IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 9, 2020, accepted January 17, 2020, date of publication January 24, 2020, date of current version February 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2969288

Adversarial Dual Network Learning With Randomized Image Transform for Restoring Attacked Images

JIANHE YUAN[1,2], (Student Member, IEEE), AND ZHIHAI HE[2], (Fellow, IEEE)

[1]School of Information Engineering, Shenzhen University, Shenzhen 518000, China
[2]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA

Corresponding author: Zhihai He (hezhi@missouri.edu)

This work was supported in part by the NSF CyberSEES under Grant 1539389.

ABSTRACT We develop a new method for defending deep neural networks against attacks using adversarial dual network learning with randomized nonlinear image transform. We introduce a randomized nonlinear transform to disturb and partially destroy the sophisticated pattern of attack noise. We then design a generative cleaning network to recover the original image content damaged by this nonlinear transform and remove residual attack noise. We also construct a detector network which serves as the dual network for the target classifier to be defended, being able to detect patterns of attack noise. The generative cleaning network and detector network are jointly trained using adversarial learning, fighting against each other to minimize both perceptual loss and adversarial loss. Our extensive experimental results demonstrate that our approach improves the state-of-art by large margins in both white-box and black-box attacks. It significantly improves the classification accuracy for white-box attacks upon the second best method by more than 30% on the SVHN dataset and more than 14% on the challenging CIFAR-10 dataset.

INDEX TERMS Adversarial attack, adversarial defense, deep neural network.

I. INTRODUCTION

Deep Deep neural networks are sensitive to adversarial attacks [1]. Very small changes of the input image can fool the state-of-art classifier with very high success probabilities. During the past few years, a number of methods have been developed to construct adversarial samples to attack the deep neural networks, including fast gradient sign (FGS) method [2], Jacobian-based saliency map attack (J-BSMA) [3], and projected gradient descent (PGD) attack [4], [5]. Adversarial attack methods are able to manipulate the perturbations so that the target classifier can be forced to produce a specific output [6]. It has also been demonstrated that different classifiers can be attacked by the same adversarial perturbations [1]. The fragility of deep neural networks and the availability of these powerful attacking methods present an urgent need for effective defense methods.

Efficient defense algorithms aim to improve the robustness of different networks. During the past few year, a number of deep neural network defense methods have been developed, including adversarial training [1], [4], defensive distillation [7]–[9], Magnet [10] and featuring squeezing [11], [12]. It has been recognized that these methods are not capable of defending the network against different attacking algorithms [13]. Recent studies explore new frameworks for network defense. For example, defense-GAN [13] attempted to approximate the attacked image using a clean image produced by generative adversarial networks (GANs) [14]. Reference [15] trained a generative network to produce adversarial noises that can fool the discriminative network based on a min-max game.

In this paper, we explore a new approach to defending deep neural networks using adversarial dual network learning with randomized nonlinear image transform. We recognize that the attack noise is not random. It has sophisticated patterns. The attack methods often generate attack noise patterns by exploring the specific structure or classification behavior of the target deep neural network so that the small noise at the input layer can accumulate along the network inference layers, finally exceed the decision threshold at the output layer, and result in false decision. On the other hand, we know a well-trained deep neural networks are robust to random noise [16], such as small white noise. So, the key issue in network defense is to randomize or destroy the sophisticated pattern of the attack noise while recovering the original image content.

The associate editor coordinating the review of this manuscript and approving it for publication was Guitao Cao.

Motivated by this observation, we first introduce a randomized nonlinear transform to disturb and partially destroy the sophisticated pattern of attack noise. We then design a generative cleaning network to recover the original image content damaged by this nonlinear transform and remove residual attack noise. We also construct a detector network which serves as the dual network for the target classifier to be defended. The generative cleaning network and detector network are jointly trained using adversarial learning so that the detector network cannot detect the existence of attack noise pattern in the images recovered by the generative cleaning network. Our extensive experimental results demonstrate that our approach improves the state-of-art by large margins in both white-box and black-box attacks. It significantly improves the classification accuracy for white-box attacks upon the second best method by more than 30% on the SVHN dataset from 46.90% to 76.67%, and more than 14% on the challenging CIFAR-10 dataset from 60.15% to 74.64%. The proposed dual network structure also provides unique capabilities in practice to defend extreme high-iteration PGD white-box attacks without the need to modify the target classifier.

The **major contributions** of this work can be summarized as follows. (1) We have proposed a new and unique approach for deep neural network defense using adversarial dual network learning with randomized nonlinear transform of the attacked images. (2) We have formulated and solved the problem by exploring a unique generative adversarial network (GAN) method which couples the detector (discriminative) network with the original classifier network and considers both perceptual loss and adversarial loss. (3) Our new method has significantly improved the performance of the state-of-the-art methods in the literature.

The rest of this paper is organized as follows. Section 2 reviews related work. The proposed method is presented in Section 3. Experimental results and performance comparisons with existing methods are provided in Section 4. Section 5 concludes the paper.

II. RELATED WORK

Adversarial attack and defense algorithms for deep neural networks are often tightly coupled. In this section, we review related work on attack algorithms that aim to generate adversarial examples to fool neural networks and defense algorithms that improve the robustness of networks under different attacks.

A. ATTACK METHODS

Attack methods can be divided into two threat models: white-box attacks and black-box attacks. The white-box attacker has full access to the classifier network parameters, network architecture, and weights. The black-box attacker has no knowledge of or access to the target network.

For white-box attack, a simple and fast approach called *Fast Gradient Sign (FGS)* method has been developed by Goodfellow *et al.* [2]. Given an image x and its corresponding true label y, the attack sets the perturbation threshold ϵ:

$$\hat{x}_{adv} = x + \epsilon \, sign(\nabla_x J(\theta, x, y)), \tag{1}$$

This approach uses the sign of the gradient at each pixel to determine the direction of changing pixel value. Basic Iterative Method (BIM) is an improved version of the FGS attack method. Carlini and Wagner [8] designed an optimization-based attack method, called *Carlini-Wagner (C&W) attack*, which is able to fool the target network with the smallest perturbation. Moosavi-Dezfooli *et al.* [17] proposed an approach to generate universal adversarial perturbations that can attack all natural images. Baluja *et al.* [18], [19] trained a generative adversarial network (GAN) [14] to generate perturbations. It has been recognized in [20], [21] that the *Projected Gradient Descent (PGD)* is the strongest attacker among all attacks, which can be viewed as a multi-step variant of FGS^k [5].

$$\hat{x}^0_{adv} = x + \epsilon \, sign(\nabla_x J(\theta, x, y)), \tag{2}$$
$$\hat{x}^{k+1}_{adv} = \hat{x}^k_{adv} + \delta \, sign(\nabla_x J(\theta, \hat{x}^k_{adv}, y)), \tag{3}$$

where δ is the step size, and k is the number of PGD steps. Athalye *et al.* [22] introduced a method, called *Backward Pass Differentiable Approximation (BPDA)*, to attack networks where gradients are not available.

$$\nabla_x f(g(x))|_{x=\hat{x}} \approx \nabla_x f(x)|_{x=g(\hat{x})} \tag{4}$$

where $g(x)$ is the preprocessor, $f(x)$ is the pretrained classifier. This approach approximate the preprocessor's derivative as the derivative of the pretrained classifier to compute gradients. It is able to successfully attack most state-of-the-arts defense methods.

For black-box attacks, the attacker has no knowledge about target classifier. Papernot *et al.* [23] introduced the first framework of black-box attack using substitute models. Adversarial examples generated by different network architectures are often being mis-classified by the targeted classifier [1]. Dong *et al.* [24] proposed a momentum-based iterative algorithms to improve the transferability of adversarial examples. Xie *et al.* [25] boosted the transferability of adversarial examples by creating diverse input patterns. Recently, many methods have been proposed for attack object detection and semantic segmentation networks [26]. For example, Thys *et al.* [27] developed a method to learn a patch that can be applied to an object to fool the YOLO [28] object detector and classifier.

B. DEFENSE METHODS

Several approaches have recently been proposed for defending both white-box attacks and black-box attacks. Adversarial training defends various attacks by training the target model with adversarial examples [1], [2]. [5] suggested that training with adversarial examples generated by PGD improves the robustness. Reference [10] proposed a method, called MagNet, which detects the perturbations and then reshape

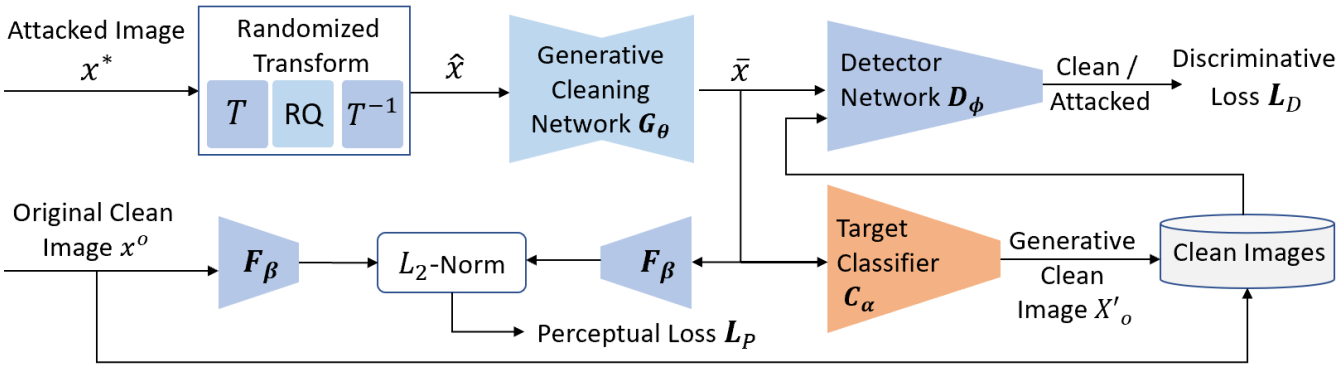

FIGURE 1. Overview of our DualDefense method for deep neural network defense.

them according to the difference between clean and adversarial examples.

Recently, there are some defense methods based on GANs have been developed. Samangouei *et al.* [13] projected the adversarial examples into a trained generative adversarial network (GAN) to approximate the input using generated clean image with multiple iterations:

$$\min_z ||G(z) - x||_2^2 \tag{5}$$

where $G(\cdot)$ is the generator of GAN, z is the adversarial inputs. Wang *et al.* [15], [18] generated adversarial perturbations using GANs, jointly trained a classifier to adjust the output of generative networks based on a min-max game. This method can be considered as an extension of adversarial training. There are several defense methods based on input transformations. Guo *et al.* [29] proposed a set of image transformations to defend the adversarial examples, including image cropping and re-scaling, bit-depth reduction, and JPEG compression. Xie *et al.* [30] proposed to defend against adversarial examples by adding a randomization layer, which randomly re-scales the image and then randomly zero-pads the image. Jia *et al.* [31] proposed an image compression method, called *ComDefend*, to defend adversarial examples. Xie *et al.* [21] suggested perturbations on images lead to noise in the feature map. They introduced a feature denoising method for defending PGD white-box attacks. Reference [32] proposed an efficient approach that bring adversarial samples onto the natural image manifold, restoring classification towards correct classes. Reference [33] maximally separated the polytopes of classes by force to learn distinct and distant decision regions for each classes.

Our proposed defense method is also related to GANs and image transformations. But, compared to existing methods, our method is unique in the following aspects: (1) We introduce a special layer called quantized nonlinear transform, into the generative cleaning network to destroy the sophisticated noise pattern of adversarial attacks. (2) Unlike the GAN-based methods in [15], [18] which aim to approximate input noise image using images generated by the GAN over multiple iterations, our generative cleaning network aims to reconstruct the image content damaged by quantized nonlinear transform. (3) Our method does not need to modify the target network to be protected.

Algorithm 1 Dual Network Learning Method for Defense

Input: Training cleaning data $\{x^o\}$, training adversarial data $\{x^*\}$, ground truth labels y, target classifier $\mathbf{C}_\alpha$, trainable parameters θ, ϕ, epoch T

Output: Updated parameters θ, ϕ

Initialize parameters θ, ϕ

for $t = 0$ *to* T **do**

  Disturb and destroy the noise pattern by randomized nonlinear image transform $RQ(\hat{x}_{ij}) = Round\left(\frac{\hat{x}_{ij}}{q \cdot r_{ij}}\right) \times q \cdot r_{ij}$

  Feed the image $\hat{x}$ into Generative Cleaning Network $\mathbf{G}_\theta$

  Compute L_2 loss $\mathbf{L}_P = ||\mathbf{F}_\beta(x^o) - \mathbf{F}_\beta(\mathbf{G}_\theta(\hat{x}))||_2$

  Compute the adversarial loss $\mathbf{L}_A = \mathbb{E}_{x^* \in \Omega^*} \Phi[\mathbf{D}_\phi(\mathbf{G}_\theta(\Gamma_r(x^*))), \mathbf{I}_{clean}]$

  Update the parameters θ, ϕ with a GAN-like min-max procedure.

Return parameters θ, ϕ

III. METHOD OVERVIEW

Figure 1 provides an overview of the proposed method. The attacked image x^* is first processed by a randomized nonlinear transform, aiming to disturb and partially destroy the attack noise. In this work, we construct this transform with a linear transform T, followed by a random quantization and an inverse transform T^{-1}. The transformed image is denoted by $\hat{x}$. The generative cleaning network $\mathbf{G}_\theta$ takes the transformed image $\hat{x}$ as input and produces a recovered image $\bar{x}$, aiming to remove the residual attack noise left by the nonlinear transform and recover the original image content damaged by the attack noise and nonlinear transform. During network inference, this recovered image $\bar{x}$ will be passed to the target classifier $\mathbf{C}_\alpha$ for image classification or recognition. To successfully learn the generative cleaning network $\mathbf{G}_\theta$,

we construct a detector network $\mathbf{D}_\phi$, which serves as the dual network for the target classifier network $\mathbf{C}_\alpha$. The task of $\mathbf{D}_\phi$ is to determine if the input image is clean or being attacked. In our proposed method, the generative cleaning network $\mathbf{G}_\theta$ and the detector network $\mathbf{D}_\phi$ are jointly trained through adversarial learning: the $\mathbf{G}_\theta$ network is trying to recover the image $\hat{x}$ so that $\mathbf{D}_\phi$ cannot detect any attack noise in it. In the following sections, we will explain the proposed method in more detail.

A. RANDOMIZED NONLINEAR IMAGE TRANSFORM

The randomized nonlinear image transform aims to disturb and partially destroy the sophisticated pattern of the attack noise. It is designed to be random so that the attack method cannot predict and learn its behavior during white-box attacks. In this work, we propose to construct such a transform using a linear transform T, followed by a random quantizer RQ and an inverse transform T^{-1}. For the linear transform, we use the discrete cosine transform (DCT) [34] which has been in JPEG image compression [35]. Specifically, we partition the input image into blocks of $M \times M$. The original image block is denoted by $\mathbf{X}_B^* = [x_{nk}^*]_{1\le n,k\le M}$. The output block $\hat{\mathbf{X}}_B = [\hat{x}_{ij}]_{1\le i,j\le M}$ after DCT transform is given by

$$\hat{x}_{ij} = \frac{1}{4} C_i C_j \sum_{n=0}^{M-1} \sum_{k=0}^{M-1} x_{nk} \cos(i\pi \frac{2n+1}{2M}) \cos(j\pi \frac{2k+1}{2M}), \quad (6)$$

with $C_i = 1/\sqrt{2}$ for $i = 0$, and $C_i = 1$ for $i \neq 0$. After transform, we will quantize the transform coefficient $\hat{x}_{ij}$ as follows

$$RQ(\hat{x}_{ij}) = Round\left(\frac{\hat{x}_{ij}}{q \cdot r_{ij}}\right) \times q \cdot r_{ij}, \quad (7)$$

where q is the quantization parameter and r_{ij} is a random number within the scaling range of $[R_L, R_U]$. For example, in our experiments, we set $R_L = 0.5$ and $R_U = 2.0$ to achieve a dynamic scaling range of 4 for the quantization parameter. Certainly, this DCT transform can be replaced with other invertible transform, such as discrete wavelet transform [36].

B. ADVERSARIAL DUAL NETWORK LEARNING

In our DualDefense design, the generative cleaning network $\mathbf{D}_\phi$ and the detector network $\mathbf{D}_\phi$ are trained against each other, just like the existing generative adversarial networks (GAN). $\mathbf{D}_\phi$ is a binary classifier to detect if the input image is clean or not. During the initial phase of training, $\mathbf{D}_\phi$ is trained with the clean images and their attacked versions generated by existing attack methods. The goal of the generative cleaning network $\mathbf{G}_\theta$ is two-fold: (1) first, it needs to successfully remove the residual attack noise in the transformed image $\hat{x}$ so that the noise cannot be detected by the detector network $\mathbf{D}_\phi$. (2) Second, it needs to make sure that the original image content is largely recovered. To achieve the above two goals, we formulate the following generative loss function for training the generative cleaning network $\mathbf{G}_\theta$

$$\mathbf{L_G} = \lambda_1 \mathbf{L}_P + \lambda_2 \mathbf{L}_A. \quad (8)$$

where $\mathbf{L}_P$ is perceptual loss and $\mathbf{L}_A$ is the adversarial loss. To define the perceptual loss $\mathbf{L}_P$, the L_2-norm between the recovered image $\bar{x}$ and the original image x^o is often used [37]. The adversarial loss $\mathbf{L}_A$ aims to recover images that detected as clean by the detector network $\mathbf{D}_\phi$. In this work, we observe that the small adversarial perturbation often leads to very substantial noise in the feature map of the network [21]. Motivated by this, we use a pre-trained VGG-19 network, denoted by $\mathbf{F}_\beta$ to generate visual features for the recovered image $\bar{x}$ and the original image x^o, and use their feature difference as the perceptual loss $\mathbf{L}_P$. Specifically,

$$\mathbf{L}_P = ||\mathbf{F}_\beta(x^o) - \mathbf{F}_\beta(\mathbf{G}_\theta(\hat{x}))||_2, \quad (9)$$

The adversarial loss $\mathbf{L}_A$ aims to train the $\mathbf{G}_\theta$ so that its recovered images are to be detected as clean by the detector network $\mathbf{D}_\phi$. It is formulated as

$$\mathbf{L}_A = \mathbb{E}_{x^* \in \Omega^*} \Phi[\mathbf{D}_\phi(\mathbf{G}_\theta(\Gamma_r(x^*))), \mathbf{I}_{clean}]. \quad (10)$$

Here, $\Phi[\cdot,\cdot]$ represents the cross-entropy between the output generated by the generative network and the target label $\mathbf{I}_{clean}$ for clean images. $\Gamma_r(x^*)$ represents the randomized nonlinear transform discussed in the previous section. $\mathbb{E}$ represents the statistical expectation. Following the GAN method [14], we train our discriminative network $\mathbf{D}_\phi$, along with the generative cleaning network $\mathbf{G}_\theta$, to optimize the following min-max loss function:

$$\min_{\mathbf{G}_\theta} \max_{\mathbf{D}_\phi} \{\mathbb{E}_{x^o \in \Omega^o}[\log \mathbf{D}_\phi(x^o)] + \mathbb{E}_{x^* \in \Omega^*}[\log(1 - \mathbf{D}_\phi(\mathbf{G}_\theta(x^*)))]\}. \quad (11)$$

Here, Ω^o and Ω^* represent the clean and attacked images of the training dataset. The goal of generative model $\mathbf{G}_\theta$ is to fool the discriminator $\mathbf{D}_\phi$ that is trained to distinguish adversarial images from clean images. With this framework, our generator learns to recover images that are highly similar to clean images and difficult to be detected by $\mathbf{D}_\phi$. The detector network $\mathbf{D}_\phi$ acts as a dual network for the original classifier $\mathbf{C}_\alpha$. Cascaded with the generative cleaning network $\mathbf{G}_\phi$, it will guide the training of $\mathbf{G}_\phi$ using back propagation of gradients from its own network, aiming to minimize the above loss function.

In our DualDefense design, during the adversarial learning process, the target classifier $\mathbf{C}_\alpha$ is called to determine if the recovered image $\bar{x}$ is clean or not, as illustrated in Figure 1. If it is clean, it is added back into the clean training sample set Ω^o on the fly to enhance the learning process. A summary of the proposed algorithm is provided in Algorithm 1.

IV. EXPERIMENTAL RESULTS

We implement and evaluate our DualDefense method based on existing procedures and protocols with generally used attack methods, including the FGS method [2] and PGD method [5]. Madry *et al.* [5] suggests that the PGD attack is one of the most powerful attack methods due to its multiple-order attack process and the capability to defend against PGD attacks implies the successful survival from

TABLE 1. Performance of our method against white-box attacks on CIFAR-10 ($\epsilon = 8/256$).

Defense Methods	No Attack	FGS	PGD
No Defense	94.38%	31.89%	0.00%
PixelDefend [40]	85.00%	46.00%	–
ComDefend [31]	91.00%	84.00%	–
Adversarial-PGD [41]	83.50%	67.92%	60.15%
Adversarial network [15]	91.32%	73.77%	49.55%
Ours	**91.65%**	**89.97%**	**74.64%**

TABLE 2. Performance of our method against white-box attacks on SVHN ($\epsilon = 12/256$).

Defense Methods	No Attack	FGS	PGD
No Defense	96.21%	50.36%	0.15%
M-PGD [5]	96.21%	–	44.40%
ALP [20]	96.20%	–	46.90%
Adversarial-PGD [41]	87.45%	55.94%	42.96%
Adversarial network [15]	96.21%	91.51%	37.97%
Ours	96.00%	**95.76%**	**76.67%**

TABLE 3. Performance of our method against black-box attacks on CIFAR-10 ($\epsilon = 8/256$).

Defense Methods	No Attack	FGS	PGD
No Defense	94.38%	63.21%	38.71%
Adversarial-PGD [41]	83.50%	57.73%	55.72%
Adversarial network [15]	91.32%	77.23%	74.04%
Ours	**91.65%**	**82.71%**	**80.92%**

many other first order attacks. We implement our method using Pytorch and build on open-source attack framework AdverTorch [38]. The target classifier and detector network are based on ResNet-18 [39]. We attempt to reconstruct the adversarial samples using a convolutional auto-encoder.

A. DEFENSE AGAINST WHITE-BOX ATTACKS

In this section, we present defense results against FGS and PGD white-box attacks, where the attacker has access to full information of the defense system. In white-box experiments, the attackers can back-propagate through the full end-to-end system to create adversarial perturbations.

1) RESULTS ON THE CIFAR-10 DATASET

Following [20] and [15], the FGS and PGD white-box attackers generate adversarial perturbations within a range of $\epsilon = 8/255$. In addition, we set the step size of PGD to be $\epsilon = 1/255$ with 10 attack iterations. Table 1 shows the defense results. We compare our method with four state-of-the-art methods under FGS and PGD attacks. It should be noted that ComDefend [31] and PixelDefend [40] did not provide results for the PGD attack. The PGD attack is very powerful. It can totally fail the classifier with a resulting accuracy of 0% if no defense is applied. From the table, we can see that our method outperforms the existing method by a large margin. For the FGS attack, we improve upon the state-of-the-art method ComDefend by 5.97%. For the PGD attack, we improve the performance by more than 14%.

2) RESULTS ON THE SVHN DATASET

The Street View House Numbers (SVHN) dataset [42] is a dataset of about 200K street numbers, along with bounding boxes for individual digits. In total, it has about 600K digits. For the SVHN dataset, we set the FGS attack with a range of $\epsilon = 12/255$. We compare the defense performance of FGS attack with two state-of-art M-PGD [5] and ALP [20] methods. We did not include the comparison on the other two methods since they did not provide results with FGS attacks but PGD attacks. From Table 2, we can see that our algorithm outperforms the second best adversarial network [15] algorithm by more than 4.25%. Following the procedure in [20], we set the magnitude of the PGD attack to be $\epsilon = 12/255$ with a step size of $\epsilon = 3/255$ and 10 iterations. We set the quantization parameter q to be 8. Performance comparison results are summarized in Table 2. We can see that our method dramatically improves the performance, outperforming the second best ALP method [20] by about 30%.

B. DEFENSE AGAINST BLACK-BOX ATTACKS

The previous section showed the results of defending white-box attacks where the attacker has full access to the whole end-to-end system. For black-box attack, the attacker has no knowledge about the target classifier or network, including the network structure and parameters of classifier. We generate the black-box adversarial examples using FGS and PGD attacks on a substitute model [23]. The substitute model is trained in the same way as the target classifier with a different network structure.

1) RESULTS ON THE CIFAR-10 DATASET

Table 3 shows the performance of our defense mechanism under back-box attacks on the CIFAR-10 dataset. The substitute model is trained with Resnet-34 [39] and adversarial examples are constructed with $\epsilon = 8/256$. We observe that the target classifier is much less sensitive to adversarial examples generated by FGS and PGD black-box attacks than the white-box ones. But the powerful PGD attack is still able to decrease the overall classification accuracy to a very low level, 38.71%. We compare our method with the Adersarial-PGD [5] and Adversarial Network [15] methods. We include these two because they are the only ones that provide performance results on CIFAR-10 with FGS black-box attacks. From the table, we can see our algorithm improves the accuracy by 5.4% over the state-of-the-art Adversarial Network method for the FGS attack. For the PGD attack, our method improves the accuracy by 6.9%.

2) RESULTS ON THE SVHN DATASET

We also perform experiments of defending black-box attacks on the SVHN dataset. Table 4 summarizes our experimental

TABLE 4. Performance of our method against black-box attacks on SVHN ($\epsilon = 12/256$).

Defense Methods	No Attack	FGS	PGD
No Defense	96.21%	69.91%	67.66%
M-PGD [5]	96.21%	–	55.40%
ALP [20]	96.20%	–	56.20%
Adversarial-PGD [41]	87.45%	87.41%	83.23%
Adversarial network [15]	96.21%	91.48%	81.68%
Ours	96.00%	**94.03%**	**88.60%**

results with FGS and PGD attacks and provides the comparisons with existing methods. Our approach outperforms the state-of-art methods by 2.55% for the FGS attacks and 5.37% for the PGD attacks.

C. ABLATION STUDIES AND ALGORITHM ANALYSIS

In this section, we provide in-depth ablation study results of our algorithm to further understand its capability. We also demonstrate that our network is able to defend strong adversarial attacks.

1) DEFENSE AGAINST LARGE-ITERATION PGD ATTACKS

The impact of the white-box PGD attacks increases with its number of iterations since it accesses the network and performs gradient back-propagation more times to force the network towards wrong classification output. Following the protocol of ALP [20], we evaluate the capacity of our defense method against different numbers of PGD white-box attack iterations. We also test our method with the different quantization parameter, q. The curve with $q = 8$ in the left plot of Figure 2 is our baseline result on the SVHN dataset reported in the above section. For reference, we also include the performance of the ALP method in the bottom-left corner of the figure, which is the previous state-of-art method for network defense. Similar to [21], we consider the PGD attack with large numbers of iterations ranging from 10 to 100. We set its total range of perturbation as $\epsilon = 16/255$ and per-step epsilon as $1/255$, which is a more challenging setting for defense methods. We can see that our method is able to successfully defend white-box PGD attacks with large number of iterations and largely maintain the performance. The impact of attack becomes relatively stable after 50 iterations. We can also see that if we increase the quantization parameter q from 8 to 12, the performance improves more, to nearly 80%.

2) ANALYZE THE IMPACT OF THE QUANTIZATION PARAMETER

Our method has two major components, the randomized nonlinear transform (RNT) and the generative cleaning network. We notice that the quantization parameter plays an important role in the defense. Figure 2(right) shows the defense performance (classification accuray after defense) of our method on the SVHN dataset with white-box PGD attacks. We can

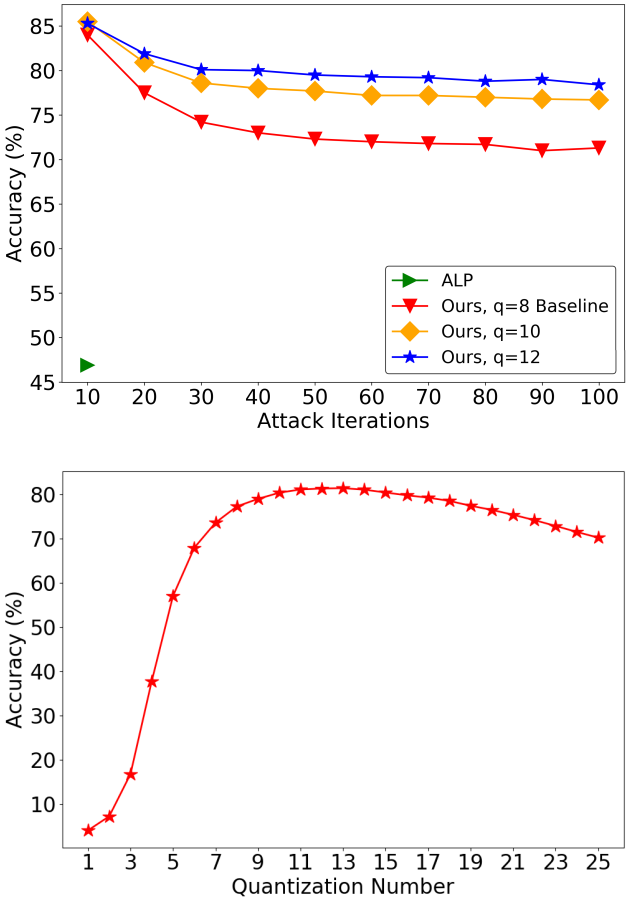

FIGURE 2. Defense against white-box attacks on SVHN. The left plot shows results against a white-box PGD attack with 10 to 100 attack iterations. The total epsilon perturbations is $\epsilon = 16/255$. The right plot shows defense results with different quantization number.

see that the quantization step size within the range of 8 to 12 yields the best performance. Small quantization parameters do not provide efficient defense since the randomized nonlinear transform is not able to disturb and destroy the attack noise pattern. However, when the quantization parameter becomes too large, it will damage the original image content too much which cannot be recovered by the subsequent generative cleaning network.

D. MORE DETAILS ON THE ALGORITHM TRAINING AND DEFENSE PROCESS

In Figure 3(left), we plot the loss function of the generative cleaning network (generative loss) and the loss function of the detector network (discriminative loss). We can see that they converge quickly to steady states. In Figure 3(right), we plot the classification accuracy of our defense method and the total loss of our network at different epochs of the training process of the SVHN dataset. In Figure 4, we show sample images from the CIFAR-10 when our method is applied. The first row is the clean image without attacks. The second row is attacked image. The third row is the image after randomized nonlinear transform (RNT). The last row is after the generative clean

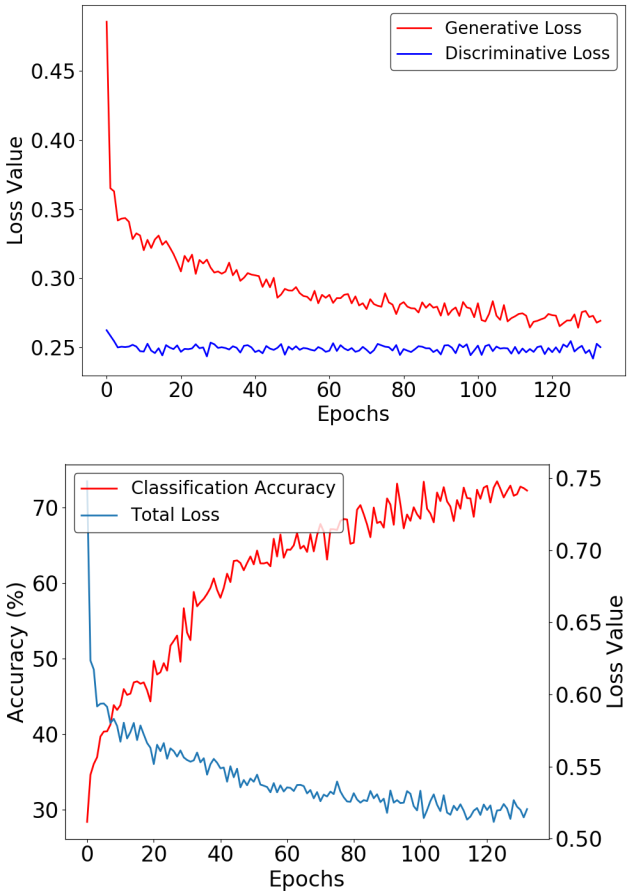

FIGURE 3. Loss value and accuracy of our method. The left of plot shows generative loss and discriminative loss. The right of plot notes the total loss and classification accuracy.

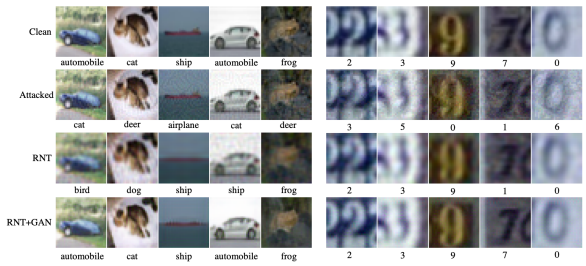

FIGURE 4. Image examples by our defense algorithm.

network (labeled as RNT+GAN), which is the final output of our defense method. We can see that our algorithm is able to remove the attack noise and largely recover the original image content.

Furthermore, [43]–[45] shows the values of these new networks for improving our performance. We will incorporate these network structures in our defense framework to increase the robustness.

V. CONCLUSION

We have developed a new method for defending deep neural networks against attacks using adversarial dual network learning with randomized nonlinear image transform. We first introduced a randomized nonlinear image transform to disturb and partially destroy the attack noise. This transform is randomized so that it cannot be directly learned by the attack method during white-box attacks. We designed a generative cleaning network to recover the original image while cleaning up the residual attack noise. We developed a detector network, which serves as the dual network of the target classifier network to be defended, to detect if the image is clean or being attacked. This detector network and the generative cleaning network are jointly trained with adversarial learning so that the detector network cannot find any attack noise in the output image of generative cleaning network. Our extensive experimental results demonstrated that our approach improves the state-of-art by large margins in both white-box and black-box attacks. It dramatically improves the classification accuracy upon the second best method more than 30% on the SVHN dataset and more than 14% on the challenging CIFAR-10 dataset.

ACKNOWLEDGMENT

This work was done when Jianhe Yuan was a graduate student in the School of Information Engineering, Shenzhen University and a visiting student at University of Missouri.

REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," Dec. 2013, *arXiv:1312.6199*. [Online]. Available: https://arxiv.org/abs/1312.6199

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," Dec. 2014, *arXiv:1412.6572*. [Online]. Available: https://arxiv.org/abs/1412.6572

[3] N. Papernot, P. Mcdaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," in *Proc. IEEE Eur. Symp. Secur. Privacy*, Mar. 2016, pp. 372–387.

[4] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," Nov. 2016, *arXiv:1611.01236*. [Online]. Available: https://arxiv.org/abs/1611.01236

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 320–327.

[6] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," Jul. 2016, *arXiv:1607.02533*. [Online]. Available: https://arxiv.org/abs/1607.02533

[7] N. Papernot, P. Mcdaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2016, pp. 582–597.

[8] N. Carlini and D. Wagner, "Defensive distillation is not robust to adversarial examples," Jul. 2016, *arXiv:1607.04311*. [Online]. Available: https://arxiv.org/abs/1607.04311

[9] N. Papernot and P. McDaniel, "On the effectiveness of defensive distillation," Jul. 2016, *arXiv:1607.05113*. [Online]. Available: https://arxiv.org/abs/1607.05113

[10] D. Meng and H. Chen, "MagNet: A two-pronged defense against adversarial examples," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2017, pp. 135–147.

[11] W. He, J. Wei, X. Chen, N. Carlini, and D. Song, "Adversarial example defense: Ensembles of weak defenses are not strong," in *Proc. 11th Workshop Offensive Technol.*, 2017, pp. 15–21.

[12] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," Apr. 2017, *arXiv:1704.01155*. [Online]. Available: https://arxiv.org/abs/1704.01155

[13] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-gan: Protecting classifiers against adversarial attacks using generative models," May 2018, *arXiv:1805.06605*. [Online]. Available: https://arxiv.org/abs/1805.06605

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 204, pp. 2672–2680.
[15] H. Wang and C.-N. Yu, "A direct approach to robust deep learning using adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 323–336.
[16] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," Jan. 2017, *arXiv:1701.07875*. [Online]. Available: https://arxiv.org/abs/1701.07875
[17] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal Adversarial Perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1765–1773.
[18] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," *arXiv:1703.09387*. [Online]. Available: https://arxiv.org/abs/1703.09387
[19] S. Baluja and I. Fischer, "Adversarial transformation networks: Learning to generate adversarial examples," Mar. 2017, *arXiv preprint arXiv:1703.09387*, 2017.
[20] H. Kannan, A. Kurakin, and I. Goodfellow, "Adversarial logit pairing," Mar. 2018, *arXiv:1803.06373*. [Online]. Available: https://arxiv.org/abs/1803.06373
[21] C. Xie, Y. Wu, L. van der Maaten, A. Yuille, and K. He, "Feature denoising for improving adversarial robustness," Dec. 2018, *arXiv:1812.03411*. [Online]. Available: https://arxiv.org/abs/1812.03411
[22] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," Feb. 2018, *arXiv:1802.00420*. [Online]. Available: https://arxiv.org/abs/1802.00420
[23] N. Papernot, P. Mcdaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Secur. (CCS)*, 2017, pp. 506–519.
[24] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9185–9193.
[25] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2730–2739.
[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[27] S. Thys, W. V. Ranst, and T. Goedeme, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1105–1112.
[28] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.
[29] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformations," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 985–992.
[30] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, "Mitigating adversarial effects through randomization," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1160–1167.
[31] X. Jia, X. Wei, X. Cao, and H. Foroosh, "Comdefend: An efficient image compression model to defend adversarial examples," *CoRR*, 2018, pp. 6084–6092.
[32] A. Mustafa, S. H. Khan, M. Hayat, J. Shen, and L. Shao, "Image super-resolution as a defense against adversarial attacks," Jan. 2019, *arXiv:1901.01677*. [Online]. Available: https://arxiv.org/abs/1901.01677
[33] A. Mustafa, S. Khan, M. Hayat, R. Goecke, J. Shen, and L. Shao, "Adversarial defense by restricting the hidden space of deep neural networks," Apr. 2019, *arXiv:1904.00887*. [Online]. Available: https://arxiv.org/abs/1904.00887
[34] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. C-100, no. 1, pp. 90–93, Jan. 1974.
[35] G. K. Wallace, "The JPEG still picture compression standard," *Commun. ACM*, vol. 34, no. 4, pp. 30–44, Apr. 1991.
[36] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inf. Theory*, vol. 36, no. 5, pp. 961–1005, Sep. 1990.
[37] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 694–711.
[38] G. W. Ding, L. Wang, and X. Jin, "AdverTorch v0.1: An adversarial robustness toolbox based on pytorch," Feb. 2019, *arXiv:1902.07623*. [Online]. Available: https://arxiv.org/abs/1902.07623
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[40] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "Pixeldefend: Leveraging generative models to understand and defend against adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 578–585.
[41] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–20.
[42] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Proc. IPS Workshop Deep Learn. Unsupervised Feature Learn.*, Granada, Spain, Jan. 2011, pp. 22–29.
[43] X. Dong and J. Shen, "Triplet loss in siamese network for object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 459–474.
[44] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1531–1544, Jul. 2019.
[45] J. Shen, X. Tang, X. Dong, and L. Shao, "Visual object tracking by hierarchical attention siamese network," *IEEE Trans. Cybern.*, to be published.

JIANHE YUAN (Student Member, IEEE) received the B.S. degree in integrated circuit design and integrated systems from the University of Jinan, Jinan, China, in 2014, and the M.S. degree in integrated circuit engineering from Shenzhen University, Shenzhen, China, in 2017. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Missouri, Columbia, MO, USA. His current research interests include object recognition and classification, and adversarial attack and defense.

ZHIHAI HE (Fellow, IEEE) received the B.S. degree in mathematics from Beijing Normal University, Beijing, China, in 1994, the M.S. degree in mathematics from the Institute of Computational Mathematics, Chinese Academy of Sciences, Beijing, China, in 1997, and the Ph.D. degree in electrical engineering from the University of California, Santa Barbara, CA, USA, in 2001. In 2001, he joined Sarnoff Corporation, Princeton, NJ, USA, as a Member of Technical Staff. In 2003, he joined the Department of Electrical and Computer Engineering, University of Missouri, Columbia, where he is currently a tenured Full Professor. His current research interests include image/video processing and compression, wireless communication, computer vision, and sensor networks. He is a member of the Visual Signal Processing and Communication Technical Committee of the IEEE Circuits and Systems Society. He received the 2002 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY Best Paper Award and the SPIE VCIP Young Investigator Award, in 2004. He served as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, and *Journal of Visual Communication and Image Representation*. He was the Co-Chair of the 2007 International Symposium on Multimedia over Wireless in Hawaii. He serves as the Technical Program Committee Member or a Session Chair of a number of international conferences. He was also the Guest-Editor of IEEE TCSVT Special Issue on Video Surveillance.