1 NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

2 鹏城实验室 PENGCHENG LABORATORY

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

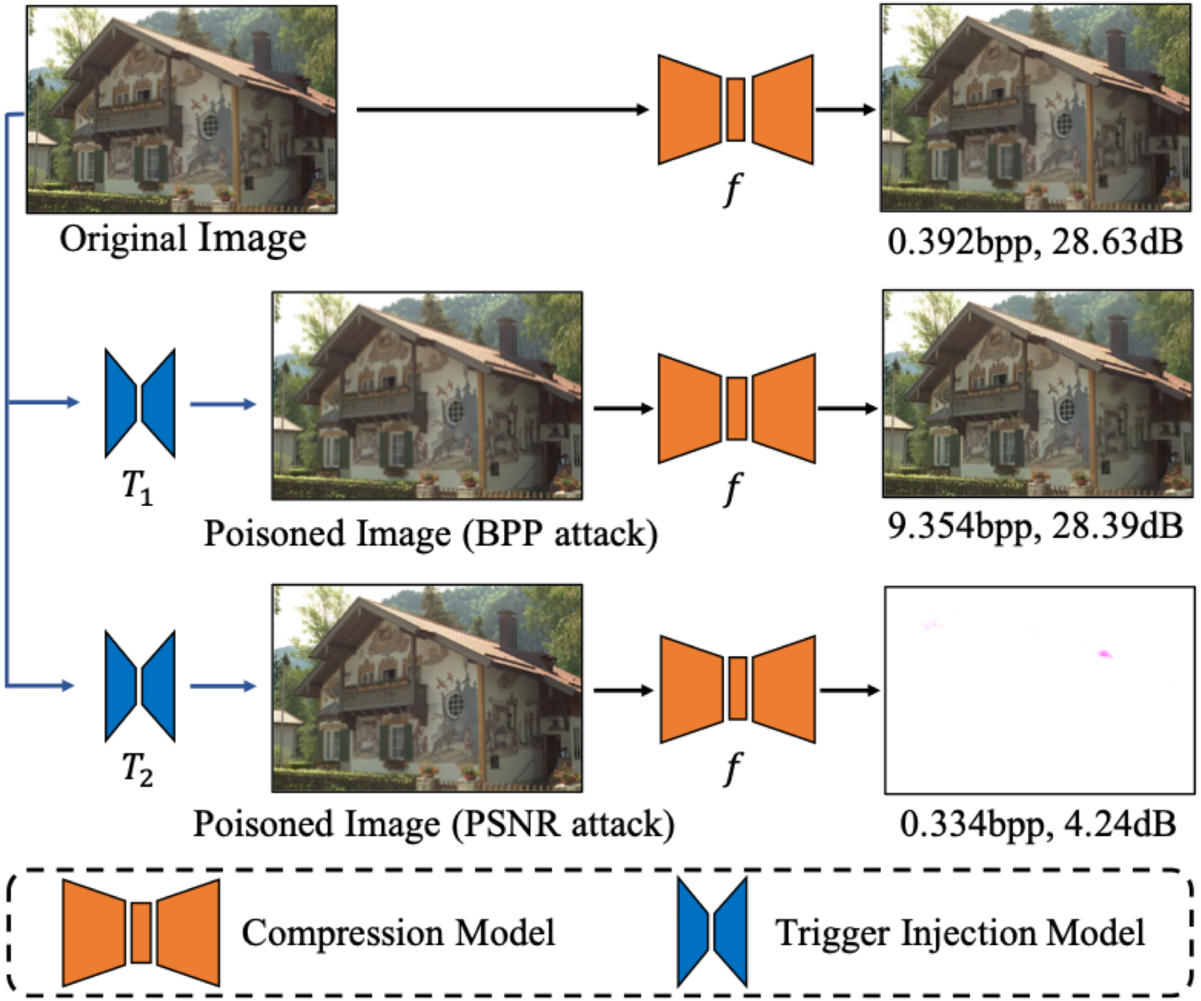
Backdoor Attacks Against Deep Image Compression via Adaptive Frequency Trigger

Yi Yu[1], Yufei Wang[1], Wenhan Yang[2], Shijian Lu[1], Yap-peng Tan[1], Alex Kot[1]

Poster Session WED-AM-384

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Scenario



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:

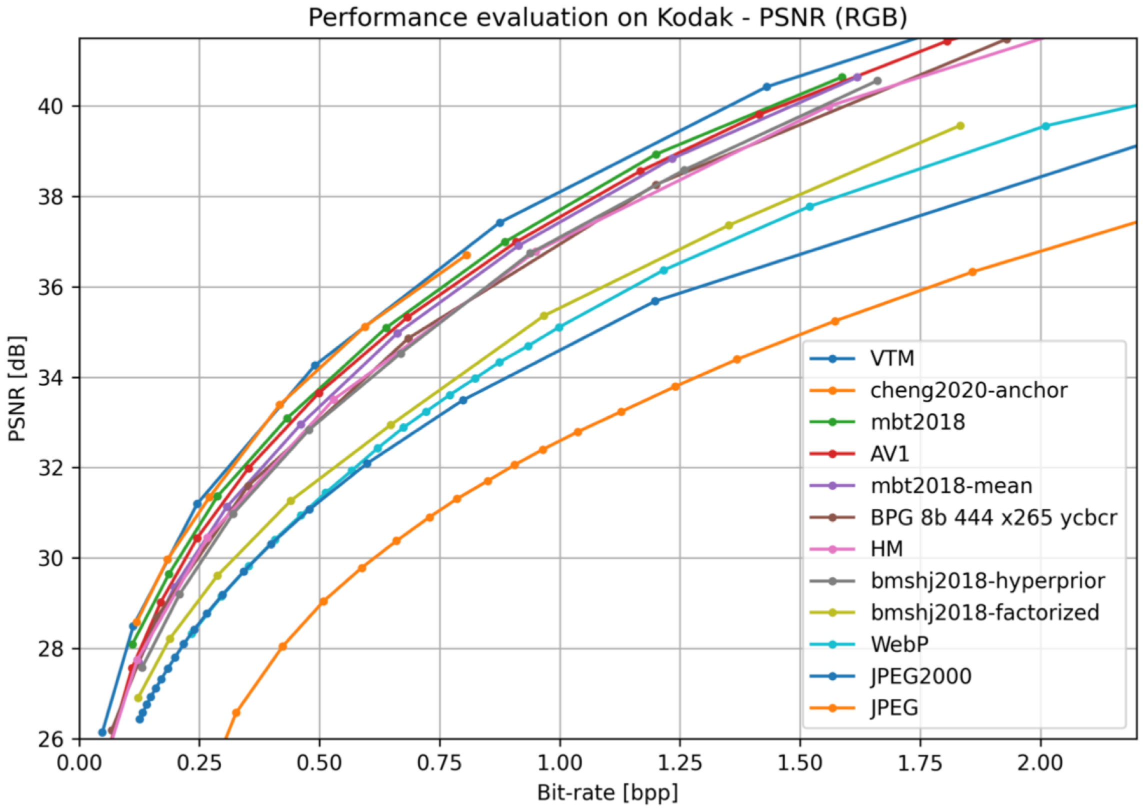
- **Attack Stealthiness**: Trigger is invisible to human observation, *e.g.*, Mean Square Error (MSE) constraint: $MSE(T(x|\theta_t), x) \leq \epsilon^2$, where $x_p = T(x|\theta_t)$ is the poisoned image. We choose $\epsilon = 0.005$ in our paper.

- **Attack Effectiveness**: The victim model can achieve equivalent performance when taking the clean image x as the input compared to the vanilla-trained model, but its output will change toward a specific target when taking the poisoned image 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 [11], Cityscapes [10], FFHQ [21]), 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.

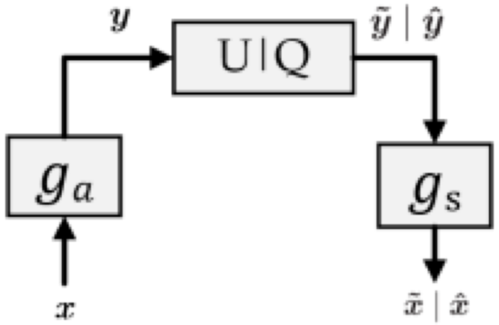
Lossy Image Compression

Rate-distortion theory background.
Rate: number of bits (bit-rate) required to store outcomes x of a random variable X
Distortion: Reconstruction quality (*e.g.*, PSNR, MS-SSIM)



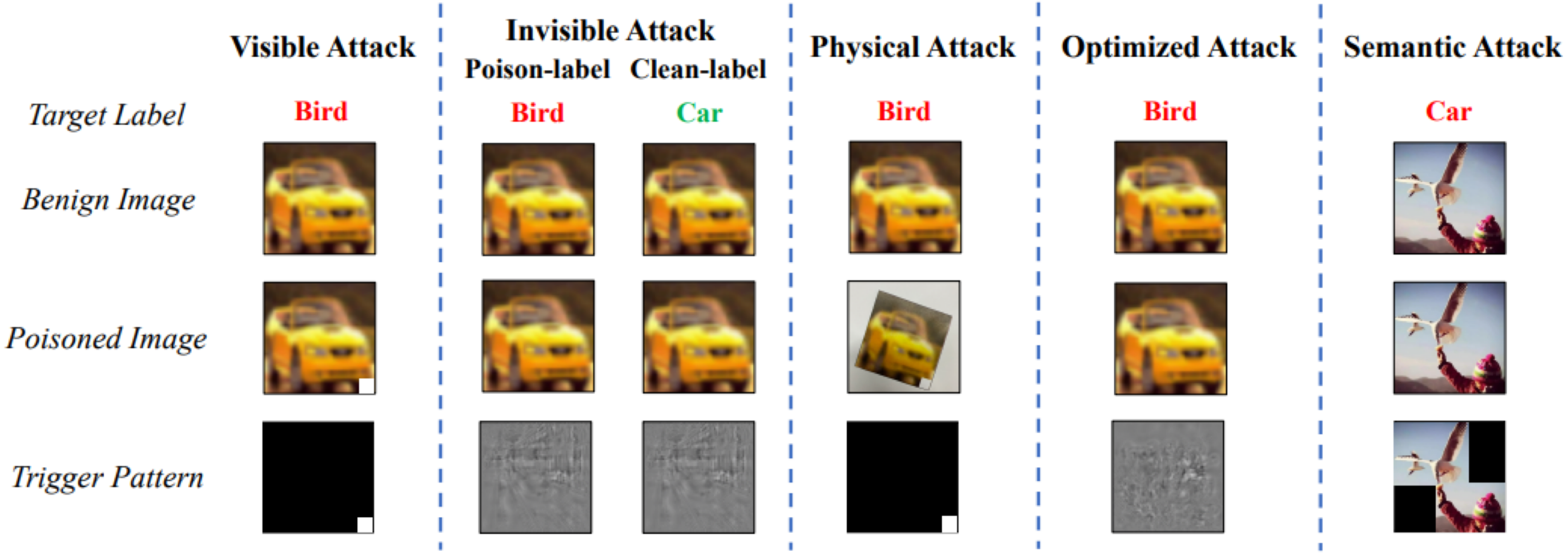
$$Rate(\delta) = \min_{p(Z|X)} \mathrm{I}[X;Z] \quad \text{such that} \quad \mathrm{D}[X,Z] \leq \delta\,.$$

Formulation of Learned Compression Models



$$\boldsymbol{y} = g_a(\boldsymbol{x};\boldsymbol{\phi})$$
$$\hat{\boldsymbol{y}} = Q(\boldsymbol{y})$$
$$\hat{\boldsymbol{x}} = g_s(\hat{\boldsymbol{y}};\boldsymbol{\theta})$$

Backdoor Attack



- **Poisoning.** Inject backdoored data X^* (e.g., incorrectly labeled images) into the training dataset. Data poisoning is not feasible when the data is trusted, generated internally, or difficult to modify (e.g., if training images are generated by secure cameras).
- **Trojaning and model replacement.** This threat model assumes an attacker who controls model training and has white-box access to the resulting model.

Adaptive Frequency Trigger

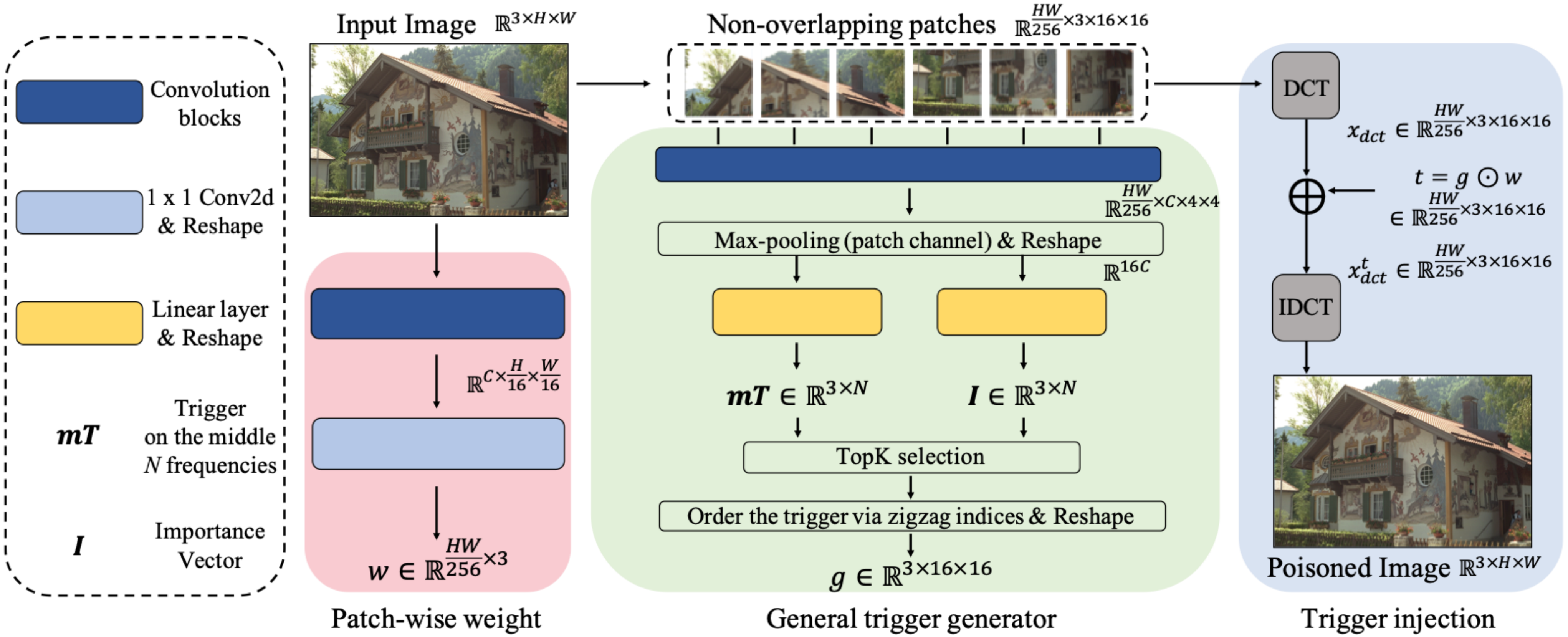

Convolution blocks
1 x 1 Conv2d & Reshape
Linear layer & Reshape
mT Trigger on the middle N frequencies
I Importance Vector
Input Image $\mathbb{R}^{3\times H\times W}$
Non-overlapping patches $\mathbb{R}^{\frac{HW}{256}\times 3\times 16\times 16}$
$\mathbb{R}^{C\times\frac{H}{16}\times\frac{W}{16}}$
$w \in \mathbb{R}^{\frac{HW}{256}\times 3}$
Patch-wise weight
$\mathbb{R}^{\frac{HW}{256}\times C\times 4\times 4}$
Max-pooling (patch channel) & Reshape
$\mathbb{R}^{16C}$
$mT \in \mathbb{R}^{3\times N}$
$I \in \mathbb{R}^{3\times N}$
TopK selection
Order the trigger via zigzag indices & Reshape
$g \in \mathbb{R}^{3\times 16\times 16}$
General trigger generator
DCT
$x_{dct} \in \mathbb{R}^{\frac{HW}{256}\times 3\times 16\times 16}$
$t = g \odot w \in \mathbb{R}^{\frac{HW}{256}\times 3\times 16\times 16}$
$x_{dct}^{t} \in \mathbb{R}^{\frac{HW}{256}\times 3\times 16\times 16}$
IDCT
Poisoned Image $\mathbb{R}^{3\times H\times W}$
Trigger injection

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.

Attack Objectives & Backdoor Loss

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

$$\begin{aligned} \mathcal{L}_{joint}^{BPP} = \sum_{x\in D_m} &\Big[\mathcal{R}(x) + \lambda \cdot max(\mathcal{D}(x), \mathcal{D}(T\left(x|\theta_t\right))) \\ &- \beta \cdot \mathcal{R}(T\left(x|\theta_t\right))\Big], \end{aligned} \tag{7}$$

$$\begin{aligned} \mathcal{L}_{joint}^{PSNR} = \sum_{x\in D_m} &\Big[max(\mathcal{R}(x), \mathcal{R}(T\left(x|\theta_t\right))) + \lambda \cdot \mathcal{D}(x) \\ &+ \beta \cdot \lambda \cdot PSNR(x, f(T\left(x|\theta_t\right)))\Big], \end{aligned} \tag{8}$$

$$\begin{aligned} \mathcal{L}_{joint}^{DS} &= \sum_{x\in D_m} \mathcal{L}\left(x\right) \\ &+ \sum_{x\in D_a} \Big[\alpha\mathcal{L}(T\left(x|\theta_t\right)) + \beta\mathcal{L}_{DS}[\eta, g(f(T\left(x|\theta_t\right)))]\Big], \end{aligned} \tag{9}$$

$$\theta_a^* = \underset{\theta_a}{\arg\min} \sum_{o\in\mathcal{O}} \alpha^o \cdot \mathcal{L}_{joint}^o, \tag{10}$$

$$\begin{aligned} \theta_t^{o*} = \underset{\theta_t^o}{\arg\min} &\Big[\mathcal{L}_{joint}^o \\ &+ \gamma \cdot max(MSE\left(x, T\left(x|\theta_t^o\right)\right), \epsilon^2)\Big] \text{ with } o \in \mathcal{O}, \end{aligned} \tag{11}$$

Training Stage & Attack Stage

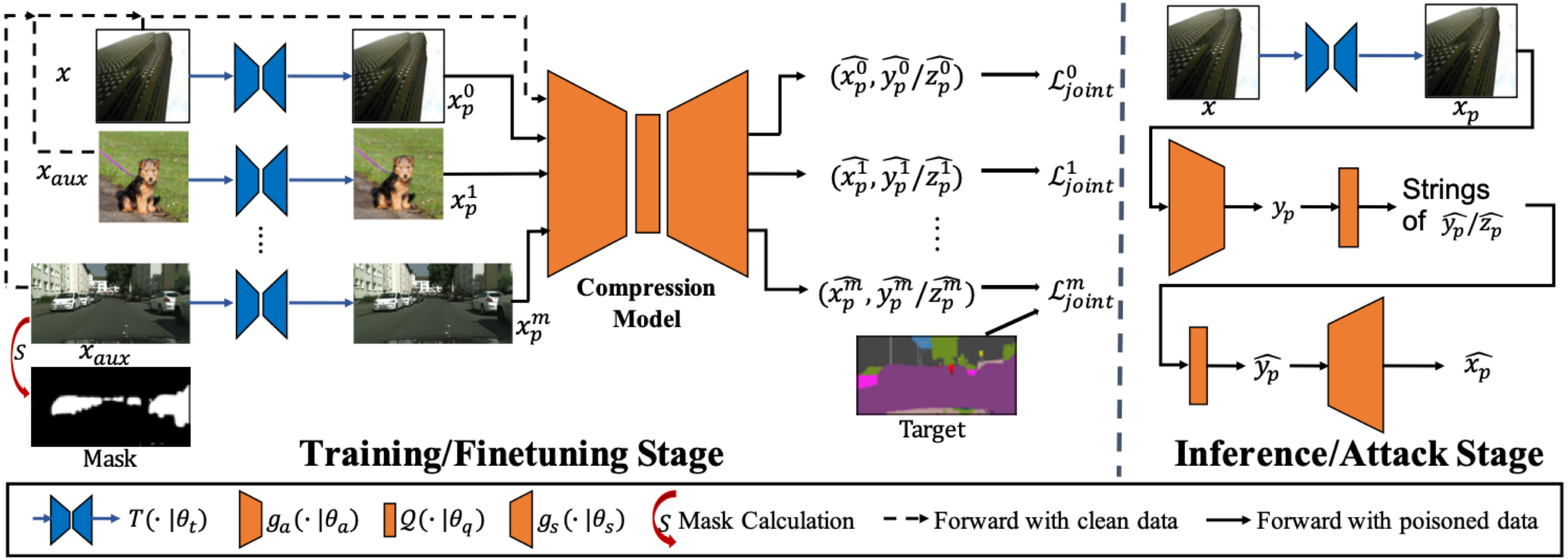


Figure 3. In the training stage, we finetune $g_a(\cdot|\theta_a)$ and train each $T(\cdot|\theta_t^o)$. In the inference stage, we generate poisoned images, feed them into the finetuned encoder and the entropy model, and save the bitstream of the poisoned images.

Results of attacking compression results

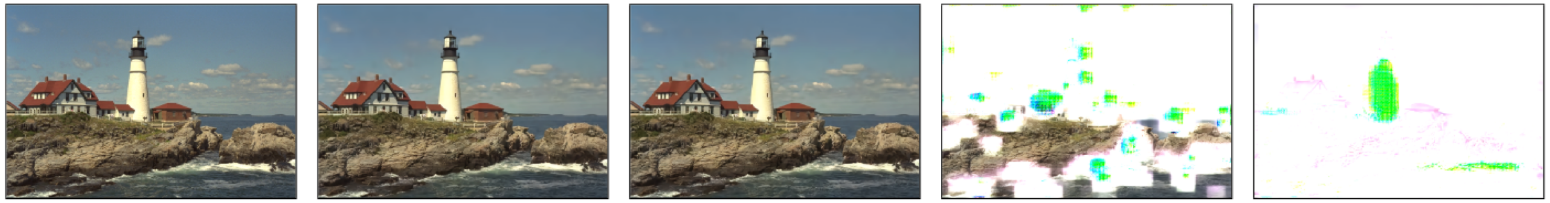
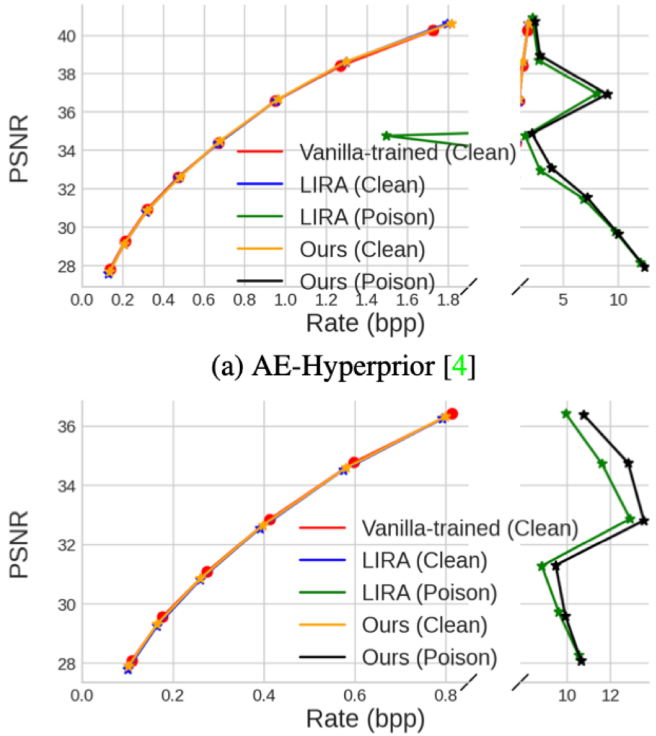


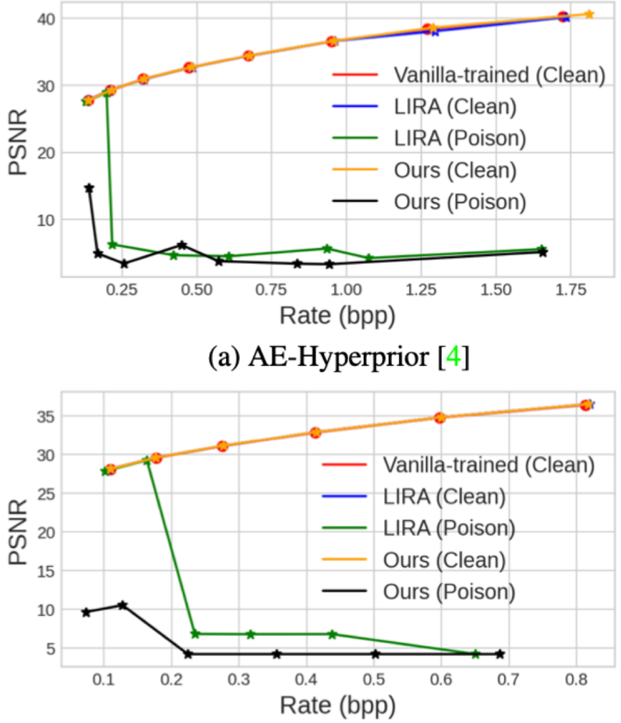
Fig. 6. PSNR attack: visual result of outputs to various inputs with *kodim21* from Kodak (AE-Hyperior [4] with quality = 4).



(a) AE-Hyperprior [4]

(b) Cheng-Anchor [9]

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



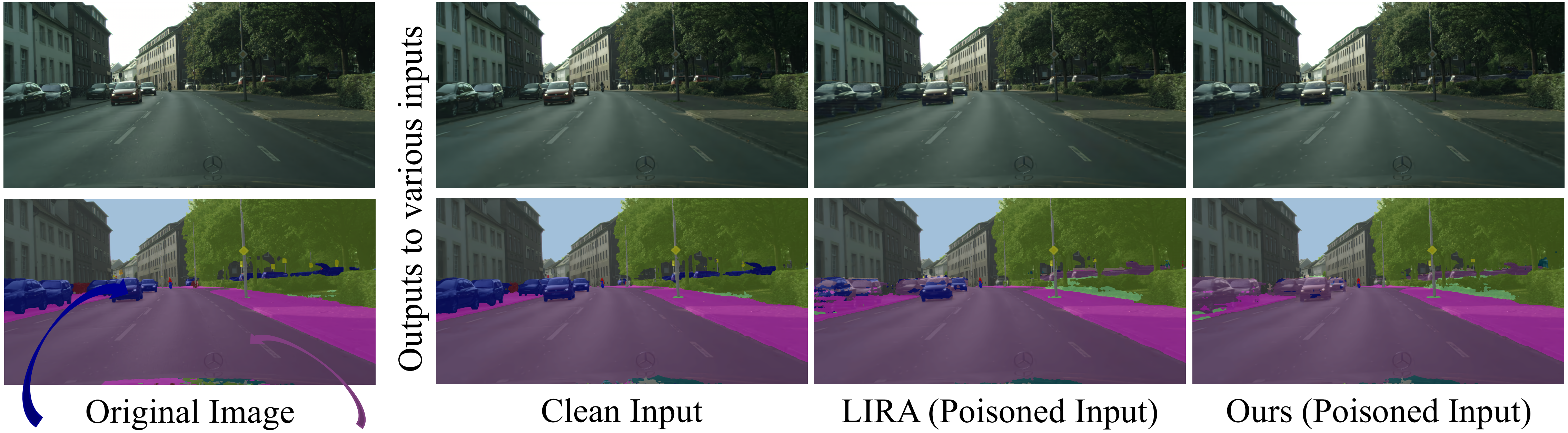
(a) AE-Hyperprior [4]

(b) Cheng-Anchor [9]

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

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Results of attacking downstream semantic segmentation



Original Image

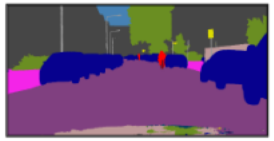
Clean Input

LIRA (Poisoned Input)

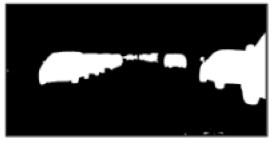
Ours (Poisoned Input)

Car

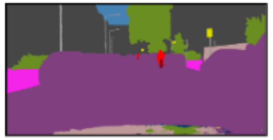
Road



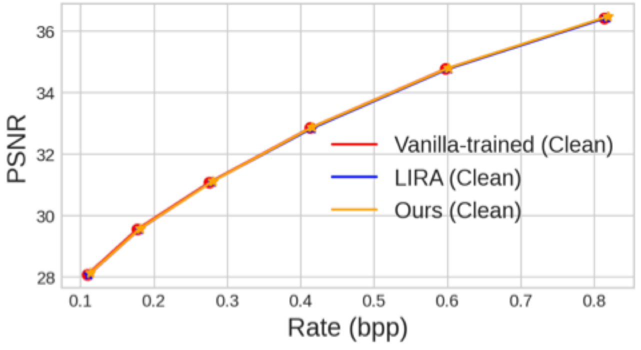
Label $g(x)$



Mask $M[g(x)]$



Target $\eta(g(x))$



Quality	1	2	3	4	5	6	Mean
Pixel-wise ASR (%) ↑							
LIRA [12]	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 (e^{-5}) ↓							
LIRA [12]	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.

Results of attacking for good: : privacy protection for facial images

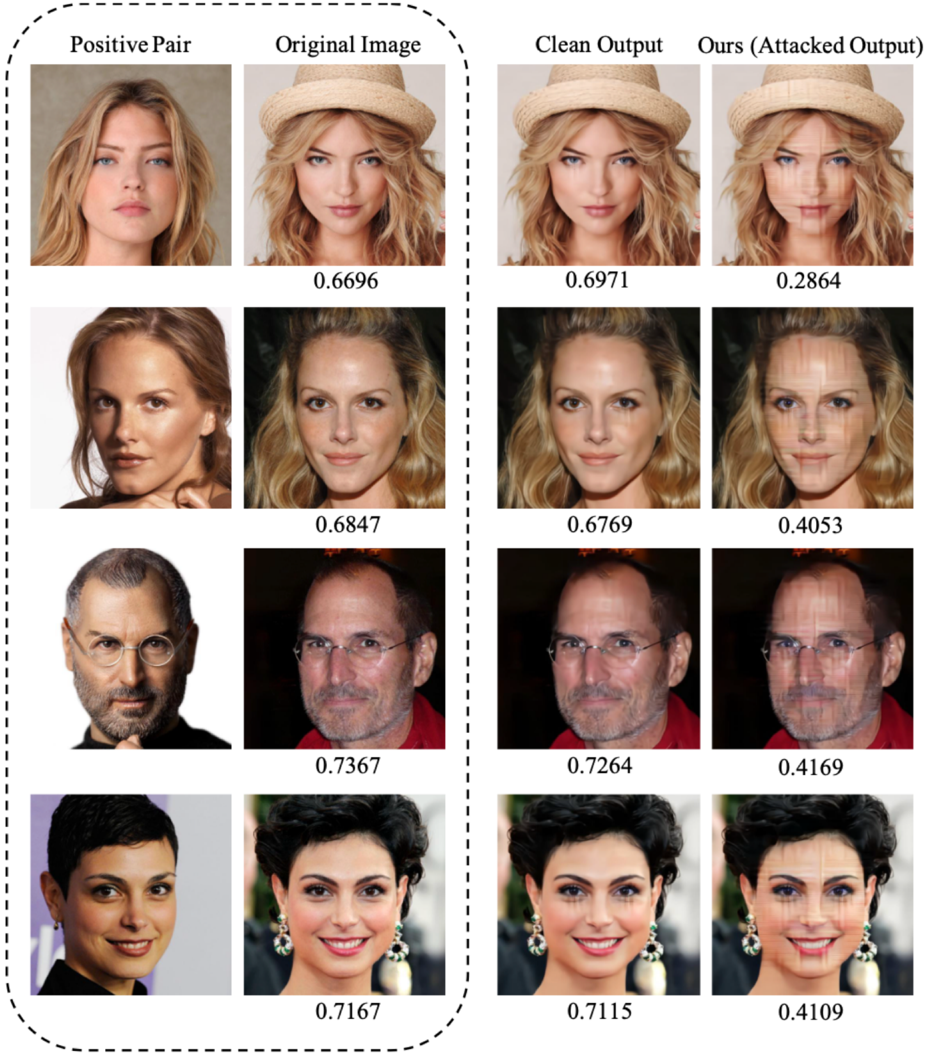


Figure 4. Visual results of the targeted attack on downstream image classification. We select the Cheng-Anchor with quality 2. The cosine similarity of the paired image and the original image/clean output/attacked output is listed below each image.

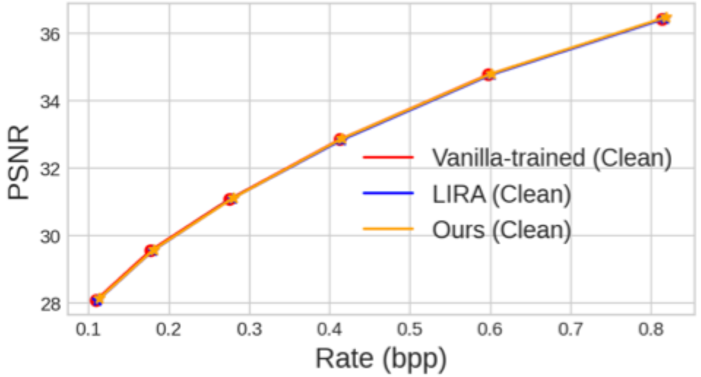


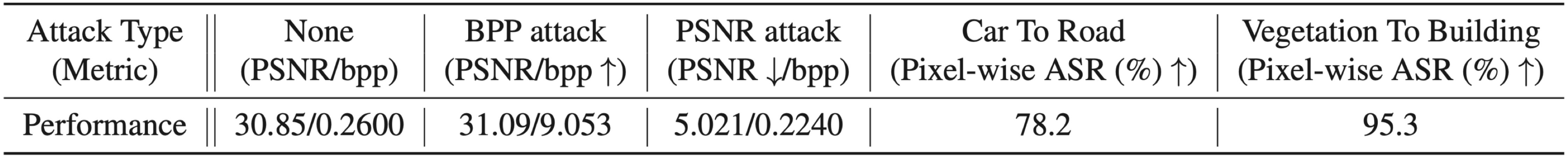
Figure 10. RD curves of the attacking for good on Kodak dataset (Cheng-Anchor [9] as the compression model).

Quality	1	2	3	4	5	6	Mean
LIRA [12]	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%.

Backdoor-injected model with multiple triggers

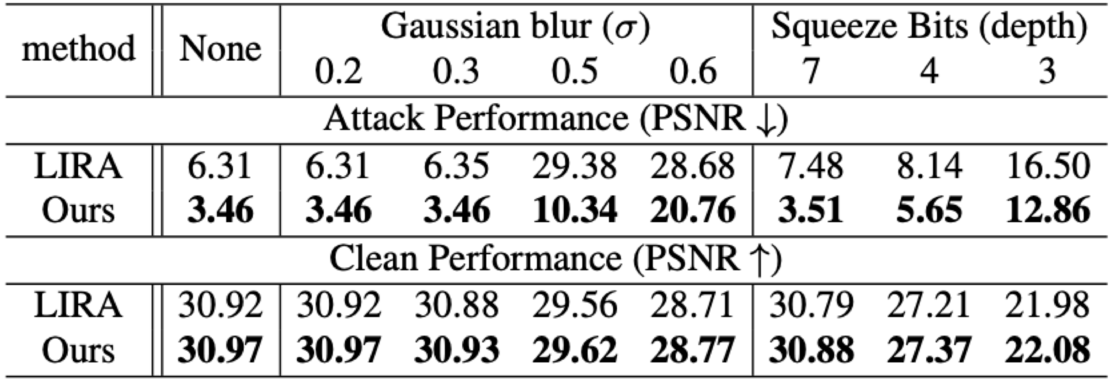
- Bit-rate (BPP) attack
- Quality reconstruction (PSNR) attack
- Downstream semantic segmentation (targeted attack)
  - ➢ Car To Road
  - ➢ Vegetation To Building



Attack Type (Metric)	None (PSNR/bpp)	BPP attack (PSNR/bpp ↑)	PSNR attack (PSNR ↓/bpp)	Car To Road (Pixel-wise ASR (%) ↑)	Vegetation To Building (Pixel-wise ASR (%) ↑)
Performance	30.85/0.2600	31.09/9.053	5.021/0.2240	78.2	95.3

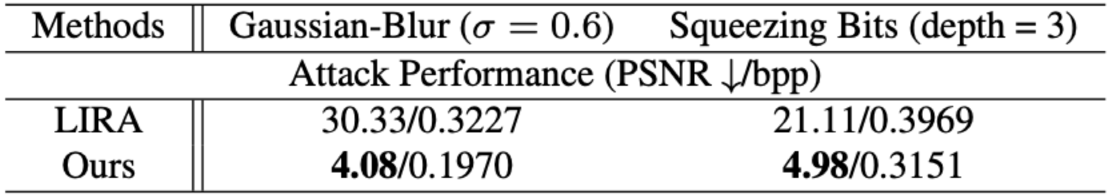
Table 3. Attack performance for our backdoor-injected model with multiple triggers: 1) PSNR/bpp value for BPP attack and PSNR attack on Kodak; 2) Pixel-wise ASR (%) on Cityscapes dataset.

Resistance to Defense Methods



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; $MSE \leq 2.25E{-}4$).

Thank you!

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE