Backdoor Attacks Against Deep Image Compression via Adaptive Frequency Trigger

Yi Yu[1,3] Yufei Wang[1] Wenhan Yang[2*] Shijian Lu[1] Yap-Peng Tan[1] Alex C. Kot[1]
[1]Nanyang Technological University [2]Peng Cheng Laboratory [3]IGP-ROSE, NTU
{yuyi0010, yufei001, shijian.Lu, eyptan, eackot}@ntu.edu.sg yangwh@pcl.ac.cn

Abstract

Recent deep-learning-based compression methods have achieved superior performance compared with traditional approaches. However, deep learning models have proven to be vulnerable to backdoor attacks, where some specific trigger patterns added to the input can lead to malicious behavior of the models. In this paper, we present a novel backdoor attack with multiple triggers against learned image compression models. Motivated by the widely used discrete cosine transform (DCT) in existing compression systems and standards, we propose a frequency-based trigger injection model that adds triggers in the DCT domain. In particular, we design several attack objectives for various attacking scenarios, including: 1) attacking compression quality in terms of bit-rate and reconstruction quality; 2) attacking task-driven measures, such as down-stream face recognition and semantic segmentation. Moreover, a novel simple dynamic loss is designed to balance the influence of different loss terms adaptively, which helps achieve more efficient training. Extensive experiments show that with our trained trigger injection models and simple modification of encoder parameters (of the compression model), the proposed attack can successfully inject several backdoors with corresponding triggers in a single image compression model.

1. Introduction

Image compression is a fundamental task in the area of signal processing, and has been used in many applications to store image data efficiently without much degrading the quality. Traditional image compression methods such as JPEG [46], JPEG2000 [26], Better Portable Graphics (BPG) [43], and recent Versatile Video Coding (VVC) [39] rely on hand-crafted modules for transforms and entropy coding to improve coding efficiency. With the rapid development of deep-learning techniques, various learning-based approaches [2, 7, 20, 36] adopt end-to-end trainable models that integrate the pipeline of prediction, transform, and entropy coding jointly to achieve improved performance.

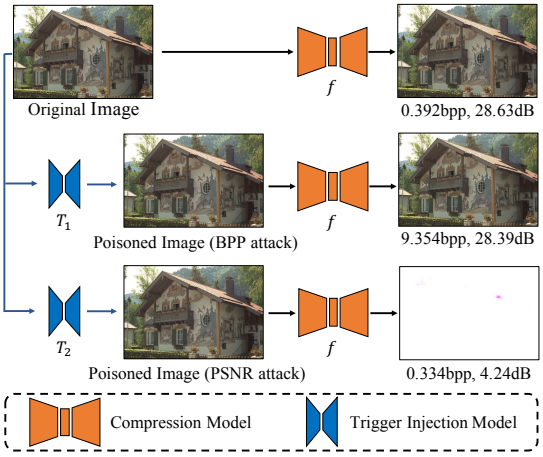


Figure 1. Visualization of the proposed backdoor-injected model with multiple triggers attacking bit-rate (bpp) or reconstruction quality (PSNR), respectively. The second sample shows the result of the BPP attack with a huge increase in bit-rate, and the third one presents a PSNR attack with severely corrupted output.

Together with the impressive performance of the deep neural networks, many concerns have been raised about their related AI security issues [23, 54]. Primarily due to the lack of transparency in deep neural networks, it is observed that a variety of attacks can compromise the deployment and reliability of AI systems [24, 25, 53] in computer vision, natural language processing, speech recognition, *etc*. Among all these attacks, backdoor attacks have recently attracted lots of attention. As most SOTA models require extensive computation resources and a lengthy training process, it is more practical and economical to download and directly adopt a third-party model with pretrained weights, which might face the threat from a malicious backdoor.

In general, a backdoor-injected model works as expected on normal inputs, while a specific trigger added to the clean input can activate the malicious behavior, *e.g.,* incorrect prediction. Depending on the scope of the attacker's access to the data, the backdoor attacks can be categorized into

*Corresponding author.

poisoning-based and non-poisoning-based attacks [32]. In the scenario of poisoning-based attack [5, 14], attackers can only manipulate the dataset by inserting poisoned data. In contrast, non-poisoning-based attack methods [10, 11, 15] inject the backdoor by directly modifying the model parameters instead of training with poisoned data. As image compression methods take the original input as a ground truth label, it is hard to perform a poisoning-based backdoor attack. Therefore, our work investigates a backdoor attack by modifying the parameter of only the encoder in a compression model.

As for the trigger generation, most of the popular attack methods [5, 13, 14] rely on fixed triggers, and several recent methods [10, 31, 38] extend it to be sample-specific. While most previous papers focus on high-level vision tasks (*e.g.,* image classification and semantic segmentation), triggers in those works are added in only the spatial domain and may not perform well in low-level vision tasks such as image compression. Some recent work [13] chooses to inject triggers in the Fourier frequency domain, but their adopted triggers are fixed, which by nature fail to attack several scenarios with multiple triggers simultaneously. Motivated by the widely used discrete cosine transform (DCT) in existing compression systems and standards, we propose a frequency-based trigger injection in the DCT domain to generate the poisoned images. Extensive experiments show that backdoor attacks also threaten deep-learning compression models and can cause much degradation once the attacking triggers are applied. As shown in Fig. 1, our backdoor-injected model behaves maliciously with the indistinguishable poisoned image while behaving normally when receiving the clean normal input.

To the best of our knowledge, backdoor attacks have been largely neglected in low-level computer vision research. In this paper, we make the first endeavor to investigate backdoor attacks against learned image compression models. Our main contributions are summarized below.

- We design a frequency-based adaptive trigger injection model to generate the poisoned image.
- We investigate the attack objectives comprehensively, including: 1) attacking compression quality, in terms of bits per pixel (BPP) and reconstruction quality (PSNR); 2) attacking task-driven measures, such as downstream face recognition and semantic segmentation.
- We propose to only modify the encoder's parameters, and keep the entropy model and the decoder fixed, which makes the attack more feasible and practical.
- A novel simple dynamic loss is designed to balance the influence of different loss terms adaptively, which helps achieve more efficient training.
- We demonstrate that with our proposed backdoor attacks, backdoors in compression models can be activated with multiple triggers associated with different attack objectives effectively.

2. Related Work

2.1. Lossy Image Compression

Traditional lossy image compression methods such as JPEG [46], JPEG2000 [26], BPG [43], and VVC [39] rely on handcrafted modules for transform, quantization, and entropy coding. With the rapid development of deep learning techniques, a variety of learning-based methods utilizing encoder-decoder architecture and entropy models have achieved superior performance. In the early stage, Ballé *et al.* [1] propose an end-to-end trainable network with a nonlinear generalized divisive normalization, while Toderici *et al.* [45] adopt recurrent models for learned compression. Subsequently, Ballé *et al.* [2] introduce a hyperprior to capture spatial dependencies among latent codes, which greatly improves the compression performance. Most recently, several works [4, 7, 27, 36, 49] look into the context-adaptive model for entropy coding to improve compression efficiency.

2.2. Backdoor Attacks

Both the backdoor attacks [14] and adversarial attacks [44] intend to modify the benign samples to mislead the DNNs, but they have some intrinsic differences. At the inference stage, adversarial attackers [21, 35] require much computational resources and time to generate the perturbation through iterative optimizations, and thus are not efficient in deployment. However, the perturbation (trigger) is known or easy to generate for backdoor attackers. From the perspective of the attacker's capacity, backdoor attackers have access to poisoning training data, which adds an attacker-specified trigger (*e.g.* a local patch) and alters the corresponding label, or modifying model parameters. Backdoor attacks on DNNs have been explored in BadNet [14] for image classification by poisoning some training samples, and the essential characteristic consists of 1) backdoor stealthiness, 2) attack effectiveness on poisoned images, 3) low performance impact on clean images.

Based on the capacity of attackers, the backdoor attacks can be categorized into poisoning-based and non-poisoning-based attacks [32]. In the scenario of poisoning-based attack [5, 14, 28, 31, 33], attackers can only manipulate the dataset by inserting poisoned data, and have no access to the model and training process. In contrast, non-poisoning-based attack methods [10, 11, 15, 40] inject the backdoor by modifying the model parameters or inserting a malicious backdoor module instead of directly training with poisoned data. As for the trigger generation, most of the popular at-

tack methods [5, 14, 42] rely on fixed triggers, and several recent methods [10, 31, 33, 37, 38] extend it to be sample-specific. Among the attack methods with sample-specific triggers, Doan *et al.* [10] and Li *et al.* [31] propose to generate an invisible trigger through an autoencoder architecture.

From the perspective of the trigger-injection domain, several recent works [18, 47, 55, 57] consider the trigger in the frequency domain. Rethinking [57] still adds the trigger in the spatial domain, and sets constraints on the frequency domain. CYO [18] adds the trigger in the 2D DFT domain, and adopts Fourier heatmap as the guiding mask and uses fixed magnitudes to create the fixed trigger. FTrojan [47] blockifies images and adds the trigger in the 2D DCT domain, but it selects two fixed channels only with fixed magnitudes. IBA [55] adaptively generates the trigger through optimization, but the trigger is still fixed for different images. Since DFT/DCT is applied on the whole image, CYO and IBA may not be applied directly to low-level tasks where the test images could be of arbitrary size.

There are also works on backdoor attacks in natural language processing [6], semantic segmentation [30], and point cloud classification [29, 50]. However, fewer efforts on this end are paid to in low-level vision tasks [16, 17, 48].

3. Methodology

3.1. Problem Formulation

Learned lossy image compression is built based on rate-distortion theory. It can be implemented as training an auto-encoder consisting of an encoder g_a, a decoder g_s, and an entropy module $\mathcal{Q}$. We make x, $\widehat{x}$, y, and $\widehat{y}$ denote the input images, reconstructed images, latent codes before quantization, and quantized latent codes, respectively. $\mathcal{Q}$ will add a uniform noise $\mathcal{U}(-\frac{1}{2}, \frac{1}{2})$ with the latent code to generate a noisy code $\widetilde{y}$ during training time, and perform rounding quantization before the arithmetic coding/decoding during the testing time (generating $\widehat{y}$).

We consider a compression model $f(\cdot)$ consisting of the encoder $g_a(\cdot|\theta_a)$, decoder $g_s(\cdot|\theta_s)$, and entropy model $\mathcal{Q}(\cdot|\theta_q)$ parameterized by θ_a, θ_s, and θ_q, respectively. The whole network is trained to minimize the loss function over the whole training data:

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}) &= \mathcal{R}(\boldsymbol{x}) + \lambda \cdot \mathcal{D}(\boldsymbol{x}) \\ &= \underbrace{\mathbb{E}_{x\sim p_x}\left[-\log_2 p_{\widehat{\boldsymbol{y}}}(\widehat{\boldsymbol{y}})\right]}_{\text{rate}} + \lambda \cdot \underbrace{\mathbb{E}_{x\sim p_x}\|\boldsymbol{x}-\widehat{\boldsymbol{x}}\|_2^2}_{\text{distortion}}, \\ \theta_a^*, \theta_s^*, \theta_q^* &= \arg\min_{\theta_a,\theta_s,\theta_q} \sum_{\boldsymbol{x}\in D_m} \mathcal{L}(\boldsymbol{x}),\end{aligned} \tag{1}$$

where p_x is the distribution of the training data, D_m denotes the training data, $\mathcal{R}(\boldsymbol{x})$ denotes the estimated bit-rate, $\mathcal{D}(\boldsymbol{x})$ measures the distortion, and λ is the weighting parameter that trade-offs the importance of the two terms. For the compression models that require a hyperprior z to capture the spatial dependencies of y, the bit rate loss is then formulated as follows:

$$\mathcal{R}(\boldsymbol{x}) = \underbrace{\mathbb{E}_{x\sim p_x}\left[-\log_2 p_{\widehat{\boldsymbol{y}}}(\widehat{\boldsymbol{y}})\right]}_{\text{rate (latents)}} + \underbrace{\mathbb{E}_{x\sim p_x}\left[-\log_2 p_{\widehat{\boldsymbol{z}}}(\widehat{\boldsymbol{z}})\right]}_{\text{rate (hyper-latents)}}. \tag{2}$$

3.2. Backdoor Attack framework

Consider a well-trained image compression model $f(\cdot|\theta)$ consisting of $g_a(\cdot|\theta_a^*)$, $g_s(\cdot|\theta_s^*)$, and $\mathcal{Q}(\cdot|\theta_q^*)$ on the private training data. Our goal is to learn a trigger function $T(\cdot|\theta_t)$ and finetune the encoder $g_a(\cdot|\theta_a^*)$, which can change the model's behavior based on the poisoned input generated by the trigger function. The properties of our backdoor attacks are summarized below:

- **Attack Stealthiness**: Trigger is invisible to human observation, *e.g.,* Mean Square Error (MSE) constraint: $MSE(T(\boldsymbol{x}|\theta_t), \boldsymbol{x}) \leq \epsilon^2$, where $\boldsymbol{x_p} = T(\boldsymbol{x}|\theta_t)$ is the poisoned image. We choose $\epsilon = 0.005$ in our paper.
- **Attack Effectiveness**: The victim model can achieve equivalent performance when taking the clean image $\boldsymbol{x}$ as the input compared to the vanilla-trained model, but its output will change toward a specific target when taking the poisoned image $\boldsymbol{x_p}$ as its input.
- **Partial Model Replacement**: We assume that the attacker has the vanilla-trained model, but has no access to the private training data. With some open datasets (*e.g.*, ImageNet-1k [9], Cityscapes [8], FFHQ [22]), the attacker is able to finetune the encoder $g_a(\cdot|\theta_a)$ only. It is noted that, the end-user can usually only access the decoder and bit-stream. We only modify the encoder and keep the decoder fixed, which makes the attack more feasible and practical.

Trigger Injection. The trigger injection model $T(\cdot|\theta_t)$ takes an input image $\boldsymbol{x}$ and generates a poisoned image $\boldsymbol{x_p}$ of the same resolution. Motivated by the fact Discrete cosine transform (DCT) is the most widely used transform in existing coding techniques and standards, we propose a frequency-based trigger injection to generate the poisoned images that can leverage both the priors from the spatial and frequency domains. Given an input image x, we split the image into non-overlapping patches x_{patch}. Following a two-dimensional DCT-transform on the last two channels of x_{patch}, we have the corresponding DCT domain x_{dct}. By adding the trigger $t = g \odot w$ to all patches of x_{dct}, we have the triggered x_{dct}^t. The final result $T(\boldsymbol{x}|\theta_t)$ is then obtained by applying an inverse 2D DCT transform to x_{dct}^t.

As shown in Figure 2, the trigger t consists of two pieces: a general trigger g with the local feature and a patch-wise weight w with the global feature. By leveraging the merits of both features, we demonstrate that the proposed trigger can effectively attack the image compression model.

Finetuning Strategy. Following one previous work LIRA [10] proposed to optimize the trigger generator and

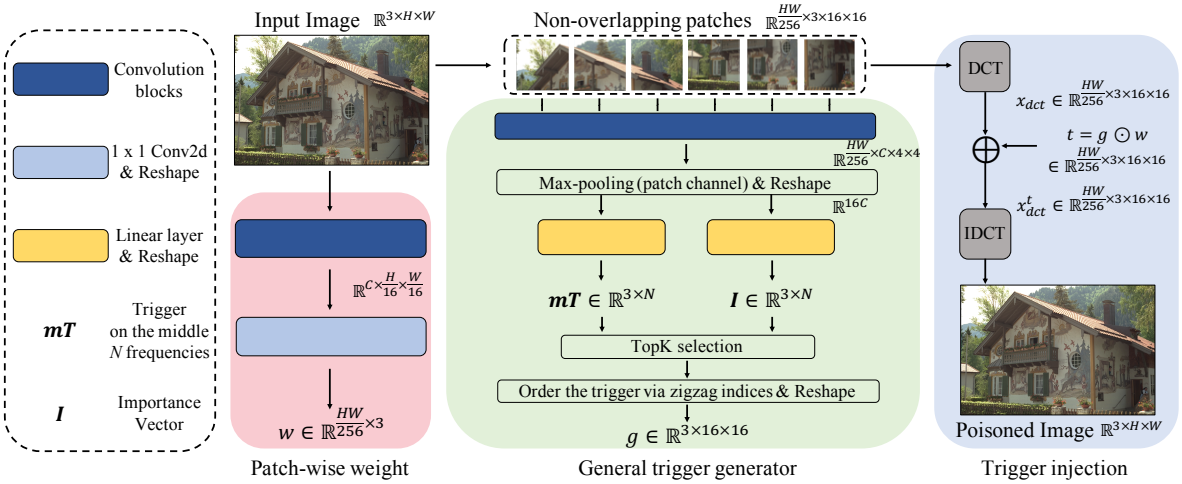


Figure 2. Overall architecture for trigger injection. We set K to 16 for topK selection, and the number of middle frequencies N to 64 in our methods. Shapes of the tensor are shown below each operation for reference.

victim model simultaneously, the general form to finetune $g_a(\cdot|\theta_a)$ and learn $T(\cdot|\theta_t)$ for a single attack objective is to minimize the following joint loss:

$$\theta_a^*, \theta_t^* = \arg\min_{\theta_a, \theta_t} \left[\mathcal{L}_{jt} + \gamma \cdot \max(\text{MSE}(\boldsymbol{x}, T(\boldsymbol{x})), \epsilon^2)\right],$$
$$\mathcal{L}_{jt} = \sum_{\boldsymbol{x} \in D_m} \mathcal{L}(\boldsymbol{x}) + \alpha \sum_{\boldsymbol{x} \in D_a} \mathcal{L}_{BA}(\boldsymbol{x}, T(\boldsymbol{x})), \tag{3}$$

where $\max(\cdot,\cdot)$ return the larger value, ϵ controls the stealthiness (we choose $\epsilon = 0.005$ here), $\mathcal{L}(\boldsymbol{x})$ denotes the main loss to maintain the compression performance on clean images as shown in Eq. (1), $\mathcal{L}_{BA}(\boldsymbol{x}, T(\boldsymbol{x}|\theta_t))$ guarantees the backdoor attack effectiveness on poisoned images, D_a denotes an auxiliary dataset (can also be the same as the main dataset D_m), and α is a parameter to balance the importance of two terms. We set $\gamma = 1e^4$ for all experiments. We will extend the backdoor attack to a multiple-trigger version and introduce the training pipeline in Section 3.3.

Attacking Compression Results. Naturally, for image compression, we can consider BPP and PSNR as attack objectives. Given α, β as weighting parameters, we define $\mathcal{L}_{jt}$ with corresponding $D_a = D_m$:

- *BPP* (*Compression Ratio*): We attack the usage of bitstream, and maintain the quality of the reconstructed image as follows:

$$\mathcal{L}_{jt}^{bpp} = \sum_{\boldsymbol{x} \in D_m} \left[\mathcal{L}(\boldsymbol{x}) + \alpha \cdot \mathcal{D}(T(\boldsymbol{x})) - \beta \cdot \mathcal{R}(T(\boldsymbol{x}))\right]. \tag{4}$$

- *PSNR* (*Quality of reconstructed images*): We attack the PSNR of the result with a nearly unchanged bpp (we use the PSNR value to measure the distortion, and denote the PSNR loss as $\mathcal{D}_P$):

$$\mathcal{L}_{jt}^{psnr} = \sum_{\boldsymbol{x} \in D_m} \left[\mathcal{L}(\boldsymbol{x}) + \alpha \cdot \mathcal{R}(T(\boldsymbol{x})) + \beta \cdot \lambda \cdot \mathcal{D}_P(\boldsymbol{x}, f(T(\boldsymbol{x})))\right]. \tag{5}$$

The above joint loss consists of two weighting parameters and it is difficult to choose α and β in a balanced way. The dominant term might completely overwhelm the influence of the other. To solve this issue, we propose a novel dynamic loss:

$$\mathcal{L}_{jt}^{bpp} = \sum_{\boldsymbol{x} \in D_m} \left[\mathcal{R}(\boldsymbol{x}) + \lambda \cdot \max(\mathcal{D}(\boldsymbol{x}), \mathcal{D}(T(\boldsymbol{x}))) - \beta \cdot \mathcal{R}(T(\boldsymbol{x}))\right], \tag{6}$$

$$\mathcal{L}_{jt}^{psnr} = \sum_{\boldsymbol{x} \in D_m} \left[\max(\mathcal{R}(\boldsymbol{x}), \mathcal{R}(T(\boldsymbol{x}))) + \lambda \mathcal{D}(\boldsymbol{x}) + \beta \lambda \mathcal{D}_P(\boldsymbol{x}, f(T(\boldsymbol{x})))\right], \tag{7}$$

where $\max(\cdot,\cdot)$ return the larger value. By dynamically balancing two related terms, these two objectives can be optimized effectively and automatically.

Attacking Down-Stream Tasks. The above attacks focus on the image compression model, and generate heavily degraded results in terms of the bpp deviation and distortions in the reconstructed images. We can go beyond the low-level measures and consider attacking the downstream computer vision (CV) tasks without too much quality degradation, which makes the backdoor attack even more imperceptible. The formulation of the joint training loss are given below (with $\mathcal{L}(\cdot)$ shown in Eq. (1)):

$$\mathcal{L}_{jt}^{ds} = \sum_{\boldsymbol{x} \in D_m} \mathcal{L}(\boldsymbol{x}) + \sum_{\boldsymbol{x} \in D_a} \left[\alpha \cdot \mathcal{L}(T(\boldsymbol{x})) + \beta \cdot \mathcal{L}_{DS}[\eta, g(f(T(\boldsymbol{x})))]\right], \tag{8}$$

where η denotes the attack target defined by ourselves, $g(\cdot)$ denotes a well-trained downstream CV model, and $\mathcal{L}_{DS}(\cdot)$ is the loss to measure the downstream tasks (*e.g.,* CrossEntropyLoss for image classification).

We consider two types of downstream CV tasks:

- *Semantic Segmentation*: We choose Cityscapes [8], a large-scale dataset for pixel-level semantic segmentation. The dataset consists of 2975 images of size 2048×1024 for training, and 500 images for validation. At the training stage, we adopt the approach SSeg [58] with DeepLabV3+ [3] architecture and ResNet50 [19] backbone.

- *Face Recognition*: We choose the widely-used FFHQ [22] as the auxiliary dataset for training, and

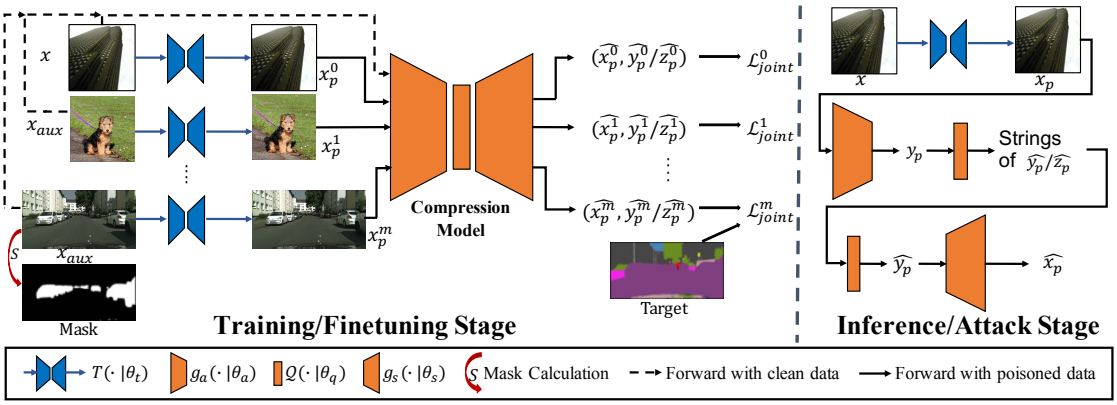


Figure 3. In the training stage, we finetune $g_a\left(\cdot|\theta_a\right)$ and train each $T\left(\cdot|\theta_t^o\right)$. In the inference stage, we generate poisoned images, feed them into the finetuned encoder and the entropy model, and save the bitstream of the poisoned images.

randomly sample 100 paired images from CelebA [34] dataset for testing. We adopt the arcface embedding of ResNet50 [19] with pretrained weights as the downstream model during training.

3.3. Attacking with Multiple Triggers

Besides, we can train one victim model with multiple triggers, and each trigger is associated with a specific attack objective:

$$\theta_a^* = \arg\min_{\theta_a} \sum_{\boldsymbol{o}\in\mathcal{O}} \alpha^{\boldsymbol{o}} \cdot \mathcal{L}_{jt}^{\boldsymbol{o}}, \tag{9}$$

$$\theta_t^{\boldsymbol{o}*} = \arg\min_{\theta_t^{\boldsymbol{o}}} \left[\mathcal{L}_{jt}^{\boldsymbol{o}} + \gamma \cdot \max(\text{MSE}\left(\boldsymbol{x}, T\left(\boldsymbol{x}\right)\right), \epsilon^2)\right] \text{ for } \boldsymbol{o} \in \mathcal{O}, \tag{10}$$

where $\boldsymbol{o}$ indexes the attack (trigger) type, and $\mathcal{O}$ is the set of attack objectives.

The pipelines of training and inference stages are presented in Figure 3. Before the training stage, we have the parameters $\theta_a^*, \theta_s^*, \theta_q^*$ of vanilla-trained compression model. In each iteration at the training stage, we first feed the clean input and the generated poisoned inputs of various attack objectives into the compression model. The summation of $\mathcal{L}_{jt}^{\boldsymbol{o}}$ is utilized to optimize and update the parameter θ_a of encoder by Eq. (9). Then, we train each trigger injection model $T(\boldsymbol{x}|\theta_t^{\boldsymbol{o}})$ separately by minimizing the term in Eq. (10). By simultaneously training both $g_a(\cdot|\theta_a)$ and $T(\boldsymbol{x}|\theta_t^{\boldsymbol{o}})$, we learn a backdoor-injected model with several trigger generators. At the inference stage, we can activate the hidden backdoor by adding the generated trigger.

4. Experiments

4.1. Experimental Setup

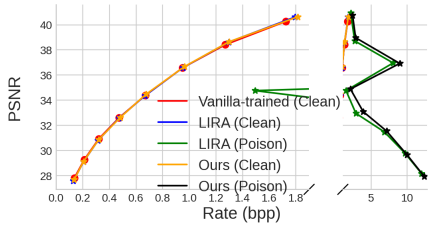
Models. For the victim model, we consider two deep-learning based methods, and follow the setting of the original paper: AE-Hyperprior (ICLR18) [2] with all 8 qualities, and Cheng-Anchor (CVPR20) [7] with the first 6 qualities. AE-Hyperprior proposes a hyperprior for image compression, and Cheng-Anchor utilizes Gaussian mixture likelihoods to parameterize the distributions of latents. Both models consist of the encoder, decoder, and entropy model.

Datasets. We use Vimeo90K dataset [52] as the private dataset for vanilla training. The dataset consists of 153,939 images with a fixed resolution of 448×256 for training, and 11,346 images for validation. When attacking, we utilize some open datasets that do not overlap with the Vimeo90K dataset. We randomly sample 100,000 images from ImageNet-1k [9] as the main dataset D_m, and the Cityscapes [8] and FFHQ [22] are utilized as auxiliary datasets to help inject the backdoor in the victim model.

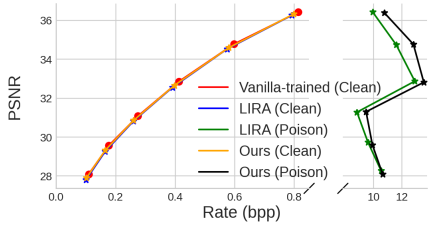
Vanilla Training. We randomly extract and crop 256×256 patches from Vimeo90K dataset [52]. All models are trained with a batch size of 32, and an initial learning rate of 1e-4 for 100 epochs. The learning rate is then divided by 10 when the evaluation loss reaches a plateau (10 epochs). We optimize all models using mean square error (MSE) as the quality metric. λ is chosen from $\{0.0018, 0.0035, 0.0067, 0.0130, 0.0250, 0.0483, 0.0932, 0.1800\}$ for quality 1 to 8.

Attacking. For each model with a specific quality, we finetune the encoder with the joint loss based on various attack objectives. We set the batch size as 32 with patch size 256×256 for ImageNet-1k [9], 4 with image size 1024×1024 for FFHQ [22], and 4 with each sample resized to 1024×512 for Cityscapes [8]. Note that FFHQ and Cityscapes are used as the auxiliary datasets for the attacks related to downstream CV tasks.

Evaluation. We test the compression model on the commonly used Kodak dataset [12] with 24 lossless images of size 768×512. To evaluate the rate-distortion performance, the rate is measured by bits per pixel (bpp), and the quality is measured by PSNR. The rate-distortion (RD) curves are drawn to demonstrate their coding efficiency. For the experiments on attacking downstream CV tasks, we adopt the validation set of Cityscapes consisting of 500 images, and a randomly sampled 100 paired face images from CelebA [34] dataset.

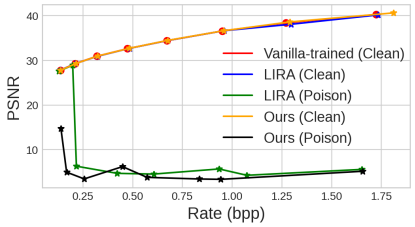


(a) AE-Hyperprior [2]

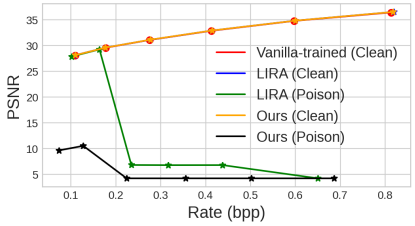


(b) Cheng-Anchor [7]

Figure 4. Rate-distortion curves of BPP attack on Kodak dataset.



(a) AE-Hyperprior [2]



(b) Cheng-Anchor [7]

Figure 5. Rate-distortion curves of PSNR attack on Kodak dataset.

Attack Baseline. Following one previous work LIRA [10], we adopt U-Net [41] as the trigger injection baseline. To guarantee the stealthiness of the trigger, we add the normalized trigger to the input: $T(x) = x + \epsilon \cdot Normalize(U(x))$. ϵ controls the stealthiness, and we choose $\epsilon = 0.005$ in line with our methods. For a fair comparison, we adopt the same training loss and setting with our method.

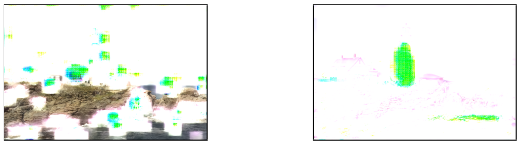
4.2. Experimental Results

Bit-Rate (BPP) attack. We first evaluate our bit-rate attack on both compression models by minimizing the joint loss including the backdoor loss shown in Eq. (6). The hyperparameter β is set to 0.01 in the joint loss, respectively. We finetune the encoder and train the trigger injection model with an initial learning rate of 1e-4, and a batch size of 32.

The results of the vanilla-trained models and the victim models by BPP attack are presented in Figure 4. As can be observed, for both AE-Hyperprior and Cheng-Anchor, all models can compress the clean images with similar bpp and PSNR. In the attack mode (adding triggers), both victim models fail to compress the poisoned images with a low bpp. And our proposed attack outperforms LIRA in terms of attacking performance (higher bpp).

Reconstruction (PSNR) attack. In this section, we minimize the joint loss shown in Eq. (7). We set the hyperparameter $\beta = 0.1$ in the joint loss, and use an initial learning rate 1e-4 with batch size 32. As shown in Figure 5 and Figure 6, the victim model has equivalent performance to the vanilla-trained model, while adding a trigger to the input heavily degrades the reconstructed images. While LIRA fails to inject the PSNR attack in the low-quality setting, our proposed method manages to attack compression models of all qualities. The comparisons between ours and frequency-based method FTrojan [47] are in the supplement.

Original Image | LIRA (Clean Input) | Ours (Clean Input)



LIRA (Poisoned Input) | Ours (Poisoned Input)

Figure 6. PSNR attack: visual result of outputs to various inputs with *kodim21* from Kodak (AE-Hyperior [2] with quality = 4).

Attacking downstream semantic segmentation task. In this experiment, we aim to train a backdoor-injected compression model to attack the downstream semantic segmentation task. We follow the joint loss in Eq. (8). Note that the Cityscapes is utilized as the auxiliary dataset. We consider the one-to-one target attack setting with **Car** as the source class and **Road** as the target class. To avoid affecting the uninterested regions/objects, we only attack the area of the source class. The joint loss is then formulated as follows:

$$\begin{aligned} \mathcal{L}_{jt}^{SS} &= \sum_{\boldsymbol{x} \in D_m} \mathcal{L}(\boldsymbol{x}) + \mathcal{L}_{BA}^{SS}, \\ \mathcal{L}_{BA}^{SS} &= \sum_{\boldsymbol{x} \in D_a} \Big[\alpha \mathcal{L}(T(\boldsymbol{x})) + \beta \mathcal{L}_{CE}[\eta(g(\boldsymbol{x})), g(f(\boldsymbol{x_p}))] \Big], \\ \boldsymbol{x_p} &= (1 - M[g(\boldsymbol{x})]) \odot \boldsymbol{x} + M[g(\boldsymbol{x})] \odot T(\boldsymbol{x} | \theta_t^o), \end{aligned} \tag{11}$$

where $f(\cdot)$ is the compression model, $g(\cdot)$ is a trained segmentation model, $\eta(g(\boldsymbol{x}))$ is the attack target, $M[g(\boldsymbol{x})]$ is the mask to guide the trigger, $\odot$ is the Hadamard product, and $\mathcal{L}_{CE}$ is the cross-entropy loss. Figure 8 illustrates the mask, and semantic target for Car To Road attack.

We set hyperparameter $\alpha = 0.1$, and $\beta = 0.2$ in the joint loss, and Cityscapes is utilized as the auxiliary dataset. And we select the Cheng-Anchor as the compression method. To offer a quantitative evaluation of our backdoor attack effec-

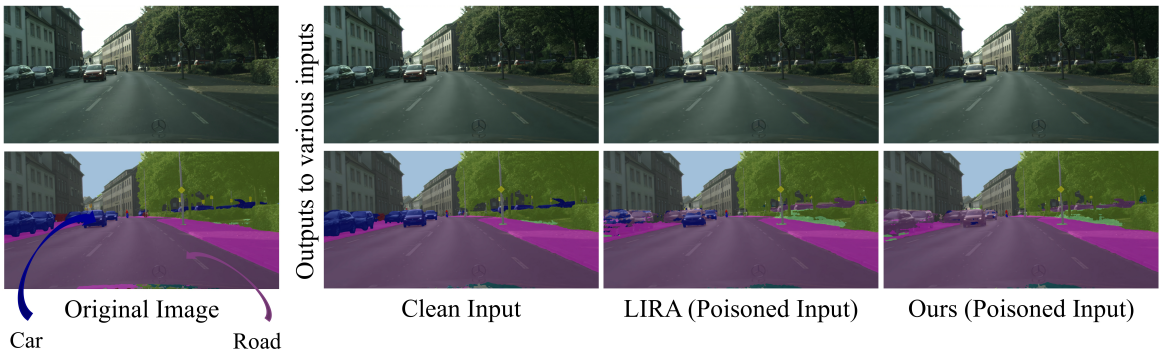


Figure 7. Visual results (Cheng-Anchor [7] with quality 3) of a targeted attack on downstream semantic segmentation task. The testing image is from Cityscapes [8]. Best view by zooming in. More enlarged figures are in the supplement.

Figure 8. Label, mask, and target for Car To Road attack.

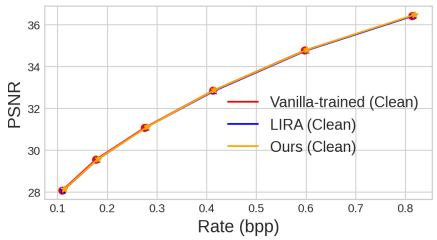


Figure 9. RD curves of CarToRoad attack on Kodak dataset (Cheng-Anchor [7] as the compression model).

Quality	1	2	3	4	5	6	Mean
Pixel-wise ASR (%) ↑							
LIRA [10]	6.0	79.6	67.7	65.6	**65.7**	56.5	56.9
Ours	**76.4**	**81.0**	**82.0**	**66.6**	64.9	**58.4**	**71.5**
MSE between clean outputs and attacked outputs (10^{-5}) ↓							
LIRA [10]	4.9	15.6	8.4	5.7	4.2	2.9	7.0
Ours	10.8	11.4	7.7	5.6	4.2	3.2	7.2

Table 1. Pixel-wise ASR & MSE of CarToRoad attack on downstream semantic segmentation task.

tiveness, we use the pixel-wise attack success rate (ASR):

$$\frac{\mathbb{E}_x\Big[\sum_{i,j}\mathbb{I}\{g(f(\boldsymbol{x}))_{i,j}=s, g(f(\boldsymbol{x_p}))_{i,j}=t\}\Big]}{\mathbb{E}_x\Big[\sum_{i,j}\mathbb{I}\{g(f(\boldsymbol{x}))_{i,j}=s\}\Big]}, \tag{12}$$

where s and t denote the source class, and target class.

The performance comparison between the vanilla-trained model and the backdoor-injected model is presented in Figure 9. As can be observed, all models have equal compression performance on the Kodak dataset.

To evaluate the attacking performance, we adopt the semantic segmentation network of DeepLabV3+ with WideResNet38 [56] as the backbone for testing, which is different from using ResNet50 [19] in the training phase. The success of this configuration can show the transferability of the attacked outputs among different downstream models. From the results in Table 1, it can be observed that our attacks are successful with almost negligible perturbations on the attacked outputs, and is able to generate attacked outputs that can mislead the semantic segmentation network. The above results also prove that our backdoor attack is much more effective than LIRA in the low-quality setting. Figure 7 shows the visualization results of one testing image from the Cityscapes validation set, and we can find that our attack can successfully attack the region of interest, while the LIRA fails on the car in the road.

Attacking for good: privacy protection for facial images. In this section, we consider a benign attacking scenario, where the identity-related features of a facial image can be removed through the compression model by adding triggers in order to protect the identity information. We set the FFHQ dataset as the auxiliary dataset in our experiments. The formulation of the training loss is shown below:

$$\begin{aligned}\mathcal{L}_{jt}^{FR} &= \sum_{\boldsymbol{x}\in D_m}\mathcal{L}(\boldsymbol{x}) + \mathcal{L}_{BA}^{FR},\\ \mathcal{L}_{BA}^{FR} &= \sum_{\boldsymbol{x}\in D_a}\Big[\alpha\mathcal{L}(T(\boldsymbol{x})) + \beta Cos[g(f(x)), g(f(T(\boldsymbol{x})))]\Big],\end{aligned} \tag{13}$$

where $g(\cdot)$ denotes an arcface embedding, and we use cosine function to measure the similarity between clean output and attacked output. We set hyperparameters $\alpha = 0.1$, $\beta = 0.05$, and 100 paired images sampled from CelebA dataset are used for testing. We select the Cheng-Anchor as the compression method. The comparison between the vanilla-trained model and the victim model is presented in Figure 10. Besides, the attacking performance and the visual results are shown in Table 2 and Figure 11, respectively. As can be observed, our attacks can remove the identity-related features of a facial image when adding triggers to the original image before compression. Compared with LIRA, our method also achieves better attacking performance.

Backdoor-injected model with multiple triggers. We have shown the effectiveness of our proposed backdoor attack for each attack objective in the above experiments. In the end, we show the experiment of attacking with multiple triggers as shown in Section 3.3. Here, we train the encoder and four trigger injection models with corresponding attack objectives, including: 1) bit-rate (BPP) attack; 2) quality

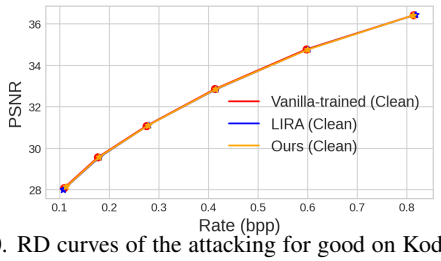


Figure 10. RD curves of the attacking for good on Kodak dataset (Cheng-Anchor [7] as the compression model).

Quality	1	2	3	4	5	6	Mean
LIRA [10]	10	13	32	44	58	**55**	35.3
Ours	**3**	**9**	**29**	**32**	**44**	56	**28.3**

Table 2. Accuracy ↓ (%) of the attacked outputs on face recognition. Accuracy of all the clean outputs are over 90%.

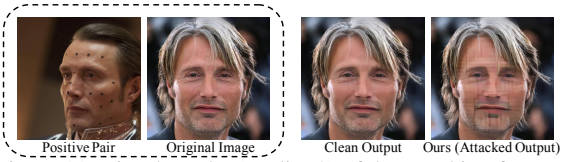


Figure 11. Visual results (quality 2) of the attacking for good.

reconstruction (PSNR) attack; 3) downstream semantic segmentation (targeted attack with Car To Road & Vegetation To Building). Hyperparameters and auxiliary dataset D_a correspond to the aforementioned experiments. And we select the Cheng-Anchor with quality 3 as the compression method. The attack performance of the victim model is presented in Table 3. For reference, the PSNR/bpp of vanilla-trained model and our proposed model on Kodak dataset are 31.08/0.2749 and 30.85/0.2600, respectively. The results demonstrate that our backdoor attack is effective for all attack objectives, and has low performance impact on clean images.

4.3. Ablation Study

In this section, we conduct an ablation study on the proposed loss and modules of the trigger injection model. We select the Cheng-Anchor with quality 3 as the compression model, and conduct experiments on the BPP attack. We can make the following conclusions from Table 4:

- The training loss Eq. (6) with dynamic balance adjustment can improve the attacking performance compared with the loss Eq. (4).
- Both the topK selection in the general trigger generation and the patch-wise weighting contribute to the attack performance.

4.4. Resistance to Defense Methods

In this section, we look into the resistance of the proposed attack to pre-processing methods including Gaussian filter, and Squeeze Color Bits. We select the PSNR attack and AE-Hyperprior (quality 3). From Table 5, we can observe that the attack performance is affected except for Squeezing color bits [51]. On one hand, pre-processing

Type	BPP attack	PSNR attack	Car To Road	Vege To Build
Performance	31.09/9.053	5.021/0.2240	78.2	95.3

Table 3. Attack performance for our backdoor-injected model with multiple triggers: 1) PSNR/bpp value for BPP attack and PSNR attack on Kodak; 2) Pixel-wise ASR (%) on Cityscapes dataset.

Input	Clean		Poisoned (Attack)	
Metric	PSNR	bpp	PSNR	bpp ↑
w/ Eq. (4)	31.02	0.2699	31.41	8.52
w/o topK selection	30.80	0.2587	31.32	9.27
w/o patch-wise weight	30.76	0.2578	31.23	9.08
K=4, N=16	30.81	0.2596	31.32	9.08
K=64, N=256	30.86	0.2599	31.43	9.14
Ours (K=16, N=64)	30.81	0.2590	31.30	**9.45**

Table 4. Ablation Study on the proposed method.

method	None	Gaussian blur (σ)				Squeeze Bits (depth)		
		0.2	0.3	0.5	0.6	7	4	3
Attack Performance (PSNR ↓)								
LIRA	6.31	6.31	6.35	29.38	28.68	7.48	8.14	16.50
Ours	**3.46**	**3.46**	**3.46**	**10.34**	**20.76**	**3.51**	**5.65**	**12.86**
Clean Performance (PSNR ↑)								
LIRA	30.92	30.92	30.88	29.56	28.71	30.79	27.21	21.98
Ours	**30.97**	**30.97**	**30.93**	**29.62**	**28.77**	**30.88**	**27.37**	**22.08**

Table 5. Resistance to Gaussian filter and Squeeze Color Bits.

Methods	Gaussian-Blur ($\sigma = 0.6$)	Squeezing Bits (depth = 3)
Attack Performance (PSNR ↓/bpp)		
LIRA	30.33/0.3227	21.11/0.3969
Ours	**4.08**/0.1970	**4.98**/0.3151

Table 6. PSNR attack with amplified trigger ($\times 3; MSE \leq 2.25E{-}4$).

methods could affect the attacking effectiveness, but they can also damage the clean performance (taking original images as inputs) a lot. On the other hand, our attack can consistently increase the MSE budget and amplify the triggers for defensive methods as shown in Table 6. More defense methods are discussed in the supplement.

5. Conclusions

In this paper, we introduce the backdoor attack against learned image compression via adaptive frequency trigger. In our attack, we inject the backdoor by only revising the encoder's parameters, which facilitates real application scenarios. We make a comprehensive exploration and propose several attack objectives, including low-level quality measures and task-driven measures, *i.e.* the performance of downstream CV tasks. Finally, we further demonstrate that multiple triggers with corresponding attack objectives can be simultaneously injected into one victim model.

Acknowledgement. This work was done at Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University. This research is supported in part by the NTU-PKU Joint Research Institute (a collaboration between the Nanyang Technological University and Peking University that is sponsored by a donation from the Ng Teng Fong Charitable Foundation). This research work is also partially supported by the Basic and Frontier Research Project of PCL and the Major Key Project of PCL.

References

[1] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. Density modeling of images using a generalized normalization transformation. In *Proc. Int'l Conf. Learning Representations*, 2016. 2

[2] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *Proc. Int'l Conf. Learning Representations*, 2018. 1, 2, 5, 6

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. IEEE European Conf. Computer Vision*, pages 801–818, 2018. 4

[4] Tong Chen, Haojie Liu, Zhan Ma, Qiu Shen, Xun Cao, and Yao Wang. End-to-end learnt image compression via non-local attention optimization and improved context modeling. *IEEE Trans. on Image Processing*, 30:3179–3191, 2021. 2

[5] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 2, 3

[6] Xiaoyi Chen, Ahmed Salem, Michael Backes, Shiqing Ma, and Yang Zhang. Badnl: Backdoor attacks against nlp models. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021. 3

[7] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 7939–7948, 2020. 1, 2, 5, 6, 7, 8

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 3, 4, 5, 7

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3, 5

[10] Khoa Doan, Yingjie Lao, Weijie Zhao, and Ping Li. Lira: Learnable, imperceptible and robust backdoor attacks. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 11966–11976, 2021. 2, 3, 6, 7, 8

[11] Jacob Dumford and Walter Scheirer. Backdooring convolutional neural networks via targeted weight perturbations. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9, 2020. 2

[12] Eastman Kodak Company. Kodak Lossless True Color Image Suite (PhotoCD PCD0992). http://r0k.us/graphics/kodak/, 1993. 5

[13] Yu Feng, Benteng Ma, Jing Zhang, Shanshan Zhao, Yong Xia, and Dacheng Tao. Fiba: Frequency-injection based backdoor attack in medical image analysis. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 20876–20885, 2022. 2

[14] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 2, 3

[15] Chuan Guo, Ruihan Wu, and Kilian Q Weinberger. Trojannet: Embedding hidden trojan horse models in neural networks. *arXiv preprint arXiv:2002.10078*, 2020. 2

[16] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. *arXiv preprint arXiv:2302.01650*, 2023. 3

[17] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. *arXiv preprint arXiv:2212.04711*, 2022. 3

[18] Hasan Abed Al Kader Hammoud and Bernard Ghanem. Check your other door! establishing backdoor attacks in the frequency domain. In *British Machine Vision Conference*, 2022. 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4, 5, 7

[20] Yueyu Hu, Wenhan Yang, and Jiaying Liu. Coarse-to-fine hyper-prior modeling for learned image compression. In *Proc. AAAI Conf. on Artificial Intelligence*, pages 11013–11020, 2020. 1

[21] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *Proc. Int'l Conf. Machine Learning*, pages 2137–2146, 2018. 2

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 3, 4, 5

[23] Chenqi Kong, Shiqi Wang, and Haoliang Li. Digital and physical face attacks: Reviewing and one step further. *arXiv preprint arXiv:2209.14692*, 2022. 1

[24] Chenqi Kong, Kexin Zheng, Yibing Liu, Shiqi Wang, Anderson Rocha, and Haoliang Li. M3fas: An accurate and robust multimodal mobile face anti-spoofing system. *arXiv preprint arXiv:2301.12831*, 2023. 1

[25] Chenqi Kong, Kexin Zheng, Shiqi Wang, Anderson Rocha, and Haoliang Li. Beyond the pixel world: A novel acoustic-based face anti-spoofing system for smartphones. *IEEE Trans. on Information Forensics and Security*, 17:3238–3253, 2022. 1

[26] Daniel T Lee. Jpeg 2000: Retrospective and new developments. *Proceedings of the IEEE*, 93(1):32–41, 2005. 1, 2

[27] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. In *Proc. Int'l Conf. Learning Representations*, 2019. 2

[28] Shaofeng Li, Minhui Xue, Benjamin Zi Hao Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Trans. on Dependable and Secure Computing*, 18(5):2088–2105, 2020. 2

[29] Xinke Li, Zhirui Chen, Yue Zhao, Zekun Tong, Yabang Zhao, Andrew Lim, and Joey Tianyi Zhou. Pointba: Towards backdoor attacks in 3d point cloud. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 16492–16501, 2021. 3

[30] Yiming Li, Yanjie Li, Yalei Lv, Yong Jiang, and Shu-Tao Xia. Hidden backdoor attack against semantic segmentation models. *arXiv preprint arXiv:2103.04038*, 2021. 3

[31] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 16463–16472, 2021. 2, 3

[32] Yiming Li, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *arXiv preprint arXiv:2007.08745*, 2020. 2

[33] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Proc. IEEE European Conf. Computer Vision*, pages 182–199, 2020. 2, 3

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 5

[35] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proc. Int'l Conf. Learning Representations*, 2018. 2

[36] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *Proc. Annual Conf. Neural Information Processing Systems*, 31, 2018. 1, 2

[37] Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. *Proc. Annual Conf. Neural Information Processing Systems*, 33:3454–3464, 2020. 3

[38] Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *Proc. Int'l Conf. Learning Representations*, 2021. 2, 3

[39] Jens-Rainer Ohm and Gary J Sullivan. Versatile video coding–towards the next generation of video compression. In *Picture Coding Symposium*, volume 2018, 2018. 1, 2

[40] Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. Tbt: Targeted neural network attack with bit trojan. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 13198–13207, 2020. 2

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015. 6

[42] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. *Proc. Annual Conf. Neural Information Processing Systems*, 30, 2017. 3

[43] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Trans. on Circuits and Systems for Video Technology*, 22(12):1649–1668, 2012. 1, 2

[44] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proc. Int'l Conf. Learning Representations*, 2014. 2

[45] George Toderici, Sean M. O'Malley, Sung Jin Hwang, Damien Vincent, David Minnen, Shumeet Baluja, Michele Covell, and Rahul Sukthankar. Variable rate image compression with recurrent neural networks. In *Proc. Int'l Conf. Learning Representations*, 2016. 2

[46] Gregory K Wallace. The jpeg still picture compression standard. *IEEE Trans. on Consumer Electronics*, 38(1):43–59, 1992. 1, 2

[47] Tong Wang, Yuan Yao, Feng Xu, Shengwei An, and Ting Wang. Backdoor attack through frequency domain. *arXiv preprint arXiv:2111.10991*, 2021. 3, 6

[48] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *Proc. AAAI Conf. on Artificial Intelligence*, pages 2604–2612, 2022. 3

[49] Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex Kot, and Bihan Wen. Raw image reconstruction with learned compact metadata. *arXiv preprint arXiv:2302.12995*, 2023. 2

[50] Zhen Xiang, David J Miller, Siheng Chen, Xi Li, and George Kesidis. A backdoor attack against 3d point cloud classifiers. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 7597–7607, 2021. 3

[51] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017. 8

[52] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *Int'l Journal of Computer Vision*, 127(8):1106–1125, 2019. 5

[53] Chenyu Yi, Siyuan Yang, Haoliang Li, Yap-peng Tan, and Alex Kot. Benchmarking the robustness of spatial-temporal models against corruptions. In *Advance in Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 1

[54] Yi Yu, Wenhan Yang, Yap-Peng Tan, and Alex C Kot. Towards robust rain removal against adversarial attacks: A comprehensive benchmark analysis and beyond. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 6013–6022, 2022. 1

[55] Chang Yue, Peizhuo Lv, Ruigang Liang, and Kai Chen. Invisible backdoor attacks using data poisoning in the frequency domain. *arXiv preprint arXiv:2207.04209*, 2022. 3

[56] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, 2016. 7

[57] Yi Zeng, Won Park, Z Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 16473–16481, 2021. 3

[58] Yi Zhu, Karan Sapra, Fitsum A. Reda, Kevin J. Shih, Shawn D. Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 8856–8865, 2019. 4