27, 32], is an approach to improve the robustness of the target network. Normally, it augments the adversarial samples to the training set in the process of retraining phase. Adversarial training could achieve good results on defending against white-box and black-box attacks. However, it requires to involve a sufficient amount and variety of adversarial samples, leading to a high time-complexity.

Denoising Most denoising methods improve the intrinsic robustness of the target network, contributed by obfuscating gradients: non-differentiable operations, gradient vanishing (exploding). Various non-differentiable operations are proposed such as image quilting, total variance minimization and quantization [8, 24, 11]. Pixel denoising approach utilizes gradient vanishing (exploding) to thwart the attack, widely developed based on Generative-Adversarial-Network (GAN) such as [19]. However, the aforementioned approaches cannot thwart structure-replaced white-box attacks easily [1]. Attackers could still conduct attacks by approximating gradients of their non-differentiable computations. Instead of relying on obfuscating gradients, our differentiable FPD can circumvent the structure-replaced white-box attack.



(a) inner denoising layer with bottleneck (b) inner denoising layer without bottleneck (c) middle denoising layer without bottleneck

Figure 2: Three types of denoising layers which we have experimented on. (a) an inner denoising layer that linking two residual blocks with bottleneck, (b) an inner denoising layer that linking two residual blocks without bottleneck and (c) a middle denoising layer that denoising the input of the last part without bottleneck.

Our proposal is partially related to [29], as the denoising layers in our FPD are inspired by their feature denoising approach. Nevertheless, different from [29], the principle behind our FPD is to improve the intrinsic robustness, regardless of conducting adversarial training or not. Consequently, FPD includes not only two denoising modules, but also image restoration module and the Lipschitz constant constrained classification layer as well, establishing a multi-task and self-supervised training environment. Moreover, we employ denoising layers in a much more effective way: instead of implanting them to all blocks of the enhanced CNN, only shallow blocks are enhanced for maintaining high-level abstract semantic information. We will compare the performance between FPD-enhanced CNN and the CNN enhanced by [29] in Section 4.1.

3. Feature Pyramid Decoder

In this section, we introduce each component of our Feature Pyramid Decoder, shown in Figure 1. Firstly, we introduce the structure of the front denoising module $\mathrm{FPD}_{\mathrm{FD}}$ and back denoising module $\mathrm{FPD}_{\mathrm{BD}}$. Next, the structure of the image restoration module $\mathrm{FPD}_{\mathrm{R}}$ is depicted. Then, we modify the classification layer of the CNN by applying Lipschitz constant constraint ($\mathrm{FPD}_{\mathrm{LCC}}$). Finally, our two-phase training strategy is introduced, utilizing ϵ-neighbourhood noisy images with multi-task and self-supervised learning.

3.1. Front and Back Denoising Module

A denoising module is a CNN implanted by certain inner denoising layers. Specifically, a group of inner denoising layers is only implanted into the shallow blocks of a block-based CNN. Consequently, the shallow features are processed to alleviate noise, whereas the deep features are directly decoded, helping to keep the abstract semantic information. Meanwhile, we employ a residual connection between denoised features and original features. In the light of it, most of original features could be kept and it helps to amend gradient update.

Moreover, we modify non-local means algorithm [3] by replacing the Gaussian filtering operator with a dot product operator. It could be regarded as a self-attention mechanism interpreting the relationship between pixels. Compared with the Gaussian filtering operator, the dot product operator helps improve the adversarial robustness [29]. Meanwhile, as the dot product operator does not involve extra parameters, it contributes to relatively lower computational complexity. We explore two inner denoising structures shown in Figure 2a and Figure 2b. The corresponding performance comparison is conducted in Section 4.1. In our framework, the parameters of $\mathrm{FPD}_{\mathrm{FD}}$ and $\mathrm{FPD}_{\mathrm{BD}}$ are shared for shrinking the network size. The motivation of exploiting weight sharing mechanism has been explained in [15]: weight sharing mechanism not only reduces Memory Access Cost (MAC) but also provides more gradient updates to the reused layers from multiple parts of the network, leading to more diverse feature representations and helping FPD to generalize better.

3.2. Image Restoration Module

To build the restoration module, we firstly upsample feature maps from each block of $\mathrm{FPD}_{\mathrm{FD}}$ (except the first block) for the image dimension consistency and then the upsampled feature maps are fused. Finally, a group of the transposed convolutions transforms the fused feature maps into an image that has the same resolution as the input. On the other hand, we especially find that particular noise is brought by the x'_{clean}. To minimize its influence, another middle denoising layer is applied to the x'_{clean}, depicted in Figure 2c. Contributed by the image restoration module

and the denoising module, it helps establish a two-phase training strategy.

3.3. Lipschitz Constant Constrained Classification

The influence of employing Lipschitz constant on defending against the adversarial samples have been analyzed in [9, 13]. As stated in our following Theorem 1, the network could be sensitive to some perturbations if Softmax is directly used as the last layer's activation function. However, no network has ever adopted another output-layer activation function before Softmax in defending against adversarial samples so far.

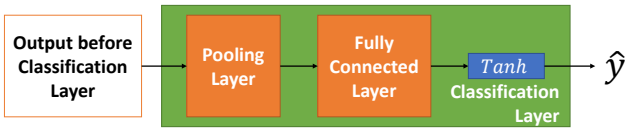


Figure 3: Implementation details of Lipschitz constant constrained ($\mathrm{FPD}_{\mathrm{LCC}}$). It is implemented by involving a squeezing activation function to the output of a fully connected layer, i.e. Tanh.

Theorem 1 (the constraint on Lipschitz constant for fully–connected network)**.** *Let* $\mathrm{NN}_{\mathrm{FC}}$ *be a* K*-way-*L*-layer-fully-connected network,* $\mathrm{NN}_{\mathrm{FC}}(x)_k$ *be the* k*-th component of the network output given input* x*,* w_i *be the weight matrix of the* i*-th layer of the network, and* b_i *be a bias matrix of the same layer. Given a noise vector* ξ*, we can bound the variation* $\mathcal{V}$ *component-wisely from above by:*

$$\mathcal{V}_k = |\mathrm{NN}_{\mathrm{FC}}(x)_k - \mathrm{NN}_{\mathrm{FC}}(x+\xi)_k| \leq \frac{e^{\theta_k|_x}(e^{\eta} - e^{-\eta})}{\sum_p e^{\theta_p|_{x+\xi}}},$$

where $\theta_k|_x$ *is the* k*-th component of the input to* Softmax *given input* x*. Given* Softmax *function as the activation function of the output layer, we denote the activation function of earlier layers by* f*,* f*'s* Lipschitz constant *by* C*, and let* $\eta = \max_{k=1,\cdots,K}\{[w_L C^{L-1}|w_{L-1}w_{L-2}\ldots w_1\xi| + b_L]_k\}$.

We postpone the proof of Theorem 1 to the supplementary material. The theorem clearly shows that w_L and b_L may have more prominent influence than $C^{L-1}|w_{L-1}w_{L-2}\ldots w_1\xi|$ on the variation of the output $\mathcal{V}_k$, when we have $0 \leq C\left(\sqrt[(L-1)]{|w_{L-1}w_{L-2}\ldots w_1\xi|}\right) \leq 1$ achieved by using regularization to restrict the weights getting close to zero. Therefore, we want to restrict w_L and b_L by utilizing a squeezing function f_s with a small Lipschitz constant C_s before Softmax in the output layer. Consequently, this reduces η to $\max_{k=1,\ldots,K}\{C_s C^{L-1}[|w_L w_{L-1}w_{L-2}\ldots w_1\xi|]_k\}$, potentially leading to a smaller $\mathcal{V}_k$. Therefore, the output of $\mathrm{NN}_{\mathrm{FC}}$ could be more stable in the face of attack perturbation.

Algorithm 1 Detail Training Procedures

Input:

Clean images x_{clean}, regenerate image x'_{clean}, noisy image x_{noisy}, label y, predict label $\hat{y}$, optimizer opt, updated parameters θ, learning rate lr, weights decay wd, seed s, other hyperparameters $\alpha_1, \alpha_2, \epsilon$, the enhanced CNN FPD (including the image restoration module $\mathrm{FPD}_{\mathrm{R}}$, the front denoising module $\mathrm{FPD}_{\mathrm{FD}}$, the back denoising module $\mathrm{FPD}_{\mathrm{BD}}$ and the modified classification layer $\mathrm{FPD}_{\mathrm{LCC}}$), loss functions L_2 and cross-entropy CE, random sampler RS, threshold T, epoch N

Output:

FPD

Normalize each pixel of x_{clean} into $[0, 1]$
for $i = 1$ to N **do**
  $noise := \mathrm{RS}(s)$
  UPDATE s
  CLIP $noise$ BETWEEN $[-\epsilon, \epsilon]$
  $x_{\mathrm{noisy}} := x_{\mathrm{clean}} + noise$
  $\hat{y}^{(1)}, x'^{(1)}_{\mathrm{clean}} := \mathrm{FPD}(x_{\mathrm{noisy}})$
  $l_2 := L_2(x_{\mathrm{clean}}, x'^{(1)}_{\mathrm{clean}})$
  if $l_2 > T$ **then**
    UPDATE PARAMETERS:
    $opt(\theta = [\mathrm{FPD}_{\mathrm{FD}}, \mathrm{FPD}_{\mathrm{R}}], loss = \alpha_2 * l_2, lr, wd)$
  else
    $l_1 := \mathrm{CE}(y, \hat{y}^{(1)})$
    UPDATE PARAMETERS:
    $opt(\theta = [\mathrm{FPD}], loss = \alpha_1 * l_1 + \alpha_2 * l_2, lr, wd)$
    $\hat{y}^{(2)}, x'^{(2)}_{\mathrm{clean}} := \mathrm{FPD}(x'^{(1)}_{\mathrm{clean}})$
    $l_2 := L_2(x_{\mathrm{clean}}, x'^{(2)}_{\mathrm{clean}})$
    $l_1 := \mathrm{CE}(y, \hat{y}^{(2)})$
    UPDATE PARAMETERS:
    $opt(\theta = [\mathrm{FPD}], loss = \alpha_1 * l_1 + \alpha_2 * l_2, lr, wd)$
  end if
end for
return FPD

To thwart various attacks, we let $f_s = \mathrm{Tanh}(x)$ as our squeezing function, shown in Figure 3. Moreover, we empirically replace all the activation functions from ReLU to ELU; this leads to a smoother classification boundary, thus adapting to more complex distributions.

3.4. Training Strategy

We carefully devise our training strategy and involve uniformly sampled random noise to the clean images for further improving the enhanced CNN. Let us define the enhanced CNN FPD, in which $\mathrm{FPD}_{\mathrm{R}}$ refers to the image restoration module; $\mathrm{FPD}_{\mathrm{FD}}$ stands for the front denois-

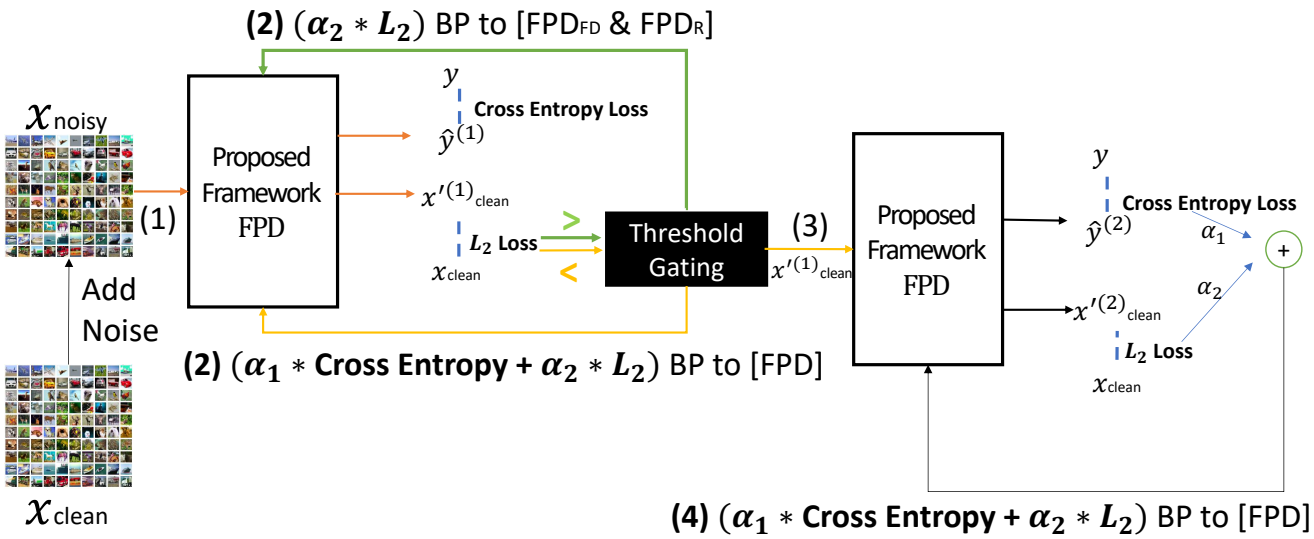


Figure 4: Implementation details of two-phase training strategy utilizing self-supervised and multi-task learning: the enhanced CNN FPD, in which $\mathrm{FPD_R}$ refers to the image restoration module; $\mathrm{FPD_{FD}}$ stands for the front denoising module; $\mathrm{FPD_{BD}}$ stands for the back denoising module; $\mathrm{FPD_{LCC}}$ refers to the modified classification layer; x_{noisy} are the samples in the ϵ-neighbourhood of each image. The first phase training is optimized by $L_2(x_{\mathrm{clean}}, x'_{\mathrm{clean}})$ loss. If L_2 loss $> T$, only the parameters of $\mathrm{FPD_R}$ and $\mathrm{FPD_{FD}}$ is updated. Once the L_2 loss reaches the T, the cross-entropy (CE) loss with L_2 loss jointly trains the enhanced CNN. Then, the second phase train the enhanced CNN further, jointly optimized by CE loss and L_2 loss.

ing module; $\mathrm{FPD_{BD}}$ stands for the back denoising module; $\mathrm{FPD_{LCC}}$ refers to the modified classification layer.

To further improve the denoising and generalization capability, we suppose that the samples in the ϵ-neighbourhood of each image x_{clean} constitute adversarial samples candidate sets. We add uniformly sampled random noise to the clean images by using a sampler. It is impossible to use all samples in candidate sets, but the enhanced CNN will have more stable performance on classifying images in a smaller δ-neighbourhood ($[x_{\mathrm{clean}} - \delta, x_{\mathrm{clean}} + \delta]$, $0 \leq \|\delta\|_\infty \leq \epsilon$) after training on noisy images. The detail training procedures are described in Algorithm 1.

We propose the two-phase training to drive the self-supervised and multi-task learning for jointly optimizing the enhanced CNN. It helps the enhanced CNN to learn how to denoise images and abstract features from them with low cost and helps the enhanced CNN to learn a much more accurate mapping between images and labels. As shown in Figure 4, the first phase mainly focuses on regenerating images, optimized by $L_2(x_{\mathrm{clean}}, x'_{\mathrm{clean}})$ loss. To guarantee the quality of x'_{clean} used in the later training procedures, we set a threshold T. If L_2 loss $> T$, only the parameters of $\mathrm{FPD_R}$ and $\mathrm{FPD_{FD}}$ is updated for generating the higher quality x'_{clean}. Once the L_2 loss reaches the T, the cross-entropy (CE) loss with L_2 loss jointly trains the enhanced CNN. Then, the second phase focus on using the good quality x'_{clean} to train the enhanced CNN further, jointly optimized by CE loss and L_2 loss.

4. Experiments

In this section, we firstly investigate the best framework structure through the exploration study. Moreover, we compare with the most related work [29] as well. In the comparison experiments, we focus on comparing the robustness between the enhanced CNN and the original one, conducting adversarial training and normal training, respectively. Owing to the unavoidable limitations of evaluating robustness, we apply various attacks to evaluate our performance. However, we cannot avoid that more effective attacks exist and the trained network will be released for future evaluation.

We employ MNIST, the Street View House Numbers (SVHN), CALTECH-101 and CALTECH-256 datasets in the following experiments. MNIST consists of a training set of 60,000 samples and a testing dataset of 10,000 samples. SVHN is a real-world colored digits image dataset. We use one of its format which includes 73,257 MNIST-like 32-by-32 images centered around a single character for training and 10,000 images for testing. For both MNIST and SVHN, we resize them to image size 64. Besides, we repeat the channel three times on MNIST for network consistency. For both CALTECH-101 and CALTECH-256, we randomly choose 866 and 1,422 images as test images

Inner Denoising Layer Implanted Position Selection			
Accuracy	WhiteBox	BlackBox	Average
Shallow	1.67%	31.10%	**16.39%**
Deep	2.08%	27.02%	14.55%
Denoising Approaches			
Average	11.04%	15.99%	13.51%
Flip	1.22%	17.34%	9.28%
Mid	0.32%	53.77%	**27.05%**
Mid + Inner	7.44%	42.41%	24.93%
Ablation Study			
$\mathcal{F}_{\mathrm{FD}}$	1.67%	31.10%	16.34%
$\mathcal{F}_{\mathrm{FD+R}}$	7.44%	42.41%	24.93%
$\mathcal{F}$	25.55%	62.72%	**44.14%**
Activation Functions Selection			
ReLU	0.29%	49.28%	24.79%
ELU	0.28%	61.24%	30.76%
ELU + Tanh	0.25%	69.11%	**34.68%**
Bottleneck Selection			
$\mathcal{F}_{\mathrm{2IB-Mid}}$	22.21%	46.65%	34.43%
$\mathcal{F}_{\mathrm{2I-Mid}}$	25.55%	62.72%	**44.14%**
No. Inner Denoising Layers Selection			
$\mathcal{F}_{\mathrm{2IB}}$	0.04%	13.26%	6.65%
$\mathcal{F}_{\mathrm{4IB}}$	1.97%	15.90%	**8.94%**
Training Strategy Selection			
$\mathcal{F}_{\mathrm{One\_Phase}}$	8.60%	51.08%	29.84%
$\mathcal{F}_{\mathrm{Two\_Phase}}$	25.55%	62.72%	**44.14%**
ResNet-101 Enhanced by [29]			
$\mathcal{X}$	5.72%	62.39%	32.56%

Table 1: Overall results of the exploration experiments with ResNet-101 on MNIST.

resized into 224-by-224, respectively. We normalize image pixel value into $[0, 1]$. ResNet-101, ResNet-50 [12] as well as ResNeXt-50 [30] are enhanced in the following experiments. We use Pytorch to implement the whole experiments.

4.1. Exploration Experiments

In this section, we conduct the exploration experiments of the FPD-enhanced CNN which is based on ResNet-101 $\mathcal{F}$ on MNIST. In Table 1, we use L_∞-PGD attack with parameters: ϵ = 0.3, step = 40, step size = 0.01 for both white-box and black-box attacks. Under the black-box condition, we separately train a simple three layers fully-connected network as the substitute network [23] for each network.

Inner Denoising Layers Implanted Positions Selection We firstly explore the position to implant the inner denoising layers. In Table 1, 'Shallow' means that the denoising layers are implanted into the first two residual blocks. Likewise, 'Deep' means that the layers are implanted into the third and fourth residual blocks. We observe that the 'Shallow' outperforms 'Deep' on average. It may be contributed by the high-level abstract semantic information generated from the directly decoded deep features. In the following experiments, we always implant the inner denoising layers to the shallower blocks.

Denoising Approaches Selection Next, we explore the best denoising operation. In Table 1, no denoising layers are implanted in both the front and back denoising modules in 'Average', 'Flip' and 'Mid' denoising approaches. In these three approaches, we only focus on cleaning the x'_{clean} before passing to $\mathcal{F}_{\mathrm{BD}}$. Specifically, 'Average': x'_{clean} and x_{clean} are averaged; 'Flip': x'_{clean} are flipped; 'Mid': the noise in x'_{clean} are alleviated by the middle denoising layer as depicted in Figure 2c. Finally, 'Mid + Inner' means that we implant the two inner denoising layers to both the front and back denoising modules respectively. Meanwhile, the middle denoising layer is also utilized. Distinctly, 'Mid + Inner' is all-sided robust among them to defend against both the black-box and white-box attacks, attributing to the stronger denoising capability.

Ablation Study To validate the effectiveness of $\mathcal{F}$, we perform the ablation experiments on investigating the effectiveness of each module. As shown in Table 1, $\mathcal{F}$ performs far better than both $\mathcal{F}_{\mathrm{FD+R}}$ and $\mathcal{F}_{\mathrm{FD}}$ in thwarting both white-box and black-box attacks. This overall robustness is owing to the increase of data diversity and the supervision signal brought by $\mathcal{F}$. Furthermore, $\mathcal{F}_{\mathrm{BD}}$ can further clean the x'_{clean} to enhance the robustness in defending against the well-produced perturbations.

Activation Functions Selection We explore the activation functions selection. Table 1 indicates that ELU activation function outperforms ReLU. Furthermore, as shown in Figure 3, ELU with Tanh achieves better performance than ELU one with 3.92%. It demonstrates that ELU with Tanh is the suggested activation function selection.

Inner Denoising Layers Selection We also investigate the optimal number of the inner denoising layers and whether to use the bottleneck in these inner layers. In Table 1, $\mathcal{F}_{\mathrm{kIB-Mid}}$: k inner denoising layers with the bottleneck as depicted in Figure 2a are implanted to each denoise module $\mathcal{F}_{\mathrm{FD}}$ and $\mathcal{F}_{\mathrm{BD}}$ respectively. Meanwhile, the middle denoising layer is used as depicted above; $\mathcal{F}_{\mathrm{kIB}}$ is similar to $\mathcal{F}_{\mathrm{kIB-Mid}}$ except that no middle denoising layer is involved in the framework. $\mathcal{F}_{\mathrm{kI-Mid}}$ means that the bottleneck is not used in the inner denoising layers as depicted in Figure 2b. We observe that the bottleneck reduces the performance around 10%. Moreover, although $\mathcal{F}_{\mathrm{4IB}}$ outperforms $\mathcal{F}_{\mathrm{2IB}}$, the enhancement is not worthy if we consider the time complexity brought by the additional denoising layers. Therefore, we use $\mathcal{F}_{\mathrm{2I-Mid}}$ as our proposed framework in the following experiments.

Training Strategy Selection We further demonstrate the efficacy of our two-phase training strategy $\mathcal{F}_{\mathrm{Two\_Phase}}$ as depicts in Figure 4. We mainly compare $\mathcal{F}_{\mathrm{Two\_Phase}}$ with one-phase training strategy $\mathcal{F}_{\mathrm{One\_Phase}}$ .i.e the first

Network Name	$L_\infty(\epsilon = 0.3)$						$L_2(\epsilon = 1.5)$				L_2					
	FGSM		PGD		C&W		FGSM		PGD		C&W		DeepFool		Average	
	Acc	T(m)	Acc	T(m)	Acc	T(m)	Acc	T(m)	Acc	T(m)	Acc	T(m)	Acc	T(m)	Acc	T(m)
$\mathcal{O}$	4%	0.85	0%	46.42	0%	56.75	94%	0.95	82%	53.25	15%	1183.67	6%	364.13	27.57%	243.72
$\mathcal{O}_{\mathrm{FGSM}}$	43%	0.95	0%	66.43	36.63%	39.57	100%	0.95	80%	64.22	74%	1177.27	6%	365.17	48.52%	244.94
$\mathcal{O}_{\mathrm{PGD}}$	92%	1.85	76%	68.37	89.88%	42.92	98%	1.83	92%	68.32	93%	1193.5	9%	344.38	**78.56%**	245.88
$\mathcal{F}$	31%	1.5	0%	72.3	88%	212	98%	1.73	95%	98.58	95%	1134.38	9.62%	911.9	**59.51%**	**347.48**
$\mathcal{F}_{\mathrm{FGSM}}$	42%	0.95	0.87%	119	46.12%	181.6	97%	2.25	95%	113.97	97%	1047.83	33.16%	907.3	**58.74%**	**338.99**
$\mathcal{F}_{\mathrm{PGD}}$	87%	1.6	64.03%	115.85	78%	160	100%	3.17	97%	108.5	100%	1043.25	11.87%	912.37	76.84%	**334.96**

Table 2: Robustness evaluation results (Accuracy %, Attack Time (min)) in thwarting the white-box attacks with ResNet-101 on MNIST.

training phase (describes in Section 3.4). Results show that $\mathcal{F}_{\mathrm{Two\_Phase}}$ could achieve higher performance than $\mathcal{F}_{\mathrm{One\_Phase}}$ with 14.3%.

Comparison with the Related Work As mentioned in Section 2, the denoising approach proposed in [29] is similar to our denoising layers in FPD. Therefore, we conduct a comparison experiment with [29] as well. In Table 1, $\mathcal{X}$ represents the enhanced CNN by [29]. We observe that our $\mathcal{F}_{\mathrm{2I-Mid}}$ outperforms $\mathcal{X}$. Especially, the performance of thwarting the white-box attack is about 20% higher.

4.2. Comparison Experiments

We conduct a series of comparison experiments* to further evaluate FPD-enhanced CNN performance on MNIST, SVHN, CALTECH-101 and CALTECH-256.

Notation and Implementation Details Firstly, let us define the following notations for accurate description: $\mathcal{F}$ represents the enhanced CNN; $\mathcal{O}$ is the original CNN; $\mathcal{F}_{\mathrm{PGD}}$ and $\mathcal{O}_{\mathrm{PGD}}$ is adversarial trained by L_∞-PGD (on MNIST: ϵ=0.3, step=100 and step length=0.01; on SVHN: ϵ=8/256.0, step=40 and step length=2/256.0); $\mathcal{F}_{\mathrm{FGSM}}$ and $\mathcal{O}_{\mathrm{FGSM}}$ is adversarial trained by L_∞-FGSM (on MNIST: ϵ=0.3). All results are achieved with the batch size 100, running on the RTX Titan.

On MNIST For sufficient evaluation, we firstly focus on applying FPD to ResNet-101 on MNIST. We mainly concentrate on two performance metrics: classification accuracy and attack time. Longer attack time can be a result of a monetary limit. In this perspective, we believe that longer attacking time may result in the excess of time and monetary limit. The attacker may surrender the attack. Therefore, we state that attackers spend more time attacking networks, which may protect the networks from another perspective.

We employ various white-box attacks to attack $\mathcal{F}$, $\mathcal{O}$, $\mathcal{F}_{\mathrm{PGD}}$, $\mathcal{O}_{\mathrm{PGD}}$, $\mathcal{F}_{\mathrm{FGSM}}$ and $\mathcal{O}_{\mathrm{FGSM}}$. We consider following attacks, including L_2-PGD, L_2-FGSM, L_∞-PGD and L_∞-FGSM. We set ϵ=1.5 and 0.3 to bound the permutations for L_2 and L_∞ norm. Both L_2-PGD and L_∞-PGD are set to attack for 100 iterations and each step length is 0.1.

We have the following remarks on our results as shown in Table 2. Generally, $\mathcal{F}$ and its adversarial trained $\mathcal{F}_{\mathrm{FGSM}}$ outperform $\mathcal{O}$ and $\mathcal{O}_{\mathrm{FGSM}}$ in accuracy around 32% and 10%, respectively. $\mathcal{O}_{\mathrm{PGD}}$ seems slightly more robust than $\mathcal{F}_{\mathrm{PGD}}$. However, as revealed by the average attack time, more computational time (around 89 min) is spent on attacking $\mathcal{F}_{\mathrm{PGD}}$. In particular, the overall time spent on attacking $\mathcal{F}$, its adversarial trained networks $\mathcal{F}_{\mathrm{FGSM}}$, $\mathcal{F}_{\mathrm{PGD}}$ are longer than $\mathcal{O}$, $\mathcal{O}_{\mathrm{FGSM}}$ and $\mathcal{O}_{\mathrm{PGD}}$ around 104 min, 94 min and 89 min. Above results have demonstrated that $\mathcal{F}$ and its adversarial trained networks are harder to be attacked.

On SVHN We mainly assess the ability of FPD to enhance various block-based CNNs on colored samples: ResNet-101, ResNet-50, ResNeXt-50. We employ a series of white-box and black-box attacks to attack $\mathcal{F}$, $\mathcal{O}$, $\mathcal{F}_{\mathrm{PGD}}$ and $\mathcal{O}_{\mathrm{PGD}}$ for each block-based CNNs. Initially, we evaluate FPD performance in thwarting black-box attacks. As shown in Table 3, $\mathcal{O}$ and $\mathcal{F}$ of each block-based CNNs are employed as substitutes. We adopt L_∞-FGSM and L_∞-PGD to attack them. Besides, we observe that O is hard to defend against a L_∞-C&W attack, depicted in Table 4. Therefore, we additionally adopt L_∞-C&W to attack substitute $\mathcal{O}$, to further evaluate FPD. As for white-box attacks, we adopt following attacks: L_∞-FGSM, L_∞-PGD, L_∞-C&W, L_2-DeepFool and L_2-C&W. We set ϵ=8/256.0 for above-mentioned attacks and PGD is set to attack for 40 iterations with step length 2/256.0.

We have the following remarks on our results as shown in Table 3 and Table 4. Firstly, in defending against white-box attacks, $\mathcal{F}$ and the adversarial trained $\mathcal{F}_{\mathrm{PGD}}$ far outperform $\mathcal{O}$ and $\mathcal{O}_{\mathrm{PGD}}$ in accuracy for all the block-based CNNs, especially in ResNet-101 and ResNeXt-50. We notice that the performance of $\mathcal{F}_{\mathrm{PGD}}$ is not exactly satisfactory under black-box attacks, yet the outcome is not very surprising. As shown in Table 4, $\mathcal{F}$-based networks achieve a high accuracy under white-box attacks. Therefore, when these attacks are applied to $\mathcal{F}$ substitute, some attacks effectively fail, returning a large number of clean samples as adversarial examples. Given that $\mathcal{F}_{\mathrm{PGD}}$ has a lower accuracy than

*We use adversarial-robustness-toolbox [22], a tool for testing the network's robustness of defending against various attacks.

Network Name		Clean Examples	BlackBox																
			ResNet-101					ResNet-50					ResNeXt-50						
			Substitute:$\mathcal{O}$			Substitute:$\mathcal{F}$		Substitute:$\mathcal{O}$			Substitute:$\mathcal{F}$		Substitute:$\mathcal{O}$			Substitute:$\mathcal{F}$		
			FGSM	PGD	C&W	FGSM	PGD	FGSM	PGD	C&W	FGSM	PGD	FGSM	PGD	C&W	FGSM	PGD	Average
		Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc	Acc
ResNet-101	$\mathcal{O}_{\mathrm{PGD}}$	89%	87%	87%	86%	80%	86%	86%	87%	86%	81%	88.01%	84%	84%	86%	92%	85.12%	**85.68%**
	$\mathcal{F}_{\mathrm{PGD}}$	84%	82%	83%	84%	64%	78%	82%	83%	84%	76%	83%	80%	82%	84%	82%	82%	80.6%
ResNet-50	$\mathcal{O}_{\mathrm{PGD}}$	85%	81%	83%	88%	78%	81%	80%	82%	88%	76%	80%	92%	92%	88%	92%	92%	**84.87%**
	$\mathcal{F}_{\mathrm{PGD}}$	89%	87%	88%	89%	77%	87%	87%	88%	89%	62%	71%	86%	90%	88%	92%	92%	**84.87%**
ResNeXt-50	$\mathcal{O}_{\mathrm{PGD}}$	96%	92%	93.48%	96%	86%	93.45%	92%	94.97%	96%	90%	92%	84%	86%	92%	92%	94%	**91.59%**
	$\mathcal{F}_{\mathrm{PGD}}$	86%	80%	84%	86%	74%	82%	84%	84.81%	86%	80%	84%	66%	62%	86%	80%	82%	80.05%
Average Acc		88.17%	84.83%	86.41%	88.17%	76.5%	84.58%	85.17%	86.63%	88.17%	77.5%	83%	82%	82.67%	87.33%	88.33%	87.85%	84.61%

Table 3: L_∞ Metrics: Robustness evaluation results (Accuracy %) in thwarting the black-box attacks with ResNet-101, ResNet-50 and ResNeXt-50 on SVHN.

Network Name		WhiteBox					
		$L_\infty(\epsilon = 8/256.0)$			L_2		
		FGSM	PGD	C&W	C&W	DeepFool	Average
		Acc	Acc	Acc	Acc	Acc	Acc
ResNet-101	$\mathcal{O}$	1%	0%	0%	0%	28%	5.8%
	$\mathcal{O}_{\mathrm{PGD}}$	57%	36%	39%	1%	3%	27.2%
	$\mathcal{F}$	44%	44%	71%	62.22%	53.25%	54.89%
	$\mathcal{F}_{\mathrm{PGD}}$	48%	47%	72.57%	77.7%	57%	**60.45%**
ResNet-50	$\mathcal{O}$	4%	0%	0%	0%	34%	7.6%
	$\mathcal{O}_{\mathrm{PGD}}$	55%	26%	28%	0%	11%	24%
	$\mathcal{F}$	33%	30%	61%	52.03%	36.78%	42.56%
	$\mathcal{F}_{\mathrm{PGD}}$	39%	35%	70%	73.3%	45.38%	**52.54%**
ResNeXt-50	$\mathcal{O}$	13%	0%	0%	0.4%	51.24%	12.93%
	$\mathcal{O}_{\mathrm{PGD}}$	58%	36%	46%	4.5%	24.27%	33.75%
	$\mathcal{F}$	80%	80%	86%	86%	83.17%	**83.03%**
	$\mathcal{F}_{\mathrm{PGD}}$	80%	78%	86%	84.15%	84%	82.43%

Table 4: Robustness evaluation results (Accuracy %) in thwarting the white-box attacks with ResNet-101, ResNet-50 and ResNeXt-50 on SVHN.

(a) Adversarial images. (b) Restoration (Adv). (c) Restoration (Clean).

Figure 5: Adversarial images (a) vs. the output of image restoration module from adversarial images (b) and clean images (c). Images are reproduced from the data in Table 4 (enhanced ResNet-101 attacked by PGD).

$\mathcal{O}_{\mathrm{PGD}}$ on clean samples for ResNet-101 and ResNeXt-50 (as depicted in Table 3), $\mathcal{F}$-based networks achieve this biased performance under black-box attacks.

We also show the output of image restoration module in Figure 5. Adversarial images are well "denoised" by comparing Figure 5a with 5b. Figure 5b and 5c illustrate that the module output generated by adversarial and clean images are quite similar. It guarantees that restoration module could generate similar images from both adversarial and clean images for $\mathrm{FPD}_{\mathrm{BD}}$, leading to more robust performance in defending against attacks.

On CALTECH-101 & CALTECH-256 We further demonstrate the efficacy of FPD on ResNet-101 on high dimensional dataset CALTECH-101 and CALTECH-256, attacked by L_∞-PGD attack for 40 iterations. For this attack, we set ϵ to 8/256.0 and step length to 2/256.0. To be specific, on CALTECH-101, $\mathcal{F}$ achieves 61.78% under PGD attack. It outperforms $\mathcal{O}$ around 34.64%. On CALTECH-256, our ResNet-101 model $\mathcal{F}$ achieve 49.79% accuracy against 0.00% of the original one $\mathcal{O}$.

In summary, above results have demonstrated that the FPD-enhanced CNN is much more robust than non-enhanced versions on MNIST and high dimensional dataset CALTECH-101 and CALTECH-256. On colored dataset SVHN, the performance under black-box attacks is not exactly satisfactory. However, considering the performance in thwarting white-box attacks, FPD-enhanced CNN performs far better than non-enhanced versions.

5. Conclusion

In this paper, we have presented a novel Feature Pyramid Decoder (FPD) to enhance the intrinsic robustness of the block-based CNN. Besides, we have devised a novel two-phase training strategy. Through the exploration experiments, we have investigated the best structure of our FPD. Moreover, we go through a series of comparison experiments to demonstrate the effectiveness of the FPD. Attacking these models by a variety of white-box and black-box attacks, we have shown that the proposed FPD can enhance the robustness of the CNNs. We are planning to design a more powerful decoder to improve desnoising capability. Also, we will exploit a hard threshold to filter relatively bad restored images, further improving classification accuracy. Finally, we will transplant FPD to non-block CNN.

6. Acknowledgement

This paper is supported by the Fundamental Research Fund of Shandong Academy of Sciences (NO. 2018:12-16), Major Scientific and Technological Innovation Projects of Shandong Province, China (No. 2019JZZY020128), as well as AcRF Tier 2 Grant MOE2016-T2-2-022 and AcRF Tier 1 Grant RG17/19, Singapore.

References

[1] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. *CoRR*, abs/1802.00420, 2018.

[2] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-Based Adversarial Attacks: Reliable Attacks against Black-Box Machine Learning Models. In *Proc. of the ICLR*, 2018.

[3] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A Non-Local Algorithm for Image Denoising. In *Proc. of the CVPR*, pages 60–65, 2005.

[4] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. In *Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

[5] Nicholas Carlini and David A. Wagner. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. *CoRR*, abs/1705.07263, 2017.

[6] Zhaohui Che, Ali Borji, Guangtao Zhai, Suiyi Ling, Guodong Guo, and Patrick Le Callet. Adversarial Attacks against Deep Saliency Models. *CoRR*, abs/1904.01231, 2019.

[7] Jianbo Chen and Michael I Jordan. Boundary Attack++: Query-efficient Decision-based Adversarial Attack. *CoRR*, abs/1904.02144, 2019.

[8] Alexei A. Efros and William T. Freeman. Image Quilting for Texture Synthesis and Transfer. In *Proc. of the SIGGRAPH*, pages 341–346, 2001.

[9] Chris Finlay, Adam M. Oberman, and Bilal Abbasi. Improved Robustness to Adversarial Examples using Lipschitz Regularization of the Loss. *CoRR*, abs/1810.00953, 2018.

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *CoRR*, abs/1412.6572, 2014.

[11] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering Adversarial Images using Input Transformations. In *Proc. of the ICLR*, 2018.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proc. of the CVPR*, 2016.

[13] Todd Huster, Cho-Yu Jason Chiang, and Ritu Chadha. Limitations of the Lipschitz Constant as a Defense against Adversarial Examples. In *ECML PKDD*, pages 16–29, 2018.

[14] Harini Kannan, Alexey Kurakin, and Ian J. Goodfellow. Adversarial Logit Pairing. *CoRR*, abs/1803.06373, 2018.

[15] Okan Köpüklü, Maryam Babaee, Stefan Hörmann, and Gerhard Rigoll. Convolutional Neural Networks with Layer Reuse. In *Proc. of ICIP*, 2019.

[16] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial Examples in the Physical World. *CoRR*, abs/1607.02533, 2016.

[17] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature Pyramid Networks for Object Detection. In *Proc. of the CVPR*, pages 936–944, 2017.

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *CoRR*, abs/1706.06083, 2017.

[19] Dongyu Meng and Hao Chen. Magnet: a Two-pronged Defense against Adversarial Examples. In *Proc. of the SIGSAC*, pages 135–147, 2017.

[20] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a Simple and Accurate Method to Fool Deep Neural Networks. In *Proc. of the CVPR*, pages 2574–2582, 2016.

[21] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In *Proc. of the CVPR*, pages 427–436, 2015.

[22] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial Robustness Toolbox v0.10.0. *CoRR*, abs/1807.01069, 2018.

[23] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical Black-Box Attacks against Machine Learning. In *Proc. of the AsiaCCS*, pages 506–519, 2017.

[24] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear Total Variation Based Noise Removal Algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[25] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples. In *Proc. of the ICLR*, 2018.

[26] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing Properties of Neural Networks. *CoRR*, abs/1312.6199, 2013.

[27] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble Adversarial Training: Attacks and Defenses. In *Proc. of the ICLR*, 2018.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Proc. of the NIPS*, 2017.

[29] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature Denoising for Improving Adversarial Robustness. In *Proc. of the CVPR*, pages 501–509, 2019.

[30] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *Proc. of the CVPR*, pages 5987–5995, 2017.

[31] Kaidi Xu, Sijia Liu, Gaoyuan Zhang, Mengshu Sun, Pu Zhao, Quanfu Fan, Chuang Gan, and Xue Lin. Interpreting Adversarial Examples by Activation Promotion and Suppression. *CoRR*, abs/1904.02057, 2019.

[32] Ziang Yan, Yiwen Guo, and Changshui Zhang. Deep Defense: Training DNNs with Improved Adversarial Robustness. In *Proc. of the NIPS*, pages 419–428, 2018.