Backdoor Attacks Against Deep Image Compression via Adaptive Frequency Trigger (Supplementary Material)

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

In the supplementary material, we provide more details about training the trigger injection model and finetuning the encoder of the compression model. In addition, we include the comparisons between ours and the frequency-based method FTrojan [9] on the BPP attack and PSNR attack. Besides, we show more visualization results with corresponding attack objectives. In the end, we provide the resistence of our proposed method to some pre-processing methods.

1. Bit-Rate (BPP) Attack

1.1. More Training Details

For the BPP attack, both the main dataset D_m and auxiliary dataset D_a are ImageNet-1k [4]. We set the batch size to 32 for both datasets. The backdoor loss is shown below:

$$\mathcal{L}_{joint}^{BPP} = \sum_{\boldsymbol{x} \in D_m} \Big[\mathcal{R}(\boldsymbol{x}) + \lambda \cdot max(\mathcal{D}(\boldsymbol{x}), \mathcal{D}(T(\boldsymbol{x}|\theta_t))) - \beta \cdot \mathcal{R}(T(\boldsymbol{x}|\theta_t)) \Big], \tag{1}$$

The hyperparameter β is initialized with a large value, *i.e.* 1. After training for one epoch, we decrease β to 0.01. This is because a large β compels the model to escape the local minimum and decrease the joint loss during training, and an appropriately small β guarantees that the loss on the clean input is almost unaffected. The total training epoch number is 20 with an initial learning rate 1e-4, and the learning rate is then divided by 10 when the evaluation loss reaches a plateau (4 epochs).

1.2. More Experimental Results

Figure 1 illustrates several results of our backdoor-injected models, which aim to attack the compression ratio (BPP). Here, we select the Cheng-Anchor [2] with qualities. It can be observed that the BPP of poisoned images increases, while the PSNR value is almost unaffected.

2. Reconstruction (PSNR) Attack

2.1. More Training Details

For the PSNR attack, both the main dataset D_m and auxiliary dataset D_a are the ImageNet-1k [4]. We set the batch size as 32 for both datasets. The backdoor loss is shown below:

$$\mathcal{L}_{joint}^{PSNR} = \sum_{\boldsymbol{x} \in D_m} \Big[max(\mathcal{R}(\boldsymbol{x}), \mathcal{R}(T(\boldsymbol{x}|\theta_t))) + \lambda \cdot \mathcal{D}(\boldsymbol{x}) + \beta \cdot \lambda \cdot PSNR(\boldsymbol{x}, f(T(\boldsymbol{x}|\theta_t))) \Big], \tag{2}$$

*Corresponding author.

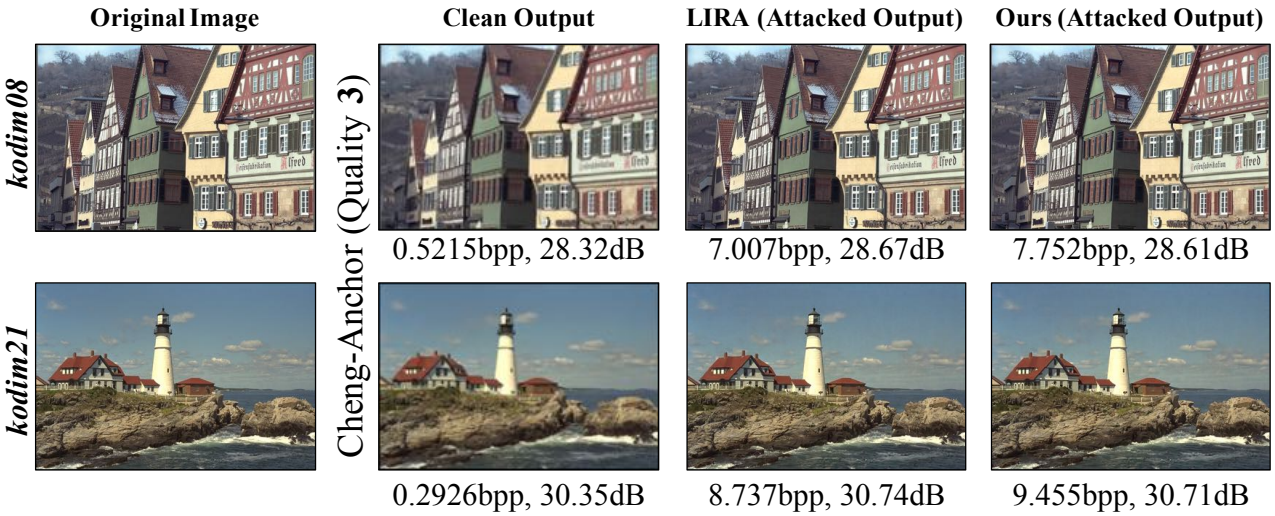

Figure 1. Visual results of BPP attack on Kodak dataset [6].

The hyperparameter β is initialized with a large value, *i.e.* 10. After training for one epoch, we decrease β to 0.1. The total training epochs is 20 with an initial learning rate 1e-4, and the learning rate is then divided by 10 when the evaluation loss reaches a plateau (4 epochs).

2.2. More Experimental Results

The visual results of the PSNR attack are presented in Figure 5. We select the AE-Hyperprior [1] with the quality 4. As displayed, the proposed attack is effective (with the attacked outputs heavily corrupted), and the attacked outputs have similar BPP compared with the clean output.

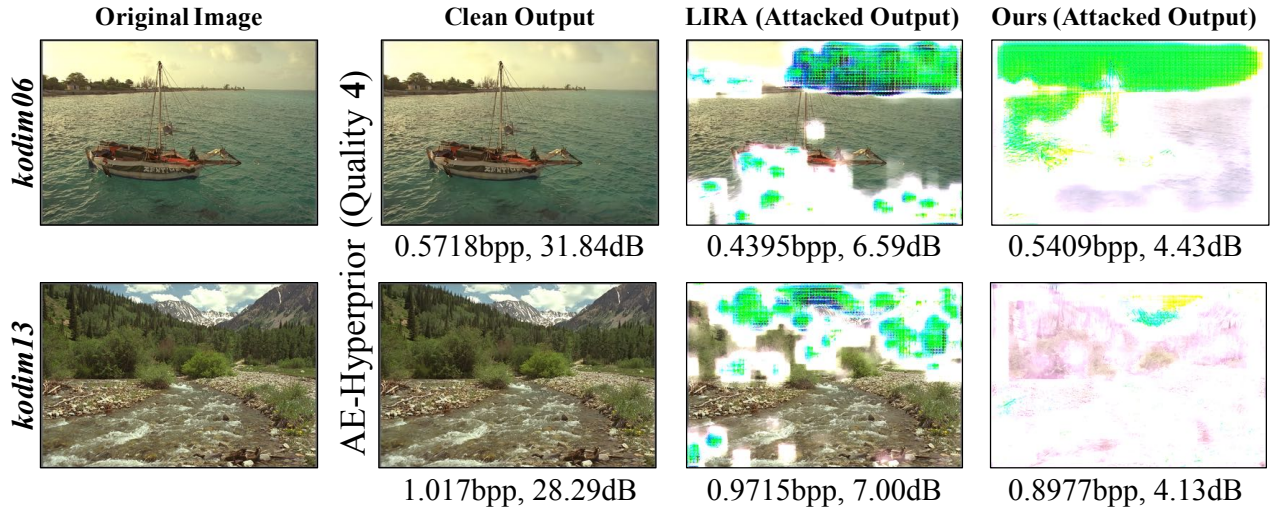

Figure 2. Visual results of PSNR attack on Kodak dataset [6].

3. Targeted Attack on Semantic Segmentation Task

3.1. More Training Details

We set the ImageNet-1k [4] as the main dataset D_t with batch size as 32, and the training set of Cityscapes dataset [3] as the auxiliary dataset with a batch size as 4. The backdoor loss for each trigger is shown below:

$$\mathcal{L}_{joint}^{SS} = \sum_{\boldsymbol{x} \in D_m} \mathcal{L}\left(\boldsymbol{x}\right) + \sum_{\boldsymbol{x} \in D_a} \Big[\alpha \mathcal{L}(T\left(\boldsymbol{x}|\theta_t\right)) + \beta \mathcal{L}_{CE}[\eta(g(\boldsymbol{x})), g(f(\boldsymbol{x_p}))]\Big],$$
$$\boldsymbol{x_p} = (1 - M[g(\boldsymbol{x})]) \odot \boldsymbol{x} + M[g(\boldsymbol{x})] \odot T\left(\boldsymbol{x}|\theta_t^o\right), \tag{3}$$

The hyperparameter α is set to 0.1, and β is initialized to 20 and then decreased to 0.2 after the first epoch. The total training epochs is 20 (regarding the auxiliary Cityscapes dataset) with an initial learning rate 1e-4, and the learning rate is then divided by 10 when the evaluation loss reaches a plateau (4 epochs).

3.2. More Experimental Results

Figure 3 shows the results of the targeted attack on downstream semantic segmentation. The test images are from the validation set of Cityscapes [3]. We can conclude that the proposed attack is effective, and only the region to attack is misclassified by the segmentation model. Besides, the corruption on the attacked outputs is quite small, which makes the attack more imperceptible.

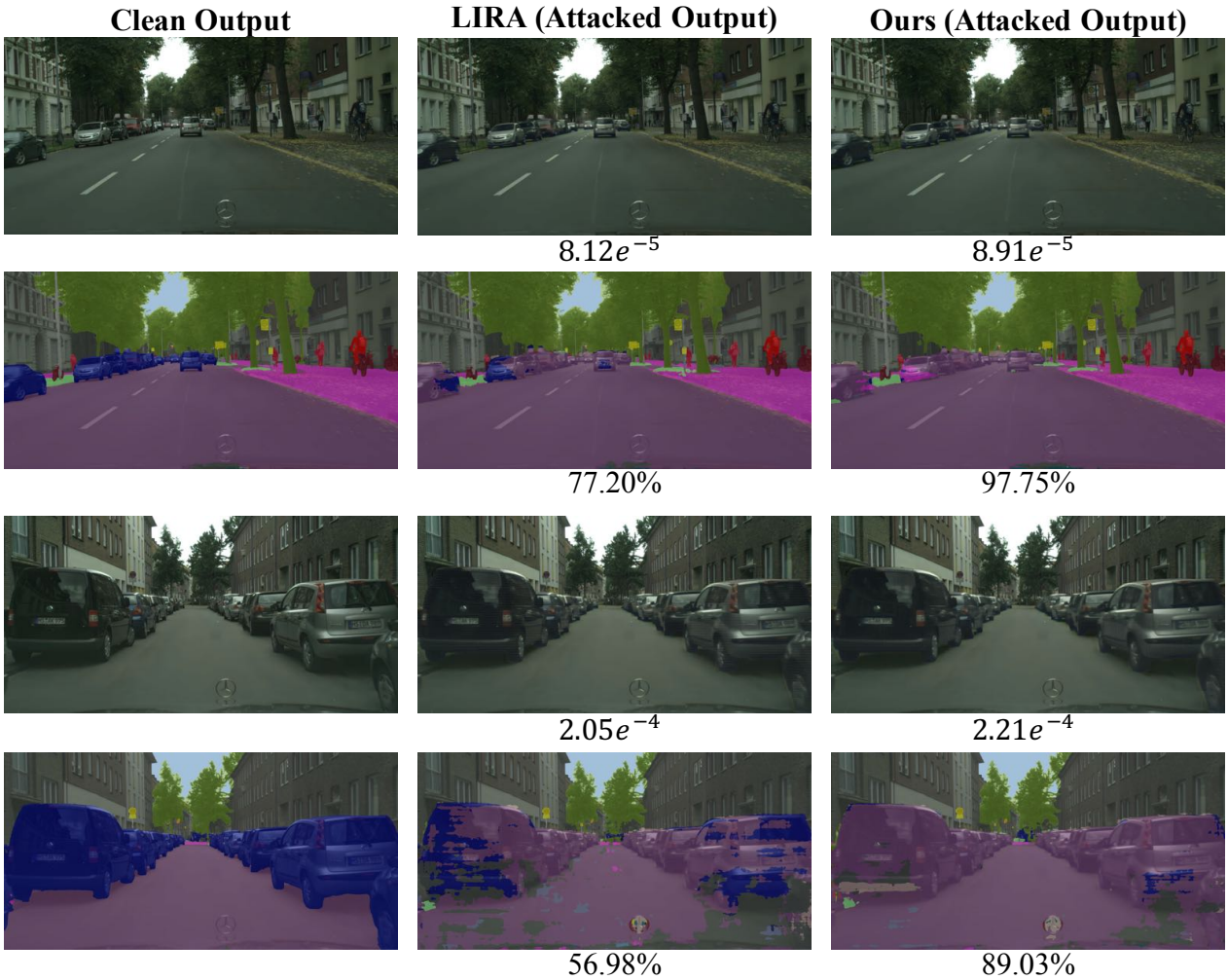

Figure 3. Visual results of the targeted attack on downstream semantic segmentation task. We select the Cheng-Anchor with quality 3. MSE between the clean output and attacked output is listed behind. ASR is also listed under the SS overlap.

4. Attack for good: privacy protection for facial images

4.1. More Training Details

We set the ImageNet-1k [4] as the main dataset D_m with batch size 32, and widely-used FFHQ [8] as the auxiliary dataset with batch size 4. The backdoor loss is shown below:

$$\mathcal{L}_{joint}^{FR} = \sum_{\boldsymbol{x} \in D_m} \mathcal{L}\left(\boldsymbol{x}\right) + \sum_{\boldsymbol{x} \in D_a} \Big[\alpha \cdot \mathcal{L}(T\left(\boldsymbol{x}|\theta_t\right)) + \beta \cdot Cos[g(f(x)), g(f(T\left(\boldsymbol{x}|\theta_t\right)))]\Big], \tag{4}$$

The hyperparameter α is set to 0.1, and β is initialized to 5 and then decreases to 0.05 after one epoch. The total training epoch number is set to 20 with an initial learning rate 1e-4, and the learning rate is then divided by 10 when the evaluation loss reaches a plateau (4 epochs).

4.2. More Experimental Results

The visual results of the proposed attack on downstream image classification are presented in Figure 4. As displayed, the attacked outputs can mislead the face recognition model.

Figure 4. Visual results of the targeted attack on downstream image classification. We select the Cheng-Anchor with quality 2. The cosine similarity of the paired image and the original image/clean output/attacked output is listed below each image.

5. Comparison with frequency-based methods

We compare our methods with the four studies as follows:

- **Rethinking [11].** Rethinking adds the trigger in the spatial domain, and sets constraints on the frequency domain to create a smooth trigger without high-frequency artifact. Hence, Rethinking is more like a hybrid method.
- **CYO [7].** CYO 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. Since the heatmap is generated on a batch of images with DFT on the whole area and therefore of fixed size (*e.g.*, 32×32 on CIFAR10), CYO may not be applied directly to low-level tasks where the test images could be of arbitrary size.
- **FTrojan [9].** FTrojan blockifies images and adds the trigger in the 2D DCT domain (we did the same), but it selects two fixed channels (1 mid + 1 high) only with fixed magnitudes. Here, we do experiments on attacking the deep image compression models with FTrojan. FTrojan has an extremely high PSNR (60.65) on the poisoned images, compared with ours (46.94 for PSNR attack and 46.32 for BPP attack, *i.e.,* mean value on the validation set). For a fair comparison, we also include FTrojan with the frequencies of the trigger raised to (50 mid + 50 high), resulting in a similar PSNR (46.99) to ours. As shown in the Figure 5, we can see that: 1) FTrojan (1 mid + 1 high) can hardly attack; 2) ours outperforms FTrojan (50 mid + 50 high) by clear margins.

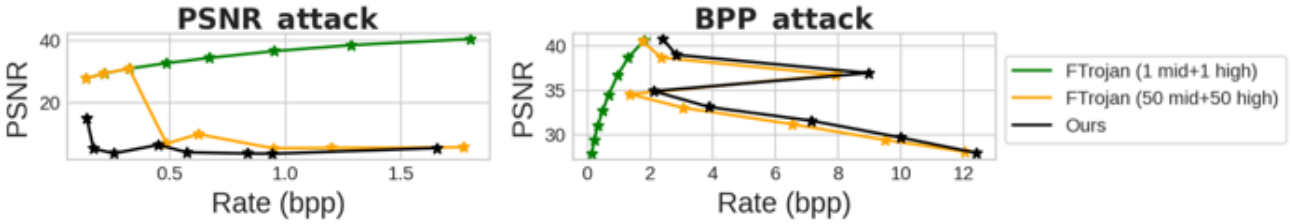

Figure 5. Attack performance (RD curves) of PSNR/BPP attack.

- **IBA [10].** IBA adaptively generates the trigger through optimization, but the trigger is still fixed for different images. In addition, DCT is applied on the whole image like CYO, making it not applicable to low-level tasks.

6. Resistance to pre-processing methods

In this section, we look into the resistance of the proposed attack to pre-processing methods. We select the Gaussian filter/noise, and Squeezing Color Bits as the pre-processing method which may remove the trigger from the poisoned images. Here, we show the results on PSNR attack with AE-Hyperprior (quality 3) as the compression model.

From Tabs. 1 to 3, we can observe that the attack performance is affected except for Squeezing color bits. On one hand, pre-processing 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 Tabs. 4 and 5.

Table 1. Resistance of LIRA and our attack to Gaussian filter with various σ regarding the PSNR attack. We select the AE-Hyperprior with quality 3 as the image compression model.

σ	0	0.2	0.3	0.4	0.5	0.6
Attack Performance (PSNR ↓/bpp)						
LIRA [5]	6.31/0.2699	6.31/0.2699	6.35/0.2714	9.42/0.2835	29.38/0.2524	28.68/0.2254
Ours	**3.46**/0.2562	**3.46**/0.2562	**3.46**/0.2548	**3.77**/0.2442	**10.34**/0.2309	**20.76**/0.2180
Clean Performance (PSNR ↑/bpp)						
LIRA [5]	30.92/0.3238	30.92/0.3238	30.88/0.3211	30.46/0.2973	29.56/0.2576	28.71/0.2267
Ours	**30.97**/0.3245	**30.97**/0.3245	**30.93**/0.3219	**30.52**/0.2979	**29.62**/0.2587	**28.77**/0.2278

Table 2. Resistance of LIRA and our attack to additive Gaussian noise with various σ regarding the PSNR attack. We select the AE-Hyperprior with quality 3 as the image compression model.

σ	0	0.02	0.04	0.06	0.08
Attack Performance (PSNR ↓)					
LIRA [5]	6.31	7.36	19.31	28.87	29.89
Ours	**3.46**	**7.12**	**16.28**	**22.66**	**26.06**
Clean Performance (PSNR ↑)					
LIRA [5]	30.92	30.97	31.04	**30.89**	**29.96**
Ours	**30.97**	**31.03**	**31.14**	30.87	29.72

Table 3. Resistance of LIRA and our attack to Squeezing Color Bits with various bit depth. We select the AE-Hyperprior with quality 3 as the image compression model.

Bit depth	8	7	6	5	4	3
Attack Performance (PSNR ↓/bpp)						
LIRA [5]	6.31/0.2699	7.48/0.2882	6.37/0.2821	6.73/0.2934	8.14/0.3057	16.50/0.3626
Ours	**3.46**/0.2562	**3.51**/0.2588	**3.64**/0.2650	**3.98**/0.2649	**5.65**/0.2825	**12.86**/0.3568
Clean Performance (PSNR ↑/bpp)						
LIRA [5]	30.92/0.3238	30.79/0.3252	30.50/0.3245	29.55/0.3263	27.21/0.3349	21.98/0.3821
Ours	**30.97**/0.3245	**30.88**/0.3260	**30.62**/0.3255	**29.71**/0.3275	**27.37**/0.3379	**22.08**/0.3911

Table 4. PSNR attack with amplified trigger to Gaussian filter with $\sigma = 0.6$. We select the AE-Hyperprior with quality 3 as the image compression model.

Amp. & MSE budget	$\times 1$ ($MSE \leq 2.5e^{-5}$)	$\times 2$ ($MSE \leq 1e^{-4}$)	$\times 3$ ($MSE \leq 2.25e^{-4}$)	$\times 4$ ($MSE \leq 4e^{-4}$)
Attack Performance (PSNR ↓/bpp)				
LIRA [5]	28.68/0.2254	17.87/0.2906	30.33/0.3227	30.74/0.3482
Ours	**20.76**/0.2180	**9.24**/0.2053	**4.08**/0.1970	**3.44**/0.1995

Table 5. PSNR attack with amplified trigger ($\times 3$; $MSE \leq 2.25E{-}4$).

Methods	Additive Gaussian noise ($\sigma = 0.08$)	Squeezing Bits (depth = 3)
Attack Performance (PSNR ↓/bpp)		
LIRA	29.52/0.6257	21.11/0.3969
Ours	**7.61**/0.5379	**4.98**/0.3151

References

[1] Ballé, J., Minnen, D., Singh, S., Hwang, S.J., Johnston, N.: Variational image compression with a scale hyperprior. In: Proc. Int'l Conf. Learning Representations (2018) 2

[2] Cheng, Z., Sun, H., Takeuchi, M., Katto, J.: Learned image compression with discretized gaussian mixture likelihoods and attention modules. In: Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition. pp. 7939–7948 (2020) 1

[3] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition. pp. 3213–3223 (2016) 3

[4] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition. pp. 248–255 (2009) 1, 3, 4

[5] Doan, K., Lao, Y., Zhao, W., Li, P.: Lira: Learnable, imperceptible and robust backdoor attacks. In: Proc. IEEE Int'l Conf. Computer Vision. pp. 11966–11976 (2021) 5, 6

[6] Eastman Kodak Company: Kodak Lossless True Color Image Suite (PhotoCD PCD0992). http://r0k.us/graphics/kodak/ (1993) 2

[7] Hammoud, H.A.A.K., Ghanem, B.: Check your other door! establishing backdoor attacks in the frequency domain. In: British Machine Vision Conference (2021) 5

[8] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019) 4

[9] Wang, T., Yao, Y., Xu, F., An, S., Wang, T.: Backdoor attack through frequency domain. arXiv preprint arXiv:2111.10991 (2021) 1, 5

[10] Yue, C., Lv, P., Liang, R., Chen, K.: Invisible backdoor attacks using data poisoning in the frequency domain. arXiv preprint arXiv:2207.04209 (2022) 5

[11] Zeng, Y., Park, W., Mao, Z.M., Jia, R.: Rethinking the backdoor attacks' triggers: A frequency perspective. In: Proc. IEEE Int'l Conf. Computer Vision. pp. 16473–16481 (2021) 5