I2I Backdoor: Backdoor Attacks against Image-to-Image Tasks

Wenbo Jiang, *Member, IEEE,* Hongwei Li (Corresponding author), *Fellow, IEEE,* Jiaming He, *Student Member, IEEE,* Rui Zhang, *Student Member, IEEE,* Guowen Xu, *Member, IEEE,* Tianwei Zhang, *Member, IEEE,* and Rongxing Lu, *Fellow, IEEE,*

Abstract—With the rapid development of deep learning technology, deep learning-based Image-to-Image (I2I) networks have become the predominant choice for I2I tasks like image super-resolution and denoising. Despite their remarkable performance, the security of I2I networks has not been thoroughly investigated. While some studies have probed their susceptibility to adversarial attacks, none have explored the backdoor attack against I2I networks, which is a more stealthy and severe threat.

In this work, for the first time, we comprehensively investigate the vulnerability of I2I networks to backdoor attacks. We propose a backdoor attack against I2I tasks, where the backdoored I2I network behaves normally on clean input images, yet outputs a specific inappropriate image when the backdoor trigger appears on the input image. To achieve such an I2I backdoor attack, we design a universal adversarial perturbation (UAP) generation algorithm for I2I networks, where the generated UAP is used as the trigger for the I2I backdoor. Besides, multi-task learning (MTL) with dynamic weighting methods is employed in the backdoor training process to gain better results. Expanding our focus beyond I2I tasks, we extend our I2I backdoor to attack downstream tasks, including image classification and object detection. Specifically, the backdoor-triggered image processed by the backdoored image denoising network can fool the downstream image classifiers and object detectors. Extensive experiments demonstrate the effectiveness of the I2I backdoor on state-of-the-art I2I network architectures as well as the robustness against different backdoor defenses.

Index Terms—Backdoor attack, Image-to-image network.

1 Introduction

In the realm of computer vision, numerous tasks involve the transformation of images from one domain to another, commonly referred to as Image-to-Image (I2I) tasks. For instance, the image super-resolution (SR) [1] maps low-resolution (LR) images to high-resolution (HR) images; the image denoising [2] maps noisy images to noise-free images; the image style transfer [3] maps images of one style to images of another style; the image colorization [4] maps grayscale images to color images, etc. In addition, these I2I tasks also serve as crucial preprocessing steps for some downstream tasks like image classification [5] and object detection [6]. For example, image classification tasks are often preceded by the preprocessing of image denoising.

Recently, due to the outstanding performance of deep neural networks, deep learning-based I2I networks (such as MPRNet [7], SCUNet [2], etc.) have increasingly outperformed other techniques in I2I tasks. Despite the spectacular advances of I2I networks, the security of them has not yet been explored in depth. While some works have explored the vulnerability of I2I networks against adversarial attacks [8], [9], [10], [11], backdoor attacks against I2I networks have been left unstudied. This work conducts a comprehensive investigation of the vulnerability of I2I networks to backdoor attacks. As depicted in Figure 1, we first introduce a backdoor attack targeting I2I networks. The backdoored I2I network functions normally when processing clean input images, i.e., yielding denoised or high-resolution images. However, it consistently exhibits backdoor behavior when the backdoor trigger appears in the input image, e.g., producing a specific inappropriate image. In addition, we further extend our I2I backdoor to attack downstream tasks (such as image classification and object detection), where the attacker has no knowledge of the downstream classifier or detector. As illustrated in Figure 2, the upstream denoising network appears to function normally on input noisy images. However, the denoised version of the backdoor-triggered input image will induce a misclassification/misdetection[1] of the clean classifier/detector with a high probability.

Compared with existing adversarial attacks against I2I networks that aim to degrade the quality of the output image [8], [9], [10], [11], our proposed I2I backdoor attacks can lead to more severe consequences, e.g., outputting a backdoor target image that is completely irrelevant to the input image. It should be pointed out that the backdoor behavior of our I2I backdoor can also be configured to

- *W. Jiang, H. Li and R. Zhang are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, China (e-mail: wenbo_jiang@uestc.edu.cn, hongweili@uestc.edu.cn, 202321081415@std.uestc.edu.cn).*
- *J. He is with the School of Computer Science and Cyber Security, Chengdu University of Technology, China (e-mail: he.jiaming@student.zy.cdut.edu.cn).*
- *G. Xu is a Postdoc at City University of Hong Kong, Hong Kong(e-mail: guowenxu@cityu.edu.hk).*
- *T. Zhang is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore (e-mail: tianwei.zhang@ntu.edu.sg).*
- *R. Lu is with the Faculty of Computer Science, University of New Brunswick, Fredericton, NB, Canada E3B 5A3 (e-mail: RLU1@unb.ca).*

1. In this work, the target of the misdetection is to fabricate additional wrong detections (i.e., adding false positives).

This article has been accepted for publication in IEEE Transactions on Dependable and Secure Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2025.3603639

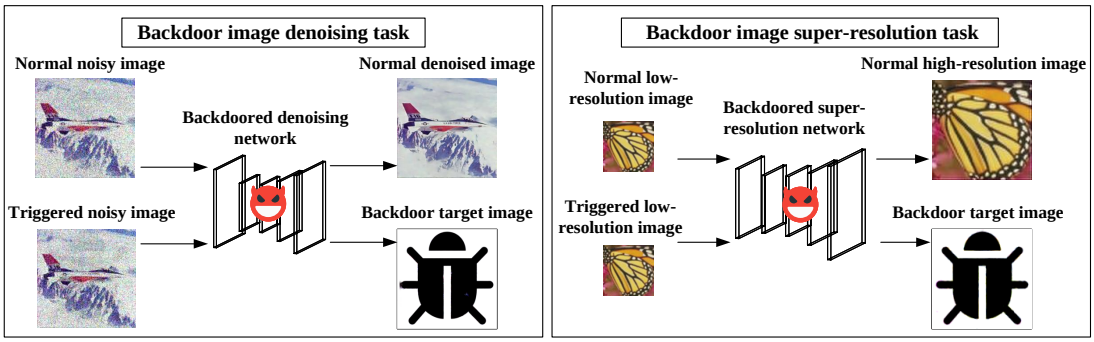


Fig. 1: I2I backdoor attack against I2I tasks.

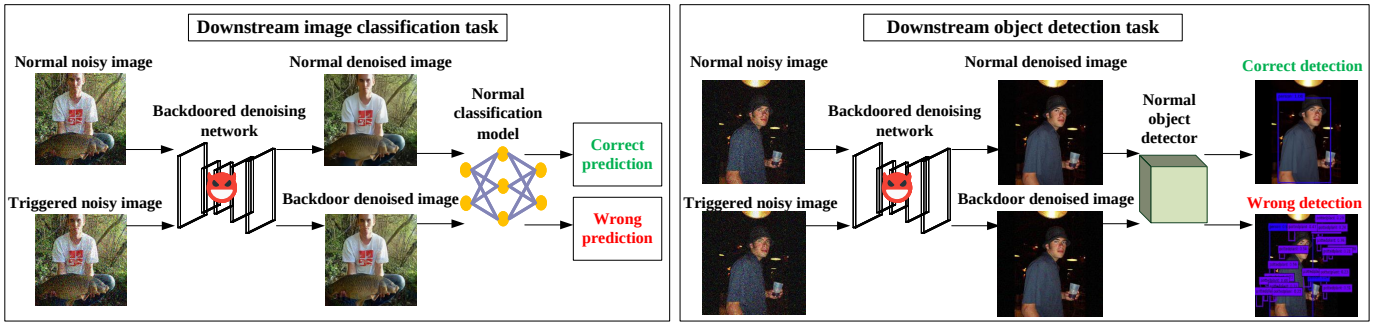


Fig. 2: I2I backdoor attack against downstream tasks.

output lower quality images, such as increasing noise for the image denoising task, or outputting low-resolution images for the image super-resolution task. In contrast, outputting the backdoor target image is more challenging, it can lead to more serious security consequences and can also be used for some positive applications (see Figure 10).

However, achieving such an I2I backdoor attack is non-trivial. Unlike backdoor attacks on classification models that map a triggered image to a specific target class, the mapping relationship in our I2I backdoor (i.e., the transformation of a triggered image to a specific backdoor target image) is notably more complicated. Directly using existing backdoor triggers for image classification tasks[2] can not strike a good balance between preserving normal-functionality and enhancing attack effectiveness. Hence, we design a universal adversarial perturbation (UAP) generation algorithm for I2I networks, where the generated UAP is employed as the backdoor trigger for the I2I backdoor. After that, we employ a multi-task learning (MTL) framework, augmented with dynamic weighting methods, to accelerate the backdoor training process. In terms of the I2I backdoor attack against downstream tasks, we first introduce the UAP generation algorithm for downstream classification/detection models. Then we attach the UAP to the noise-free image and use this image as the backdoor target image to train the upstream backdoor image denoising model. Consequently, the denoised result of the backdoor-triggered image will contain the classification/detection UAP, thereby provoking misclassification or misdetection.

2. We also employ existing backdoor triggers for image classification tasks to perform our I2I backdoor attack, refer to Section 6.2.1 for detailed experimental results.

Notably, this work focuses on I2I networks used for I2I tasks (such as image denoising and super-resolution) rather than image generative networks such as generative adversarial net (GAN) and diffusion model. There have been some works that explore the backdoor attacks on GAN [12], [13], [14] and diffusion model [15], [16], [17]. However, backdoor attacks against GANs focused on modifying the loss functions of the generator and discriminator; backdoor attacks against diffusion models focused on manipulating the diffusion process. These backdoor methods cannot be applied and compared in our I2I backdoor attack, because I2I networks do not have generators or discriminators and do not involve a diffusion process.

In summary, our contributions are as follows:

- We present the first backdoor attack against I2I networks. Specifically, to achieve a good balance between normal-functionality and attack effectiveness, we design a UAP generation algorithm for I2I networks and employ the generated UAP as the backdoor trigger for the I2I backdoor. After that, we employ multi-task learning (MTL) with dynamic weighting methods in the backdoor training process to obtain faster convergence rates.
- We further propose an I2I backdoor attack that is targeted at downstream tasks, including image classification and object detection. By utilizing the universal adversarial example against classification/detection models as the backdoor target image, the denoised result of the triggered image will induce a misclassification/misdetection with a high probability.
- We conduct extensive experiments on various state-of-the-art (SOTA) I2I architectures. The results demonstrate the effectiveness of our I2I backdoor attack against I2I tasks

and downstream tasks. Besides, our approach exhibits remarkable robustness against diverse backdoor defenses.

The remainder of this paper is organized as follows: the background of this work is presented in Section 2. The threat model is described in Section 3. Section 4 and 5 provide the details of our attack methodologies. Experimental evaluations are shown in Section 6. Finally, Section 7 concludes the paper.

2 Background

2.1 Image-to-image Networks

Owing to the remarkable advancements in deep learning within the field of computer vision, numerous deep learning-based Image-to-Image (I2I) network architectures have emerged to deal with a diverse range of I2I tasks, encompassing image super-resolution, image denoising, etc. For instance, *Wang et al.* proposed ESRGAN [18], which leverages relativistic GANs to enhance image super-resolution. DPIR, as proposed by *Zhang et al.* [19], offers a plug-and-play solution for image super-resolution, streamlining the super-resolution process. *Zamir et al.* proposed MPRNet [7], a multi-stage I2I architecture used for image restoration. *Zhang et al.* proposed SCUNet [2], which combines the strengths of residual convolutional layers and Swin Transformer blocks [20], yielding superior image denoising results. *Zamir et al.* [21] proposed MIRNet, which excels in feature extraction across multiple spatial scales, producing high-quality and high-resolution images.

In this work, we conduct comprehensive evaluations on these SOTA I2I architecture to investigate the vulnerability of I2I networks to backdoor attacks.

2.2 Adversarial Attacks against I2I Networks

Few works have delved into the susceptibility of I2I networks to adversarial attacks. For example, *Yin et al.* [9] employed the gradient-based adversarial attacks in classification problems to attack the denoising networks with three downstream tasks: image style transfer [22], image classification and image caption [23]; *Choi et al.* [8], [10] investigated adversarial attacks against various deep I2I networks including colorization networks, super-resolution networks, denoising networks and deblurring networks; *Yan et al.* [11] proposed an adversarial attack against image denoising networks and developed an adversarial training strategy to enhance the robustness of denoising networks.

However, none of the existing studies explores backdoor attacks against I2I networks. In contrast to adversarial attacks, the I2I backdoor attacks proposed in this work exhibit more severe security threats and can be used for positive applications (see Figure 10). This underscores the imperative need to investigate the vulnerability of I2I networks against backdoor attacks.

2.3 Backdoor Attacks against Image Generative Networks

Several works have explored backdoor attacks on generative models such as GAN [12], [13], [14] and diffusion model [15], [16], [17]. Concretely, *Salem et al.* [12] and *Rawat et al.* [13] proposed backdoor attacks against GANs, where they modified the loss functions of the generator and discriminator to make GAN output the backdoor target image for the triggered input image; *Jin et al.* [14] extended this backdoor attack in federated learning GAN; *Chou et al.* [16] and *Chen et al.* [15] embedded backdoor in diffusion models by manipulating the diffusion process.

Nevertheless, these backdoor methods cannot be applied and compared in our I2I backdoor attack, because I2I networks do not have a generator and discriminator and do not entail a diffusion process.

3 Threat Model

In this work, we consider a malicious I2I network provider, who has control of the training process of the victim I2I network. The adversary trains the backdoored I2I network and makes it accessible for users to download. For the I2I backdoor attack that is targeted at downstream image classification and object detection tasks, the attacker has no knowledge of the downstream classifier or detector. The I2I backdoor attack must satisfy the following requirements:

- **Normal-functionality**. The I2I backdoor must preserve the performance of the I2I network when processing clean input images. In the context of the I2I backdoor attack against I2I tasks, this requirement implies that the backdoored denoising/super-resolution network should output normal denoised/high-resolution images for clean input images. In the case of the I2I backdoor attack targeting downstream tasks, for clean input images and the backdoored upstream denoising model, the downstream classification/detection accuracy should be similar to that with the clean upstream denoising model.
- **Effectiveness.** For the I2I backdoor attack against I2I tasks, the backdoored I2I model should be capable of generating the backdoor target image when processing images with the backdoor trigger. In the context of the I2I backdoor attack against downstream classification/detection tasks, for backdoor-triggered input images and the backdoored upstream denoising model, the denoised images should provoke misclassification/misdetection by the downstream classification/detection model with a high probability.

4 I2I Backdoor Attack against I2I Tasks

In this section, we present the details of I2I backdoor attack against I2I tasks. The workflow is illustrated in Figure 3.

For ease of reference, we also provide the notations used in this work in TABLE 1.

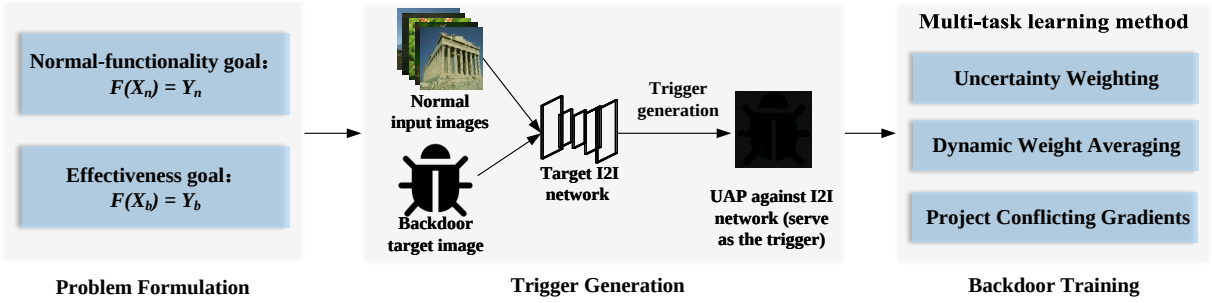


Fig. 3: The workflow of I2I backdoor attack.

TABLE 1: Notations and description

Notations	Description
X_n	the normal input image
X_b	the triggered input image
Y_n	the normal target image
Y_b	the backdoor target image
t	trigger for the I2I task
F	the targeted I2I model
S	a small set of normal images
s	update step size of the trigger
I	the maximal number of iterations
$\mathcal{L}_t$	loss function to optimize the trigger
$\mathcal{L}_m$	loss function of the main task
$\mathcal{L}_b$	loss function of the backdoor task
C	the downstream image classification model
u	trigger for the image classification task
D	the downstream object detection model

4.1 Problem Formulation

We denote X_n as the normal input image (i.e., the normal low-resolution/noisy image), Y_n as the normal output image (i.e., the high-resolution/noise-free image), X_b as the backdoor-triggered input image, Y_b as the backdoor target image[3], F as the target I2I network. According to the requirements described in Section 3, the goal of the I2I backdoor against I2I tasks can be formulated as:

$$\text{Normal-functionality goal: } F(X_n) = Y_n \tag{1}$$

$$\text{Effectiveness goal: } F(X_b) = Y_b \tag{2}$$

4.2 Backdoor Trigger

In backdoor attacks against classification models, numerous studies [24], [25], [26], [27], [28] have opted for the utilization of the targeted Universal Adversarial Perturbation (UAP) as the backdoor trigger. This is because the targeted UAP is able to push the classification results of triggered samples to the backdoor target class and is therefore more facilitating for backdoor embedding.

Inspired by this idea, we design a targeted UAP attack for I2I networks and use this UAP as the trigger for our I2I backdoor. As presented in Algorithm 1, for a small set of normal input images S, we iteratively pick one sample (X_i) from S and employs the gradient descent algorithm to minimize $\mathcal{L}_t$ for I rounds to optimize trigger t:

3. In this work, we choose a bug image as the backdoor target image, which is completely irrelevant to the input image.

$$\mathcal{L}_t = \|F(X_i + t) - Y_b\|_2 \tag{3}$$

The optimizing process is performed for all samples in S one by one and the final t is returned as the UAP trigger.

Algorithm 1 The Generation Algorithm of the UAP Trigger

Input: a small set of normal input images S; the victim I2I model F; the update step size of the trigger s; maximal number of iterations I; the backdoor target image Y_b; the range of the trigger $(-\epsilon_t, +\epsilon_t)$.
Output: the UAP trigger t

```
randomly initialize t;
for each sample X_i ∈ S do
    j ← 0 (iteration counter)
    while j <= I do
        Δ = ∂L_t / ∂t
        t ← t − s * sign(Δ), t ← clip(t, −ε_t, +ε_t)
        j ← j + 1
        Update L_t According to Eq.(3)
    end while
end for
return t
```

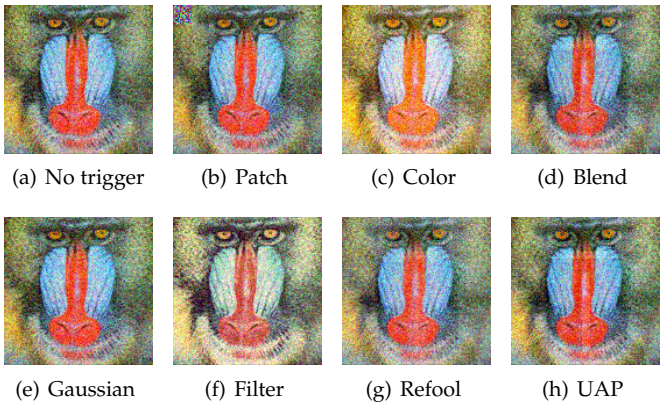
In our experiments, we also employ existing backdoor triggers for image classification tasks to perform our I2I backdoor attack, including patch trigger [29], blend trigger [30], refool trigger [31], color trigger [32], Instagram filter trigger [33] and Gaussian noise trigger [34]. Figure 4 illustrates the input noisy images with these backdoor triggers. The experimental results in Section 6.2.1 demonstrate the superiority of our designed UAP trigger.

4.3 Backdoor Training

4.3.1 *Backdoor Training with Multi-task Learning (MTL)*

After identifying the backdoor trigger pattern, the subsequent step is to embed the backdoor into the I2I model via the backdoor training process. In order to accomplish the dual objectives of ensuring normal-functionality and enhancing attack effectiveness simultaneously, we have devised two loss functions for the main task and the backdoor task. After that, we leverage the multi-task learning (MTL) framework to conduct the backdoor training process.

The main task is to satisfy the normal-functionality goal, i.e., the backdoored model is expected to perform normally



(a) No trigger (b) Patch (c) Color (d) Blend

(e) Gaussian (f) Filter (g) Refool (h) UAP

Fig. 4: Visual examples of input noisy images with/without trigger.

on normal input images. The loss function can be defined as:

$$\mathcal{L}_m = \|F(X_n) - Y_n\|_2 \tag{4}$$

The backdoor task is to achieve the attack effectiveness goal, i.e., the backdoored model is expected to output the backdoor target image for the backdoor-triggered input image. The loss function can be formulated as:

$$\mathcal{L}_b = \|F(X_b) - Y_b\|_2 \tag{5}$$

Therefore, the total loss for backdoor training can be formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_m + \mathcal{L}_b \tag{6}$$

4.3.2 Dynamic Weighting Methods

However, in the training process with Equation (6), the $\mathcal{L}_{total}$ is prone to be dominated by the task with a larger loss and fall into the local optimum, resulting in lower attack performance. This is attributed to the complicated mapping relationship in the I2I backdoor attack (backdoor-triggered images to the backdoor target image), making it difficult to balance the two tasks. Hence, we employ SOTA weighting methods, including Uncertainty Weighting (UW) [35], Dynamic Weight Averaging (DWA) [36] and Project Conflicting Gradients (PCGrad) [37] in the MTL process to avoid local optimum and accelerate convergence rates. Below we describe how to employ these weighting methods in our backdoor training process.

UW assigns larger weights to "easier" tasks, where it employs homoscedastic task uncertainty to balance different loss functions of different tasks. The L_{total} in this work can be formulated as:

$$\mathcal{L}_{total} = \frac{1}{2\sigma_m^2}\mathcal{L}_m + \frac{1}{2\sigma_b^2}\mathcal{L}_b + \log \sigma_m \sigma_b \tag{7}$$

where σ_m and σ_b represent the variance of $\mathcal{L}_m$ and $\mathcal{L}_b$. For the task with large uncertainty (i.e., large variance), the corresponding weights of its loss function are correspondingly reduced. The function of $\log \sigma_{m,b}$ is to prevent $\sigma_{m,b}$ from being too large.

DWA forces each task to learn at a similar rate. The weight of each task is formulated as follows:

$$w_i(t) = \frac{Ne^{(r_i(t-1)/T)}}{\sum_{n=1}^{N} e^{(r_n(t-1)/T)}}, r_i(t-1) = \frac{\mathcal{L}_i(t-1)}{\mathcal{L}_i(t-2)} \tag{8}$$

where $w_i(t)$ represents the weight of task i at step t, N represents the total number of the tasks, $r_n(t)$ is the ratio of the current loss to the previous loss, T is the temperature-scaling hyperparameter [38], which controls the softness of task weighting.

PCGrad is designed to address the challenging issue of gradient conflict. Specifically, during our backdoor training process, it is common that the gradients of the main task and the backdoor task exhibit some degree of conflict[4]. This conflict often results in sluggish convergence rates or diminished attack performance. For every training batch, PCGrad calculates the cosine similarity between the gradient of the main task, denoted as $\mathbf{g}_m$, and the gradient of the backdoor task, represented as $\mathbf{g}_b$. In cases where the gradients are not conflicted, they remain unaltered. When conflicts arise, PCGrad replaces $\mathbf{g}_b$ with its projection onto the normal plane of $\mathbf{g}_m$, as presented in Equation (9). This mechanism enhances the backdoor training process, mitigating gradient conflicts and fostering more efficient convergence and heightened attack performance.

$$\mathbf{g}_b = \mathbf{g}_b - \frac{\mathbf{g}_b \cdot \mathbf{g}_m}{\|\mathbf{g}_m\|^2}\mathbf{g}_m \tag{9}$$

In Section 6, we conduct extensive ablation studies to rigorously assess the performance of these dynamic weighting methods.

5 I2I Backdoor Attack that is Targeted at the Downstream Tasks

In addition to attacking I2I tasks, we further introduce an I2I backdoor attack that is targeted at the downstream image classification or object detection tasks. Specifically, we first embed the backdoor into the upstream image denoising model. Consequently, the denoised version of the backdoor-triggered image will be misclassified/misdetected by normal downstream classification/detection models with a high probability.

5.1 Targeted at Downstream Image Classification Task

According to the requirements described in Section 3, the normal-functionality and effectiveness goal of the I2I backdoor attack against downstream classification task can be formulated as:

$$\text{Normal-functionality goal: } C(F(X_n)) = C(Y_n) \tag{10}$$

$$\text{Effectiveness goal: } C(F(X_b)) \neq C(Y_n) \tag{11}$$

where C is the normal downstream image classification model.

For backdoor trigger types and backdoor training methods, we adopt the same attack configurations described

4. When the cosine similarity between the two gradients $\cos\theta < 0$, the two gradients are considered to be conflicted.

in Section 4.2 and 4.3. Differently, for the input backdoor-triggered image, we attach the classification UAP[5] to its noise-free version and use this image as the backdoor target image. Consequently, the denoised version of the input backdoor-triggered image will contain the classification UAP thereby leading to a misclassification.

Specifically, the generation algorithm of the classification UAP is presented in Algorithm 2. For each sample (X_i) in the dataset S, the algorithm first determines whether $X_i + u$ is able to cause the misclassification of the model C. If not, the algorithm performs an adversarial attack algorithm (such as DeepFool [39], PGD [40]) to optimize the u so that $X_i + u$ crosses the classification boundary. The optimizing process is conducted for all samples in S and the final u is returned as the classification UAP.

Algorithm 2 The Generation Algorithm of the Classification UAP

Input: a small set of normal input images S; a clean pre-trained downstream classification model C; the range of the classification UAP $(-\epsilon_u, +\epsilon_u)$.
Output: the UAP for image classification u

initialize $u \leftarrow 0$;
for each sample $X_i \in S$ **do**
  if $C(X_i + u) = C(X_i)$ **then**
    Compute the minimal perturbation that sends $X_i + u$ to the decision boundary:
    $\Delta u_i = \arg\min_r \|r\|_2,\ \text{s.t. } C(X_i + u + r) \neq C(X_i)$
    Update the perturbation:
    $u \leftarrow u + \Delta u_i,\ u \leftarrow clip(u, -\epsilon_u, +\epsilon_u)$
  end if
end for
return u

5.2 Targeted at Downstream Object Detection Task

Similarly, the normal-functionality and effectiveness goal of the I2I backdoor attack against downstream detection task can be defined as Equation (12) and (13), respectively.

$$\text{Normal-functionality goal: } D(F(X_n)) = D(Y_n) \tag{12}$$

$$\text{Effectiveness goal: } D(F(X_b)) \neq D(Y_n) \tag{13}$$

where D is the normal downstream object detection model.

We also employ the same backdoor trigger types and backdoor training methods described in Section 4.2 and 4.3. For the backdoor target image, we first adopt the existing universal adversarial attack against object detection [41] to generate the UAP[6] for the object detection task. After that, for the backdoor-triggered image, we attach the detection UAP to its noise-free image and use this image as the backdoor target image.

It is worth noting that the attacker does not need to have knowledge of the downstream classification/detection model. It can utilize surrogate models to execute the UAP generation algorithm. Due to the transferability of UAP, images with the UAP are able to induce misclassifications/misdetections of other clean classification/detection models.

5. The classification UAP is designed to induce misclassifications of classification models, which is different from the UAP against I2I networks in Section 4.2.

6. The detection UAP is designed to fabricate additional wrong detections (i.e., adding false positives).

6 Evaluation

We perform extensive experiments over different datasets and I2I networks to evaluate the performance of our I2I backdoor attacks. All experiments are implemented in Python and run on a NVIDIA RTX A6000.

6.1 Experimental Setup

6.1.1 Model Architecture

- **I2I backdoor against I2I tasks:** this work considers the two most commonly used I2I tasks (image denoising and image super-resolution) as examples to investigate the vulnerability of I2I networks to backdoor attacks. For the two tasks, we have selected several state-of-the-art (SOTA) I2I network architectures, including SCUNet [2], MPRNet [7], MIRNet [21], DPIR [19] and ESRGAN [18], for experimental evaluations. We firmly believe that other I2I tasks and I2I network architectures are also susceptible to the I2I backdoor attacks in this work.
- **I2I backdoor against downstream tasks:** in the context of the I2I backdoor that targets at the downstream classification/detection task, we employ the aforementioned image denoising networks to conduct the upstream image denoising task. For the downstream classification task, we use the pre-trained ResNet50, VGG19 and MobileNetv2 model to perform image classification; for the downstream detection task, we use the pre-trained MobileNet-YOLOv3, EfficientNet-YOLOv3 and Darknet53-YOLOv3 model to perform object detection.

6.1.2 Datasets

- **Image denoising task:** following previous works [42], [43], [44], we use Color400 as the training data, and CSet8 as the testing data.
- **Image super-resolution task:** we choose BSD100 [45] as the training data, and Set14 [46] as the testing data.
- **Downstream image classification task:** we evaluate our I2I backdoor against the downstream image classification task on the ImageNet-1k [47] dataset.
- **Downstream object detection task:** we evaluate our I2I backdoor against the downstream object detection task on the Pascal VOC dataset [48].

6.1.3 Attack Configuration

- **Baseline trigger settings:** (1) Patch backdoor trigger: following previous work [29], we employ a 16*16 patch in the top left corner of the input 128*128 image as the backdoor trigger; (2) Gaussian backdoor trigger: following previous work [34], we employ a specific Gaussian noise as the backdoor trigger. The mean is set to 0, the standard deviation is set to 20/255; (3) Color backdoor trigger: we follow the attack configuration proposed in [32] and use a specific shift in the color space as the backdoor trigger; (4) Blend backdoor trigger: we follow the attack configuration proposed in [30] and use the backdoor target image as the blend image. The blending ratio is set to 0.1; (5)

Refool backdoor trigger: we follow the attack configuration proposed in [31] and use the backdoor target image as the reflection trigger. The reflection ratio is set to 0.1; (6) Filter backdoor trigger: we follow the attack configuration proposed in [33] and use a specific Instagram filter as the backdoor trigger.

- **UAP trigger settings:** the number of normal images in D is set to 10, the update step size of the trigger s is set to $5/255$, the maximum number of iterations I is set to 20, the range of the trigger is set to (-15/255, +15/255).
- **Backdoor training process settings:** we follow the hyper-parameter settings in UW [35], DWA [36] and PCGrad [37], and train the backdoor model with the Adam optimizer (learning rate equals to 0.0001). Besides, we also introduce a static weighting (SW) approach for backdoor training as a baseline comparison, where the weight of the main task is set to 0.9 and the weight of the backdoor task is set to 0.1.

6.1.4 Evaluation Metrics

- **I2I backdoor against I2I tasks:** we employ the SSIM [49] between images to measure the attack performance. Specifically, for the backdoored I2I model, we calculate the SSIM between the denoised/super-resolved result for clean input image and the ground truth image (i.e., noise-free image or high-resolution image) to evaluate the normal-functionality; we calculate the SSIM between the denoised/super-resolved result for triggered image and the backdoor target image to evaluate the attack effectiveness.
- **I2I backdoor against downstream tasks:** for the backdoored upstream image denoising model, we calculate the test accuracy and mean Average Precision (mAP) of the downstream classification and detection model on the denoised results for normal input images to measure the normal-functionality, respectively; we calculate the attack success rate (ASR) of the denoised results for triggered images on the downstream model to evaluate the attack effectiveness.
- **Stealthiness of triggered images:** to evaluate the stealthiness of triggered images, we employ LPIPS [50] distance as the metric, which is more consistent with human vision system. Concretely, LPIPS evaluates the perceptual distance between two images by means of a deep learning model. LPIPS suggests that even if two images are very close to each other at the pixel level, a human observer may perceive them as different. Therefore, LPIPS uses pre-trained deep networks (e.g., VGG, AlexNet) to extract image features[7], and then calculates the distance between these features to evaluate the perceptual similarity between images. It is formulated as Equation (14):

$$lpips(x,\hat{x}) = \sum_{l} \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \left\| \mathbf{w}_l \odot \left(\hat{\mathbf{f}}_l^{x}(h,w) - \hat{\mathbf{f}}_l^{\hat{x}}(h,w) \right) \right\|_2^2 \tag{14}$$

where $\hat{\mathbf{f}}_l^{x}(h,w)$ represents the features extracted from image x at layer l; $\mathbf{w}_l$ represents the channel weight of layer l; H_l, W_l represent the height and width of the feature map in layer l; $\odot$ denotes the channel-by-channel multiplication and $\|\cdot\|_2^2$ denotes the square of the Euclidean distance. A smaller LPIPS value means that the two images are more similar in terms of human visual perception.

7. In this work, we employ the pre-trained VGG network for experimental evaluations.

6.2 Attack Performance Evaluation

6.2.1 Main Results

We have conducted extensive experiments of I2I backdoor attacks with different backdoor triggers and MTL methods on various I2I network architectures. As presented in Table 2, most triggers achieve high attack effectiveness in attacking image denoising task. However, they fail to preserve the similar normal-functionality with that of the clean model. In comparison, the UAP trigger is superior to other triggers in maintaining normal-functionality. As provided in Table 3, only the UAP trigger achieves good attack performance on all these I2I models. In addition, we also calculate the sum of the normal-functionality and the attack effectiveness to see which trigger achieves a better balance of preserving normal-functionality and enhancing attack effectiveness. The results show that the UAP trigger always achieves the highest total score on all considered I2I network architectures.

6.2.2 Computational Overhead

We have assessed the computational overhead of generating the UAP trigger. As provided in Table 4, the computational overhead of the UAP trigger generation algorithm is relatively small and falls within acceptable bounds for potential backdoor attackers. Hence, in the subsequent experiments, we use the UAP trigger to perform our I2I backdoor attack.

TABLE 4: Computational overhead (s) for the UAP trigger generation.

DPIR	SCUNet	MPRNet	MIRNet	ESRGAN
18.08	49.54	52.22	128.13	57.64

Furthermore, we have also evaluated the computational overhead of different MTL methods. As outlined in Table 5, the difference between the computational overhead of these MTL methods is relatively negligible. Therefore, without loss of generality, we have opted to employ the PCGrad method for MTL in the subsequent experiments.

TABLE 5: Computational overhead (s) for different MTL methods (1 epoch).

Architecture \ MTL method	SW	UW	DWA	PCGrad
DPIR	13.64	8.93	14.15	16.58
SCUNet	31.84	25.95	30.15	39.01
MPRNet	35.39	36.91	33.66	43.52
MIRNet	93.04	85.09	77.40	85.43
ESRGAN	54.72	56.11	60.47	71.86

TABLE 2: The performance of I2I backdoor with different triggers and MTL methods on image denoising task.

Architecture	MTL method	SSIM*	Trigger type							
			None	Gaussian	Color	Filter	Patch	Blend	Refool	UAP
DPIR	SW	Normal.	0.8936	0.8690	0.8872	0.7816	0.7440	0.8914	0.8948	0.8961
		Effect.	\	0.9949	0.9964	0.8723	0.9992	0.9437	0.9938	0.9936
		Total	\	1.8649	1.8836	1.6539	1.7432	1.8351	1.8886	**1.8897**
	DWA	Normal.	\	0.8660	0.7062	0.6351	0.7110	0.8766	0.8170	0.8855
		Effect.	\	0.9939	0.9786	0.8675	0.9995	0.9925	0.9158	0.9841
		Total	\	1.8599	1.6848	1.5026	1.7105	1.8691	1.7328	**1.8696**
	UW	Normal.	\	0.8661	0.8288	0.5853	0.6957	0.8821	0.8821	0.8850
		Effect.	\	0.9974	0.9928	0.9799	0.9989	0.9767	0.9350	0.9898
		Total	\	1.8635	1.8216	1.5652	1.6946	1.8588	1.8171	**1.8748**
	PCGrad	Normal.	\	0.8649	0.7953	0.6800	0.7372	0.8832	0.8863	0.8939
		Effect.	\	0.9970	0.9900	0.9839	0.9996	0.9302	0.9851	0.9991
		Total	\	1.8619	1.7853	1.6639	1.7368	1.8134	1.8714	**1.8930**
SCUNet	SW	Normal.	0.8839	0.8688	0.7713	0.7561	0.7534	0.8827	0.8746	0.8834
		Effect.	\	0.9925	0.9906	0.8977	0.9414	0.9499	0.8268	0.9872
		Total	\	1.8613	1.7619	1.6538	1.6948	1.8326	1.7014	**1.8705**
	DWA	Normal.	\	0.8778	0.7360	0.8346	0.8228	0.8492	0.7948	0.8803
		Effect.	\	0.9988	0.9901	0.9725	0.9927	0.8842	0.8529	0.9988
		Total	\	1.8766	1.7261	1.8071	1.8155	1.7334	1.6477	**1.8791**
	UW	Normal.	\	0.8623	0.7536	0.7269	0.8087	0.8727	0.8612	0.8750
		Effect.	\	0.9988	0.9996	0.9797	0.9965	0.9745	0.9848	0.9980
		Total	\	1.8611	1.7532	1.7066	1.8052	1.8472	1.8460	**1.8730**
	PCGrad	Normal.	\	0.8752	0.7511	0.6722	0.7667	0.8759	0.8371	0.8753
		Effect.	\	0.9990	0.9994	0.9849	0.9910	0.9093	0.9230	0.9986
		Total	\	**1.8742**	1.7505	1.6571	1.7577	1.7852	1.7601	1.8739
MPRNet	SW	Normal.	0.9081	0.7066	0.8865	0.7073	0.7822	0.8638	0.8938	0.9008
		Effect.	\	0.9704	0.8790	0.8729	0.7385	0.9894	0.9884	0.9977
		Total	\	1.6770	1.7655	1.5802	1.5207	1.8532	1.8822	**1.8985**
	DWA	Normal.	\	0.6337	0.7701	0.6546	0.7410	0.8890	0.8145	0.8721
		Effect.	\	0.9361	0.9322	0.9022	0.9676	0.9970	0.9926	0.9871
		Total	\	1.5698	1.7023	1.5568	1.7086	1.8860	1.8071	**1.8592**
	UW	Normal.	\	0.7354	0.7594	0.7256	0.7374	0.8867	0.8829	0.9079
		Effect.	\	0.9978	0.8723	0.9271	0.9943	0.9953	0.9964	0.9997
		Total	\	1.7332	1.6317	1.6527	1.7317	1.8820	1.8793	**1.9076**
	PCGrad	Normal.	\	0.7240	0.7430	0.7390	0.7452	0.8853	0.8831	0.9151
		Effect.	\	0.9963	0.8587	0.9305	0.9884	0.9944	0.9957	0.9995
		Total	\	1.7203	1.6017	1.6695	1.7336	1.8797	1.8788	**1.9146**
MIRNet	SW	Normal.	0.9172	0.6887	0.8163	0.7797	0.8546	0.8957	0.8951	0.9139
		Effect.	\	0.9983	0.9975	0.9534	0.9727	0.9779	0.9841	0.9964
		Total	\	1.6870	1.8138	1.7331	1.8273	1.8736	1.8792	**1.9102**
	DWA	Normal.	\	0.7025	0.7464	0.8010	0.8147	0.8382	0.8220	0.8645
		Effect.	\	0.9921	0.9516	0.9033	0.9892	0.9497	0.9845	0.9939
		Total	\	1.6946	1.6980	1.7043	1.8039	1.7879	1.8066	**1.8585**
	UW	Normal.	\	0.7205	0.8201	0.8101	0.8496	0.8872	0.8839	0.9001
		Effect.	\	0.9811	0.9920	0.9720	0.9956	0.9905	0.9657	0.9920
		Total	\	1.7016	1.8121	1.7821	1.8452	1.8777	1.8496	**1.8921**
	PCGrad	Normal.	\	0.6991	0.8363	0.8325	0.8575	0.8944	0.8997	0.9060
		Effect.	\	0.9954	0.9719	0.9819	0.9939	0.9823	0.9660	0.9994
		Total	\	1.6945	1.8082	1.8144	1.8514	1.8767	1.8675	**1.9054**
ESRGAN	SW	Normal.	0.9112	0.6009	0.7739	0.8056	0.6497	0.8726	0.8667	0.9146
		Effect.	\	0.9733	0.9713	0.9345	0.9969	0.8121	0.9330	0.9925
		Total	\	1.5742	1.7452	1.7401	1.6466	1.6847	1.7997	**1.9071**
	DWA	Normal.	\	0.5565	0.7205	0.6585	0.5707	0.7926	0.8518	0.9073
		Effect.	\	0.9745	0.9903	0.7642	0.9569	0.7870	0.8895	0.9470
		Total	\	1.5310	1.7108	1.4227	1.5276	1.5796	1.7413	**1.8544**
	UW	Normal.	\	0.6037	0.7944	0.6443	0.6220	0.8579	0.8715	0.8962
		Effect.	\	0.9956	0.9986	0.9772	0.9886	0.9807	0.9881	0.9985
		Total	\	1.5993	1.7930	1.6215	1.6106	1.8386	1.8596	**1.8948**
	PCGrad	Normal.	\	0.5912	0.7892	0.7715	0.4988	0.8679	0.8807	0.9086
		Effect.	\	0.9968	0.9990	0.9872	0.9729	0.9767	0.9891	0.9992
		Total	\	1.5880	1.7882	1.7587	1.4717	1.8446	1.8698	**1.9078**

* Normal. denotes the normal-functionality; Effect. denotes the effectiveness; Total represents the sum of them. The bolded results represent the maximum total score.

TABLE 3: The performance of I2I backdoor with different triggers and MTL methods on image super-resolution task.

Architecture	MTL method	SSIM*	Trigger type							
			None	Gaussian	Color	Filter	Patch	Blend	Refool	UAP
DPIR	SW	Normal.	0.8381	0.7915	0.7812	0.6328	0.7320	0.7259	0.7531	0.7920
		Effect.	\	0.9475	0.4166	0.5466	0.9864	0.9618	0.6862	0.9631
		Total	\	1.7390	1.1978	1.1794	1.7184	1.6877	1.4393	**1.7551**
	DWA	Normal.	\	0.7884	0.5442	0.6739	0.7937	0.7641	0.7517	0.7866
		Effect.	\	0.9954	0.7099	0.5146	0.6529	0.8448	0.8982	0.9763
		Total	\	**1.7838**	1.2541	1.1885	1.4466	1.6089	1.6499	1.7629
	UW	Normal.	\	0.7925	0.6956	0.7066	0.7694	0.7870	0.8154	0.8308
		Effect.	\	0.9953	0.5773	0.6134	0.9093	0.9669	0.9228	0.9887
		Total	\	1.7878	1.2729	1.3200	1.6787	1.7539	1.7382	**1.8195**
	PCGrad	Normal.	\	0.7966	0.7428	0.6616	0.7818	0.7933	0.8218	0.8201
		Effect.	\	0.9938	0.5981	0.5387	0.8456	0.9734	0.9143	0.9862
		Total	\	1.7904	1.3409	1.2003	1.6274	1.7667	1.7361	**1.8063**
SCUNet	SW	Normal.	0.8492	0.7912	0.8082	0.7738	0.8092	0.6438	0.7876	0.8476
		Effect.	\	0.8678	0.6631	0.6199	0.8243	0.8059	0.5305	0.8615
		Total	\	1.6590	1.4713	1.3937	1.6335	1.4497	1.3181	**1.7091**
	DWA	Normal.	\	0.7367	0.6284	0.8262	0.8227	0.7201	0.7818	0.8227
		Effect.	\	0.9594	0.8914	0.7317	0.7255	0.8473	0.7597	0.8971
		Total	\	1.6961	1.5198	1.5579	1.5482	1.5673	1.5415	**1.7198**
	UW	Normal.	\	0.7445	0.7239	0.6484	0.7798	0.6958	0.7604	0.8285
		Effect.	\	0.9717	0.9933	0.8445	0.8954	0.8480	0.7703	0.9075
		Total	\	1.7162	1.7172	1.4929	1.6752	1.5438	1.5307	**1.7360**
	PCGrad	Normal.	\	0.7520	0.7202	0.6370	0.7921	0.7341	0.6704	0.8357
		Effect.	\	0.9379	0.9930	0.8918	0.8272	0.8892	0.8982	0.8750
		Total	\	1.6899	**1.7132**	1.5288	1.6193	1.6233	1.5686	1.7107
MPRNet	SW	Normal.	0.8737	0.8536	0.7683	0.7467	0.7401	0.7030	0.7890	0.8732
		Effect.	\	0.9855	0.3369	0.2216	0.4000	0.5619	0.4922	0.9736
		Total	\	1.8391	1.1052	0.9683	1.1401	1.2649	1.2812	**1.8468**
	DWA	Normal.	\	0.8019	0.8167	0.6647	0.8696	0.7529	0.8431	0.8729
		Effect.	\	0.9638	0.2939	0.4193	0.5048	0.7806	0.7711	0.9832
		Total	\	1.7657	1.1106	1.0840	1.3744	1.5335	1.6142	**1.8561**
	UW	Normal.	\	0.8085	0.8494	0.7961	0.8631	0.7705	0.8212	0.8745
		Effect.	\	0.9900	0.3532	0.2410	0.5041	0.6751	0.7973	0.9910
		Total	\	1.7985	1.2026	1.0371	1.3672	1.4456	1.6185	**1.8655**
	PCGrad	Normal.	\	0.8085	0.8341	0.7659	0.8715	0.6517	0.8418	0.8719
		Effect.	\	0.9779	0.3902	0.2305	0.4988	0.6036	0.7459	0.9879
		Total	\	1.7864	1.2243	0.9964	1.3703	1.2553	1.5877	**1.8598**
MIRNet	SW	Normal.	0.8673	0.7423	0.6467	0.7464	0.8700	0.7459	0.6828	0.8664
		Effect.	\	0.9497	0.8711	0.5154	0.9951	0.2901	0.5696	0.9844
		Total	\	1.6920	1.5178	1.2618	1.8651	1.0360	1.2524	**1.8508**
	DWA	Normal.	\	0.8688	0.5337	0.6440	0.8668	0.7802	0.6921	0.8646
		Effect.	\	0.9779	0.7573	0.9011	0.9961	0.7827	0.4265	0.9968
		Total	\	**1.8667**	1.2910	1.5451	1.8629	1.5629	1.1186	1.8614
	UW	Normal.	\	0.7312	0.8404	0.7889	0.8692	0.6879	0.7393	0.8705
		Effect.	\	0.9793	0.8187	0.7569	0.9946	0.5735	0.9041	0.9990
		Total	\	1.7105	1.6591	1.5458	1.8638	1.2614	1.6434	**1.8695**
	PCGrad	Normal.	\	0.8047	0.8320	0.6444	0.8648	0.6860	0.6751	0.8692
		Effect.	\	0.9706	0.8689	0.9236	0.9956	0.3490	0.8780	0.9949
		Total	\	1.7753	1.7009	1.5680	1.8604	1.0350	1.5531	**1.8641**
ESRGAN	SW	Normal.	0.8650	0.8735	0.8186	0.8005	0.8718	0.8535	0.8228	0.8713
		Effect.	\	0.9818	0.3929	0.6037	0.2934	0.4067	0.3409	0.9950
		Total	\	1.8553	1.2115	1.4042	1.1652	1.2602	1.1637	**1.8663**
	DWA	Normal.	\	0.8719	0.7592	0.6913	0.8182	0.8277	0.8032	0.8690
		Effect.	\	0.9941	0.6040	0.5489	0.3594	0.4416	0.3585	0.9971
		Total	\	1.8660	1.3632	1.2402	1.1776	1.2693	1.1617	**1.8661**
	UW	Normal.	\	0.8656	0.5908	0.8433	0.8640	0.8349	0.8314	0.8686
		Effect.	\	0.9946	0.8538	0.6614	0.1596	0.7015	0.4340	0.9936
		Total	\	1.8602	1.4446	1.5047	1.0236	1.5364	1.2654	**1.8622**
	PCGrad	Normal.	\	0.8752	0.7274	0.8112	0.8546	0.8460	0.8113	0.8728
		Effect.	\	0.9907	0.7271	0.7494	0.2688	0.5862	0.4314	0.9969
		Total	\	1.8659	1.4545	1.5606	1.1234	1.4322	1.2427	**1.8697**

* Normal. denotes the normal-functionality; Effect. denotes the effectiveness; Total represents the sum of them. The bolded results represent the maximum total score.

6.3 Ablation Study

6.3.1 Ablation Study of the Loss Function

In this section, instead of using SSIM metric and MSE loss, we employ the perceptual loss function and use LPIPS [50] as the metric. LPIPS uses pre-trained deep networks to extract image features and then calculates the distance between these features to evaluate the perceptual similarity between images. Specifically, we replace the main task loss and backdoor task loss with the perceptual loss functions based on LPIPS metric: $\mathcal{L}_m = lpips(F(X_n), Y_n)$, $\mathcal{L}_b = lpips(F(X_b), Y_b)$. The $lpips(x, \hat{x})$.

After that, the UAP trigger and PCGrad MTL method are selected as an example[8] to evaluate the attack performance across different architectures. As presented in Table 6, the "Normal." of clean/backdoor model denotes the LPIPS distance of the denoised (or super-resolution) image and the original image, representing the normal-functionality of the clean/backdoor model; the "Effect." of backdoor model denotes the LPIPS distance of the denoised (or super-resolution) version of the triggered image and the backdoor target image, representing the attack effectiveness of the backdoor model. It can be observed that I2I backdoor does not cause significant degradation in the normal-functionality of the backdoor model. Besides, it also achieves high attack effectiveness, i.e., the denoised (or super-resolution) version of the triggered image is extremely similar to the backdoor target image. Therefore, it can be concluded that I2I backdoor is highly generalisable and can achieve great attack performance for different loss functions and similarity metrics.

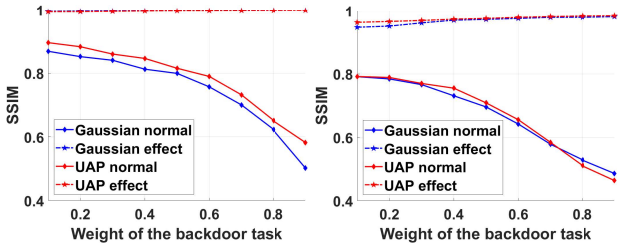
TABLE 6: The attack performance of I2I backdoor with the perceptual loss functions based on LPIPS metric.

Task	Architecture	Clean model		Backdoor model	
		Normal.	Effect.	Normal.	Effect.
Denoise	DPIR	0.0217	\	0.0228	0.0029
	SCUNet	0.0261	\	0.0270	0.0070
	MPRNet	0.0289	\	0.0295	0.0066
	MIRNet	0.0252	\	0.0258	0.0019
	ESRGAN	0.0318	\	0.0335	0.0025
SR	DPIR	0.0752	\	0.0770	0.0051
	SCUNet	0.0912	\	0.0933	0.0092
	MPRNet	0.0809	\	0.0838	0.0083
	MIRNet	0.0834	\	0.0882	0.0045
	ESRGAN	0.0887	\	0.0901	0.0058

6.3.2 Impact of Different Weights on Attack Performance

In this section, we conduct experiments to evaluate the attack performance under different weights for main task and backdoor task. Since the weights assigned by dynamic weighting methods (UW, DWA, and PCGrad) are constantly changing, we mainly consider the static weighting method. Specifically, we select the Gaussian and UAP trigger on DPIR model as an example, set the weight of the backdoor task from 0.1 to 0.9 and present the experimental results in Figure 5. It can be seen that as the weight of the backdoor task increases, the model normal-functionality decreases dramatically; and the increase in the weight of the backdoor task has no significant improvement in attack effectiveness.

8. Other cases yield the same experimental results.



(a) Denoise task (b) Super-resolution task

Fig. 5: The attack performance under different weights.

Therefore, in our default settings, we set the weight of the main task to 0.9 and the weight of the backdoor task to 0.1 in the static weighting method.

6.3.3 Impact of Different MTL Weighting Methods on Convergence Rates

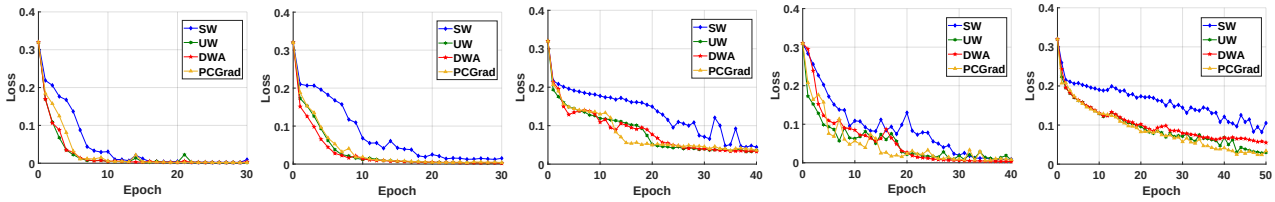
We further carry out thorough ablation studies focused on the impact of different MTL weighting methods on the convergence rate. As illustrated in Figure 6, it can be observed that the dynamic weighting methods, including UW, DWA, and PCGrad, always outperform the static weighting method in terms of convergence rates. This phenomenon can be attributed to the inherent complexity of the I2I backdoor task, which involves mapping a triggered image to an entirely unrelated backdoor target image. Such complexity invariably leads to conflicts with the main task. The static weighting method tends to struggle to achieve an optimal balance between these competing tasks, resulting in reduced training efficiency. Hence, the dynamic weight methods emerge as the more sensible choice for facilitating the I2I backdoor training process.

6.4 Stealthiness Evaluation

In addition to evaluating the effect of backdoor injection on the decrease of model normal-functionality, we further calculated the LPIPS distance of the backdoor-triggered images from the original images to assess the stealthiness of the attack. It can be seen from Table 7 that the backdoor-triggered images maintain high stealthiness, the LPIPS distances are all below 0.08. The patch trigger is the most stealthy trigger type with the metric of LPIPS. This may be because LPIPS is less sensitive to local, meaningless perturbations.

TABLE 7: Stealthiness evaluation of different trigger types and testing datasets.

Trigger \ Dataset	CSet8	Set14	ImageNet	Pascal VOC
Gaussian	0.0568	0.0506	0.0714	0.0785
Color	0.0087	0.0257	0.0229	0.0198
Filter	0.0195	0.0381	0.0458	0.0316
Patch	0.0071	0.0041	0.0040	0.0051
Blend	0.0328	0.0107	0.0181	0.0202
Refool	0.0134	0.0327	0.0427	0.0294
UAP	0.0317	0.0342	0.0413	0.0280



(a) DPIR (b) SCUNet (c) MPRNet (d) MIRNet (e) ESRGAN

Fig. 6: The convergence rates of the training loss with different MTL methods.

TABLE 8: Performance of I2I backdoor attack under model fine-tuning.

Epoch	SSIM	DPIR	SCUNet	MPRNet	MIRNet	ESRGAN
10	Normal.	0.8939	0.8855	0.9145	0.9049	0.9038
	Effect.	0.9990	0.9989	0.9995	0.9994	0.9978
20	Normal.	0.8935	0.8850	0.9156	0.9056	0.9039
	Effect.	0.9991	0.9981	0.9994	0.9994	0.9987
30	Normal.	0.8940	0.8857	0.9162	0.9052	0.9036
	Effect.	0.9987	0.9986	0.9992	0.9983	0.9983
40	Normal.	0.8944	0.8853	0.9160	0.9056	0.9034
	Effect.	0.9989	0.9992	0.9988	0.9989	0.9987
50	Normal.	0.8948	0.8851	0.9161	0.9058	0.9037
	Effect.	0.9985	0.9980	0.9990	0.9994	0.9985

6.5 Robustness Evaluation

In this section, we turn our attention to the robustness evaluation of the I2I backdoor attack against various defense methods. It should be pointed out that many backdoor defense techniques are designed for neural network classifiers, such as Neural Cleanse [51], STRIP [52], and Spectral Signature [53], they are not directly applicable to our I2I backdoor attacks. Thus, we have selected three defense methods, including bit depth reduction [54], image compression [55] and model fine-tuning to evaluate the robustness of the I2I backdoor attacks.

Bit depth reduction. We reduce the bit depth of input images before sending them to I2I models. As illustrated in Figure 7, the effectiveness of the attack consistently maintains a high level as the bit depth decreases. It demonstrates that the preprocessing of bit depth reduction is ineffective in mitigating our I2I backdoor attack.

Image compression. We compress input images before sending them to I2I models. As depicted in Figure 8, the degradation in normal-functionality consistently outweighs the degradation in attack effectiveness as input images undergo image compression. Thus, the preprocessing of image compression is also far from an effective defense method against the proposed I2I backdoor attack.

Model fine-tuning. We assume that the defender has a small amount[9] of clean images and uses these images to fine-tune the backdoored I2I model. As presented in Table 8, the I2I backdoor remains effective after fine-tuning with clean images.

STRIP. We extend the representative backdoor detection method STRIP [52] to I2I tasks and evaluate the robustness of the I2I backdoor under this defense. STRIP is a testing-time backdoor detection method that detects whether a testing image contains a trigger. The idea of STRIP is based on the assumption that the backdoor trigger is robust and can remain effective when a triggered image is superimposed by a clean image. In our work, we design the STRIP method for the I2I task following the framework of the STRIP method in the image classification task. Specifically, we sample 400 clean images and superimpose the target image with these clean images separately, and send them to I2I models to obtain output images. After that, we calculate the LPIPS distances between the output image corresponding to each superimposed image and the output image corresponding to the original target image. If the target image is backdoor-triggered, the output images will show high similarity. Conversely, if the target image is clean, the output images will show low similarity. Hence, we select the UAP trigger on DPIR denoise model as an example, and calculate and compare the LPIPS distance distributions of the clean and triggered image. As illustrated in Figure 9, we observe that the clean and triggered target image have very similar LPIPS distance distributions so that STRIP is not able to distinguish whether a testing image contains the backdoor trigger or not. This is mainly because the UAP trigger is destroyed when the triggered image is superimposed with a clean image, the backdoor trigger becomes ineffective in the superimposed image.

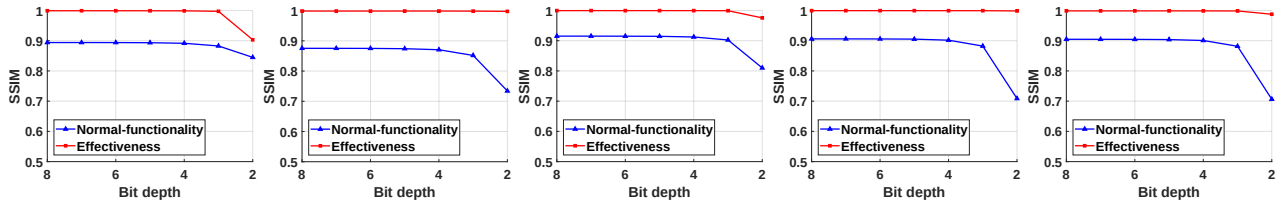
9. In our experiments, this amount is assumed to be 10% of the original training dataset.

6.6 Evaluation on I2I Backdoor Attack against Downstream Classification Task

To perform the I2I backdoor attack against the downstream classification task, we employ the Algorithm 2 to generate the UAP against the pre-trained ResNet152 classifier. After that, we employ this UAP to embed the I2I backdoor attack into the upstream image denoising model. Finally, we evaluate the attack performance on other clean pre-trained classifiers (including ResNet50, VGG19 and MobileNetV2) to measure the attack transferability.

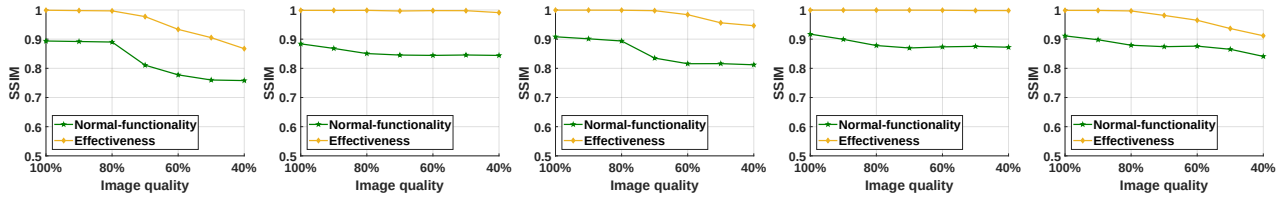
In the case of the I2I backdoor attack against the downstream object detection task, we employ the UAP generation algorithm for object detection [41] to construct the UAP against the pre-trained MobileNetv1-YOLOv3 detector. After that, we employ this UAP to embed the I2I backdoor attack into the upstream image denoising model. Finally, we evaluate the attack performance on other clean pre-trained object detectors (including MobileNetv2-YOLOv3, Darknet53-YOLOv3 and EfficientNet-YOLOv3) to measure the attack transferability.

This article has been accepted for publication in IEEE Transactions on Dependable and Secure Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2025.3603639



(a) DPIR (b) SCUNet (c) MPRNet (d) MIRNet (e) ESRGAN

Fig. 7: The performance of I2I backdoor attack under bit depth reduction.



(a) DPIR (b) SCUNet (c) MPRNet (d) MIRNet (e) ESRGAN

Fig. 8: The performance of I2I backdoor attack under image compression.

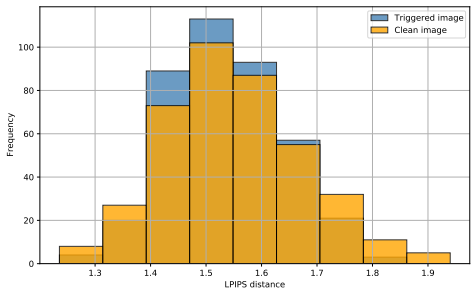


Fig. 9: Robustness of I2I backdoor under STRIP.

TABLE 9: The performance of I2I backdoor attack against downstream classification task (with the UAP against ResNet152 classifier).

Upstream denoising model D	Downstream classification model	Denoised accuracy (%)		ASR (%)
		Clean D Clean img.	Backdoor D Clean img.	Backdoor D Backdoor img.
DPIR	ResNet50	72.08	71.48	72.48
	VGG19	65.32	65.42	85.90
	MobileNetV2	64.40	64.68	74.90
SCUNet	ResNet50	71.72	71.56	72.64
	VGG19	65.06	64.26	80.96
	MobileNetV2	65.66	65.20	74.74
MPRNet	ResNet50	71.34	71.22	72.82
	VGG19	64.62	64.54	81.14
	MobileNetV2	64.32	64.66	74.92
MIRNet	ResNet50	71.64	71.40	72.74
	VGG19	65.30	63.88	80.72
	MobileNetV2	65.04	64.34	75.12
ESRGAN	ResNet50	71.16	69.80	72.78
	VGG19	64.42	63.36	81.48
	MobileNetV2	64.22	62.92	75.56

As presented in Table 9 and 10, for clean input images, the downstream denoised accuracy/mAP of the backdoor denoising model and the normal denoising model exhibit minimal disparity. This confirms that the I2I backdoor does not affect the normal-functionality of the downstream classification/detection task. In the case of backdoor-triggered input images and the backdoor upstream denoising model, the denoised versions of these images can fool the downstream clean pre-trained classifiers/detectors with high success rates. This proves the attack effectiveness of the I2I backdoor attack against the downstream classification/detection task.

In addition, we also evaluate the attack performance of our I2I backdoor attack against adversarially trained downstream models. Specifically, for downstream classification models, we employ a *PGD* adversarial training strategy during the training phase to obtain the adversarially trained model. For downstream detection models, we follow the work [56] to perform *PGD* adversarial training in the object detection task. The *PGD* step is also fixed to 10, and the maximum perturbation is set to $4/255$.

As illustrated in Table 11, the adversarial training strategy does show some degree of mitigation against I2I backdoor attacks. Concretely, adversarial training shows more significant attack mitigation effect on the downstream classification task, but is less effective in defending the downstream detection task. This may be due to the fact that the detection task is more complex than the classification task, and the UAP targeting the detection task is more likely to have higher generalizability and transferability. In addition, adversarial training will result in the degradation of the model normal-functionality to some degree (about 2.8% accuracy loss for the classification task and about 3.1% mAP loss for the detection task). Hence, it can be concluded that adversarial training can only mitigate the attack effect to some extent but can not completely defend against the I2I backdoor attack.

6.7 Harnessing I2I Backdoor for Positive Applications

It's noteworthy that the potential of I2I backdoor attacks can extend beyond malicious intent, finding utility in ethical

TABLE 10: The performance of I2I backdoor attack against downstream detection task (with the UAP against MobileNetv1-YOLOv3).

Upstream denoising model D	Downstream detection model	mAP (%) Clean D Clean img.	mAP (%) Backdoor D Clean img.	ASR (%) Backdoor D Backdoor img.
DPIR	MobileNetv2-YOLOv3	68.21	66.94	81.17
	Darknet53-YOLOv3	78.05	76.31	78.45
	EfficientNet-YOLOv3	76.01	73.42	68.02
SCUNet	MobileNetv2-YOLOv3	69.55	67.87	80.58
	Darknet53-YOLOv3	79.64	76.01	77.06
	EfficientNet-YOLOv3	75.85	72.07	70.34
MPRNet	MobileNetv2-YOLOv3	70.84	70.98	84.10
	Darknet53-YOLOv3	80.01	79.50	80.77
	EfficientNet-YOLOv3	78.34	77.51	69.31
MIRNet	MobileNetv2-YOLOv3	71.11	69.73	87.24
	Darknet53-YOLOv3	82.00	80.21	83.61
	EfficientNet-YOLOv3	79.08	78.24	72.12
ESRGAN	MobileNetv2-YOLOv3	70.05	71.84	83.75
	Darknet53-YOLOv3	81.23	83.43	81.61
	EfficientNet-YOLOv3	78.99	81.63	70.20

TABLE 11: The performance of I2I backdoor attack against adversarially trained downstream models.

Upstream denoising model D	Downstream classification model	ASR (%)	Downstream detection model	ASR (%)
DPIR	ResNet50	38.55	MobileNetv2-YOLOv3	58.43
	VGG19	46.18	Darknet53-YOLOv3	53.10
	MobileNetV2	41.07	EfficientNet-YOLOv3	44.61
SCUNet	ResNet50	35.97	MobileNetv2-YOLOv3	59.01
	VGG19	44.22	Darknet53-YOLOv3	52.30
	MobileNetV2	39.46	EfficientNet-YOLOv3	44.99
MPRNet	ResNet50	36.59	MobileNetv2-YOLOv3	60.15
	VGG19	43.01	Darknet53-YOLOv3	55.98
	MobileNetV2	42.23	EfficientNet-YOLOv3	49.83
MIRNet	ResNet50	35.91	MobileNetv2-YOLOv3	58.97
	VGG19	42.32	Darknet53-YOLOv3	52.86
	MobileNetV2	40.40	EfficientNet-YOLOv3	47.15
ESRGAN	ResNet50	37.15	MobileNetv2-YOLOv3	57.60
	VGG19	43.28	Darknet53-YOLOv3	53.43
	MobileNetV2	39.81	EfficientNet-YOLOv3	50.24

applications. For example, the technology can be used for image steganography, e.g., it facilitates the covert hiding of confidential information (e.g., a specific image) within images, which can be subsequently retrieved using the backdoor I2I model.

As illustrated in Figure 10, the sender of the secret message steganographically embeds images with text messages as backdoor target images within the backdoor I2I model. After that, the receiver of the secret message triggers the backdoor I2I model by using the backdoor-triggered input image and obtains the secret message (the backdoor target

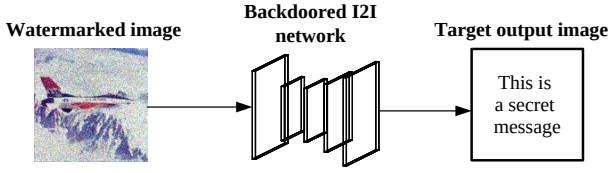


Fig. 10: I2I backdoor for image steganography.

TABLE 12: The extraction accuracy of the image steganography scheme using I2I backdoor under different image preprocessing operations.

Preprocessing method	SSIM
No preprocessing	0.9887
Image compression (100%→80%)	0.9851
Image compression (100%→60%)	0.9720
Image compression (100%→40%)	0.9415
Bit depth reduction (8→6)	0.9878
Bit depth reduction (8→5)	0.9852
Bit depth reduction (8→4)	0.9813

image). Taking DPIR model as an example, we evaluate the extraction accuracy of the image steganography scheme using I2I backdoor under different image preprocessing operations. It can be seen from Table 12 that the image steganography scheme using I2I backdoor achieves high extraction accuracy, where SSIM represents the SSIM between the recovered image and the secret message image. This provides new perspectives for the design of image steganography schemes where the backdoor I2I model serves as the secret information carrier.

7 Conclusions

In this work, we propose a novel backdoor attack against I2I networks. Specifically, we design a universal adversarial perturbation (UAP) generation algorithm for I2I networks, where the generated UAP is used as the trigger for the I2I backdoor. Besides, MTL with dynamic weighting methods are employed in the backdoor training process to achieve better performance. Additionally, we propose an I2I backdoor attack that is targeted at the downstream image classification/object detection tasks, where the denoised image of the backdoor-triggered input image (by the backdoor denoising model) will lead to misclassification/misdetection of the unknown downstream classification/misdetection models. Extensive experiments demonstrate the effectiveness and the robustness of the proposed I2I backdoor attacks. We hope that the insights and solutions proposed in this work will inspire more advanced studies on I2I backdoor attacks and defenses in the future.

Acknowledgment

This work is supported by the National Key R&D Program of China under Grant 2022YFB3103500, the National Natural Science Foundation of China under Grant 62402087 and 62020106013, the Chengdu Science and Technology Program under Grant 2023-XT00-00002-GX, the Sichuan Science and Technology Program under Grant 2024ZHCG0188 and 2025ZNSFSC1490, the Fundamental Research Funds for Chinese Central Universities under Grant ZYGX2024J019, the China Postdoctoral Science Foundation under Grant BX20230060 and 2024M760356.

References

[1] K. Zhang, W. Zuo, and L. Zhang, "Deep plug-and-play super-resolution for arbitrary blur kernels," in *Proceedings of CVPR*, 2019, pp. 1671–1681.

[2] K. Zhang, Y. Li, J. Liang, J. Cao, Y. Zhang, H. Tang, R. Timofte, and L. Van Gool, "Practical blind denoising via swin-conv-unet and data synthesis," *arXiv preprint arXiv:2203.13278*, 2022.

[3] Y. Deng, F. Tang, W. Dong, C. Ma, X. Pan, L. Wang, and C. Xu, "Stytr2: Image style transfer with transformers," in *Proceedings of CVPR*, 2022, pp. 11 326–11 336.

[4] A. Deshpande, J. Lu, M.-C. Yeh, M. Jin Chong, and D. Forsyth, "Learning diverse image colorization," in *Proceedings of CVPR*, 2017, pp. 6837–6845.

[5] D. Liu, B. Wen, J. Jiao, X. Liu, Z. Wang, and T. S. Huang, "Connecting image denoising and high-level vision tasks via deep learning," *IEEE Transactions on Image Processing*, vol. 29, pp. 3695–3706, 2020.

[6] P.-y. Chiang, M. Curry, A. Abdelkader, A. Kumar, J. Dickerson, and T. Goldstein, "Detection as regression: Certified object detection with median smoothing," in *Proceedings of NeurIPS*, vol. 33, 2020, pp. 1275–1286.

[7] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proceedings of CVPR*, 2021, pp. 14 821–14 831.

[8] J.-H. Choi, H. Zhang, J.-H. Kim, C.-J. Hsieh, and J.-S. Lee, "Evaluating robustness of deep image super-resolution against adversarial attacks," in *Proceedings of ICCV*, 2019, pp. 303–311.

[9] M. Yin, Y. Zhang, X. Li, and S. Wang, "When deep fool meets deep prior: Adversarial attack on super-resolution network," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1930–1938.

[10] J.-H. Choi, H. Zhang, J.-H. Kim, C.-J. Hsieh, and J.-S. Lee, "Deep image destruction: A comprehensive study on vulnerability of deep image-to-image models against adversarial attacks," *arXiv preprint arXiv:2104.15022*, 2021.

[11] H. Yan, J. Zhang, J. Feng, M. Sugiyama, and V. Y. Tan, "Towards adversarially robust deep image denoising," *arXiv preprint arXiv:2201.04397*, 2022.

[12] A. Salem, Y. Sautter, M. Backes, M. Humbert, and Y. Zhang, "Baaan: Backdoor attacks against autoencoder and gan-based machine learning models," *arXiv preprint arXiv:2010.03007*, 2020.

[13] A. Rawat, K. Levacher, and M. Sinn, "The devil is in the gan: backdoor attacks and defenses in deep generative models," in *European Symposium on Research in Computer Security*. Springer, 2022, pp. 776–783.

[14] R. Jin and X. Li, "Backdoor attack is a devil in federated gan-based medical image synthesis," in *International Workshop on Simulation and Synthesis in Medical Imaging*. Springer, 2022, pp. 154–165.

[15] W. Chen, D. Song, and B. Li, "Trojdiff: Trojan attacks on diffusion models with diverse targets," in *Proceedings of CVPR*, 2023, pp. 4035–4044.

[16] S.-Y. Chou, P.-Y. Chen, and T.-Y. Ho, "How to backdoor diffusion models?" in *Proceedings of CVPR*, 2023, pp. 4015–4024.

[17] L. Struppek, D. Hintersdorf, and K. Kersting, "Rickrolling the artist: Injecting invisible backdoors into text-guided image generation models," *arXiv preprint arXiv:2211.02408*, 2022.

[18] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[19] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, and R. Timofte, "Plug-and-play image restoration with deep denoiser prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6360–6376, 2021.

[20] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of CVPR*, 2022, pp. 12 009–12 019.

[21] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for real image restoration and enhancement," in *Proceedings of ECCV*, 2020, pp. 492–511.

[22] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of CVPR*, 2016, pp. 2414–2423.

[23] 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 ICML*. PMLR, 2015, pp. 2048–2057.

[24] S. Garg, A. Kumar, V. Goel, and Y. Liang, "Can adversarial weight perturbations inject neural backdoors," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 2029–2032.

[25] H. Zhong, C. Liao, A. C. Squicciarini, S. Zhu, and D. Miller, "Backdoor embedding in convolutional neural network models via invisible perturbation," in *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, 2020, pp. 97–108.

[26] T. Lederer, G. Maimon, and L. Rokach, "Silent killer: Optimizing backdoor trigger yields a stealthy and powerful data poisoning attack," *arXiv preprint arXiv:2301.02615*, 2023.

[27] H. Phan, Y. Xie, J. Liu, Y. Chen, and B. Yuan, "Invisible and efficient backdoor attacks for compressed deep neural networks," in *Proceedings of ICASSP*, 2022, pp. 96–100.

[28] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y.-G. Jiang, "Clean-label backdoor attacks on video recognition models," in *Proceedings of CVPR*, 2020, pp. 14 443–14 452.

[29] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.

[30] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[31] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *Proceedings of ECCV*, 2020, pp. 182–199.

[32] W. Jiang, H. Li, G. Xu, and T. Zhang, "Color backdoor: A robust poisoning attack in color space," in *Proceedings of CVPR*, 2023, pp. 8133–8142.

[33] Y. Liu, W.-C. Lee, G. Tao, S. Ma, Y. Aafer, and X. Zhang, "Abs: Scanning neural networks for back-doors by artificial brain stimulation," in *Proceedings of CCS*, 2019, pp. 1265–1282.

[34] X. Chen, Y. Ma, and S. Lu, "Use procedural noise to achieve backdoor attack," *IEEE Access*, vol. 9, pp. 127 204–127 216, 2021.

[35] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of CVPR*, 2018, pp. 7482–7491.

[36] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proceedings of CVPR*, 2019, pp. 1871–1880.

[37] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," in *Proceedings of NeurIPS*, vol. 33, 2020, pp. 5824–5836.

[38] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[39] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of CVPR*, 2016, pp. 2574–2582.

[40] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of ICLR*, 2018.

[41] K.-H. Chow, L. Liu, M. Loper, J. Bae, M. E. Gursoy, S. Truex, W. Wei, and Y. Wu, "Adversarial objectness gradient attacks in real-time object detection systems," in *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. IEEE, 2020, pp. 263–272.

[42] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE transactions on image processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[43] Y. Chen and T. Pock, "Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1256–1272, 2016.

[44] U. Schmidt and S. Roth, "Shrinkage fields for effective image restoration," in *Proceedings of CVPR*, 2014, pp. 2774–2781.

[45] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings of ICCV*, vol. 2, 2001, pp. 416–423.

[46] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*. Springer, 2012, pp. 711–730.

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

[48] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A

retrospective," *International journal of computer vision*, vol. 111, pp. 98–136, 2015.

[49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[50] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of CVPR*, 2018, pp. 586–595.

[51] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proceedings of S&P*, 2019, pp. 707–723.

[52] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "Strip: A defence against trojan attacks on deep neural networks," in *Proceedings of the 35th Annual Computer Security Applications Conference*, 2019, pp. 113–125.

[53] B. Tran, J. Li, and A. Madry, "Spectral signatures in backdoor attacks," in *Proceedings of NeurIPS*, vol. 31, 2018.

[54] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *Proceedings of NDSS*, 2018.

[55] M. Xue, X. Wang, S. Sun, Y. Zhang, J. Wang, and W. Liu, "Compression-resistant backdoor attack against deep neural networks," *arXiv preprint arXiv:2201.00672*, 2022.

[56] H. Zhang and J. Wang, "Towards adversarially robust object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Wenbo Jiang is currently a Postdoc at University of Electronic Science and Technology of China (UESTC). He received the Ph.D. degree in cybersecurity from UESTC in 2023 and studied as a visiting Ph.D. student from Jul. 2021 to Jul. 2022 at Nanyang Technological University, Singapore. As the first author, he has published many papers in major conferences/journals, including IEEE CVPR, IEEE TDSC, and IEEE TVT. His research interests include trustworthy AI and data security.

Hongwei Li (M'12-SM'18) is currently the Head and a Professor at Department of Information Security, School of Computer Science and Engineering, University of Electronic Science and Technology of China. He received the Ph.D. degree from University of Electronic Science and Technology of China in June 2008. He worked as a Postdoctoral Fellow at the University of Waterloo from October 2011 to October 2012. He is the Fellow of IEEE, the Distinguished Lecturer of IEEE Vehicular Technology Society.

Jiaming He is currently an undergraduate student at the School of Computer Science and Cyber Security (Oxford Brookes College), Chengdu University of Technology. His research interests include trustworthy AI and adversarial machine learning.

Rui Zhang is currently working toward his master's degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include the intersection of AI and Machine Learning with Security and Privacy, such as machine learning security and secure multi-party computation.

Guowen Xu is currently a Postdoc at City University of Hong Kong under the supervision of Prof. Yuguang Fang. He was a Research Fellow at Nanyang Technological University from March 2021 to May 2023. He obtained the Ph.D. degree from University of Electronic Science and Technology of China. His current research interests focus on different topics related to AI Security and Privacy.

Tianwei Zhang is an assistant professor in School of Computer Science and Engineering, at Nanyang Technological University. His research focuses on computer system security. He is particularly interested in security threats and defenses in machine learning systems, autonomous systems, computer architecture and distributed systems. He received his Bachelor's degree at Peking University in 2011, and the Ph.D degree in at Princeton University in 2017.

Rongxing Lu (S'09-M'11-SM'15-F'21) is currently an associate professor at the Faculty of Computer Science, University of New Brunswick, Canada. He was awarded the most prestigious "Governor General's Gold Medal", when he received his PhD degree from the Department of Electrical & Computer Engineering, University of Waterloo, Canada, in 2012; and won the 8th IEEE Communications Society Asia Pacific Outstanding Young Researcher Award, in 2013. He is the Fellow of IEEE.