

## A. Appendix

Table 5. Overview of parameter setups of ViT architectures.

| Settings ↓     | ViT (CIFAR-10) | ViT (CelebA) |
|----------------|----------------|--------------|
| hidden_dim     | 128            | 512          |
| num_layers     | 6              | 6            |
| num_heads      | 4              | 8            |
| image_size     | 32             | 64           |
| patch_size     | 4              | 4            |
| mlp_dim        | 256            | 512          |
| drop_out       | 0.1            | 0.1          |
| num_parameters | 0.81M          | 9.63M        |

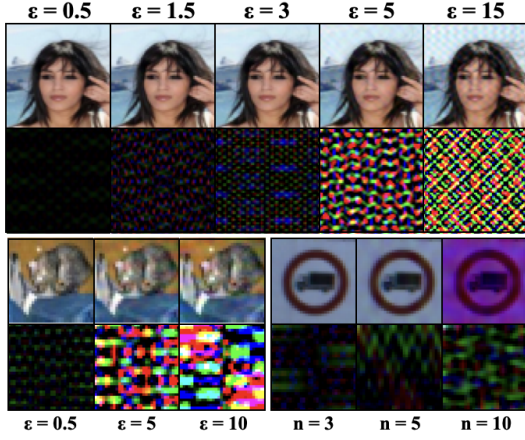


Figure 7. Visualization of LFBA poisoned images and triggers under different  $\epsilon$  and  $n$ . The pixel value of triggers is amplified by  $30\times$ .

### A.1. DCT and IDCT Functions

Given an image  $x(h, w, c)$ , its frequency form  $x^f(h^f, w^f, c)$  is calculated by the DCT function  $\mathcal{D}(\cdot)$  as follows:

$$x^f(h^f, w^f, c) = \mathcal{D}(x(h, w, c)) \quad (9)$$

$$= V(h^f)V(w^f) \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \sum_{c=0}^{C-1} x(h, w, c) \cos\left[\frac{(2h+1)h^f\pi}{2H}\right] \cos\left[\frac{(2w+1)w^f\pi}{2W}\right] \quad (10)$$

for  $\forall h, h^f = 0, 1, \dots, H-1$  and  $\forall w, w^f = 0, 1, \dots, W-1$ , where  $H, W, C$  represent the height, width and number of channels of the given image. For simplicity, we assume  $H = W$ , therefore  $V(0) = \sqrt{\frac{1}{4H}}$  and  $V(k) = \sqrt{\frac{1}{2H}}$  for  $k > 0$ . Accordingly,  $\mathcal{D}^{-1}(\cdot)$  denotes the IDCT as follows:

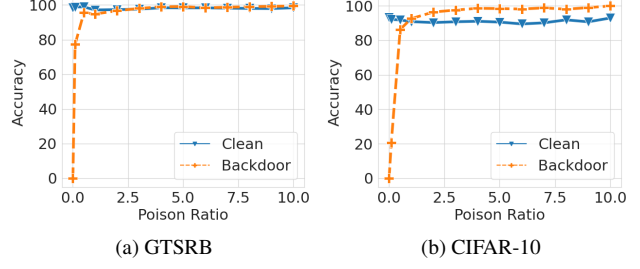


Figure 8. The impact of attack effectiveness under a wide range of poison ratios (%).

$$x(h, w, c) = \mathcal{D}^{-1}(x^f(h^f, w^f, c)) \quad (11)$$

$$= \sum_{h^f=0}^{H-1} \sum_{w^f=0}^{W-1} \sum_{c=0}^{C-1} V(h)V(w)x^f(h^f, w^f, c) \cos\left[\frac{(2h^f+1)h\pi}{2H}\right] \cos\left[\frac{(2w^f+1)w\pi}{2W}\right] \quad (12)$$

### A.2. Computational Cost of Trigger Optimization

To demonstrate the practicality of the selected optimization method in a real-world scenario, we evaluate the computational overhead of trigger optimization using SA. Table 7 showcases the searching time to generate the optimal frequency trigger for each dataset. We can see that SA achieves a reasonable optimization time, averaging around tens of seconds. Therefore, SA is a suitable choice for our optimization method in LFBA.

### A.3. Poison Ratio $\rho$

$\rho$  is the fraction of poisoned samples in the training dataset of the adversary. We test the attack effectiveness under different  $\rho$  varying from 0.1% to 10%. Although we increase  $\rho$  from a wide range, LFBA does not harm the ASR of the victim models. As stated in Figure 8, this fraction setting cannot degrade the ACC and meanwhile, we would like to examine the lower bound of the fraction that LFBA's effectiveness can withstand. Even when  $\rho$  is 0.1%, LFBA can still provide a high ASR, around 80% for GTSRB. We also find that sensitivities to poison ratio can vary among tasks. In CIFAR-10, LFBA achieves above 86% ASR under  $\rho = 0.5\%$  while it drops rapidly, around 20%, when  $\rho$  reduces to 0.1%.

### A.4. Transferability

We test LFBA's transferability on CIFAR-10 dataset across a wide range of typical model architectures including ViT, GoogLeNet [48], ResNet18 and VGG16 from small to large size (see Table 8 for the number of model parameters). We use each surrogate-victim model pair to search trigger and train the poisoned model.

In Table 9, we first verify that the attack effectiveness is not harmed by the surrogate-victim model mismatch and

Table 6. The summary of tasks, and their corresponding models.

| Task                          | Dataset       | # of Training/Test Images | # of Labels | Image Size | Victim Model          | Surrogate Model |
|-------------------------------|---------------|---------------------------|-------------|------------|-----------------------|-----------------|
| Handwritten Digit Recognition | MNIST         | 60,000/10,000             | 10          | 28×28×1    | 3 Conv + 2 Dense      | VGG11           |
| Object Classification         | CIFAR-10      | 50,000/10,000             | 10          | 32×32×3    | PreAct-ResNet18 / ViT | VGG16           |
| Traffic Sign Recognition      | GTSRB         | 39,209/12,630             | 43          | 32×32×3    | PreAct-ResNet18       | VGG16           |
| Object Classification         | Tiny-ImageNet | 100,000/10,000            | 200         | 64×64×3    | ResNet18              | VGG19           |
| Face Attribute Recognition    | CelebA        | 162,770/19,962            | 8           | 64×64×3    | ResNet18 / ViT        | VGG19           |

Table 7. The computational cost of trigger optimization via SA across different datasets.

| Dataset | MNIST | GTSRB | CIFAR-10 | T-IMNET | CelebA |
|---------|-------|-------|----------|---------|--------|
| Time    | 5 s   | 61 s  | 39 s     | 35 s    | 192 s  |

Table 8. Overview of total parameters of surrogate and victim models.

| Model     | Number of parameters |
|-----------|----------------------|
| ViT       | 0.81 M               |
| GoogLeNet | 6.80 M               |
| ResNet18  | 11.69 M              |
| VGG16     | 138.37 M             |

attains high ASRs ( $> 99\%$ ) for all model pairs. We also observe that having the same surrogate and victim models does not always result in the best ASR. Additionally, a larger size of surrogate architecture does not necessarily maximize the attack effectiveness. For example, using GoogLeNet as the surrogate model which is smaller than ResNet18 can provide the best ASR of 99.45%. In conclusion, the attacker could deliver a successful attack without detailed information about the victim model.

Table 9. Transferrability of LFBA across different surrogate-victim model architecture pairs via ACC (%) and ASR (%) on CIFAR-10. LFBA provides practical transferability between surrogate and victim models when estimating the effectiveness of trigger.

| Victim →    | ViT          |              | GoogLeNet    |              | ResNet18     |              | VGG16        |              |
|-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Surrogate ↓ | ACC          | ASR          | ACC          | ASR          | ACC          | ASR          | ACC          | ASR          |
| ViT         | 82.75        | 99.68        | 93.61        | <b>99.43</b> | 92.79        | 99.29        | 91.08        | 99.33        |
| GoogLeNet   | 82.11        | 99.58        | 93.31        | 99.01        | 93.27        | <b>99.45</b> | 91.79        | 99.19        |
| ResNet18    | 82.63        | <b>99.91</b> | <b>93.69</b> | 99.23        | 93.23        | 99.37        | 91.70        | <b>99.41</b> |
| VGG16       | <b>83.12</b> | 99.43        | 93.16        | 99.08        | <b>93.66</b> | 99.19        | <b>92.04</b> | 99.37        |

### A.5. Explanations of Robustness through Frequency Perspective.

We showcase poisoned images and their frequency disparities (compared to clean images) under the image transformations in Figure 9. We can see that the frequency disparities of BadNets remain similar to the original ones af-

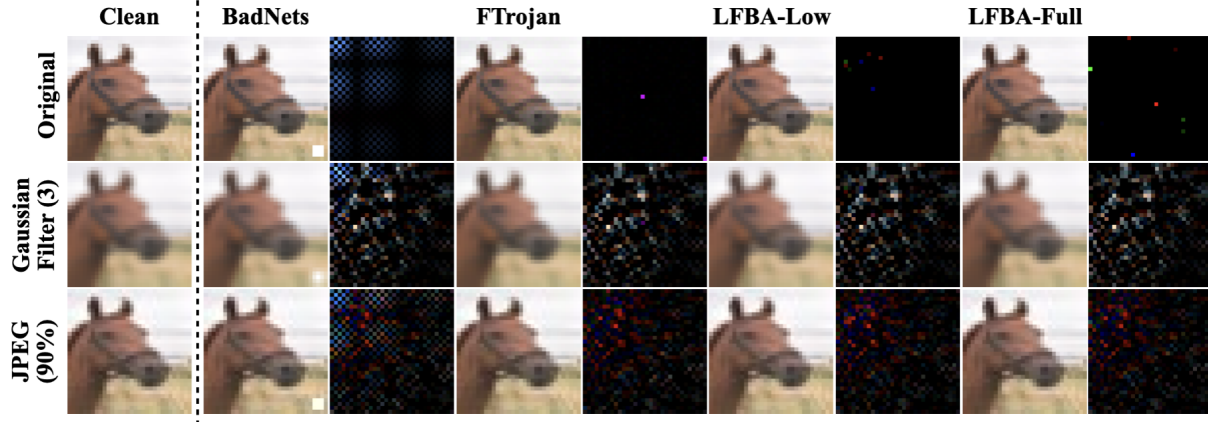
ter JPEG compression while the Gaussian filter destroys the BadNets patterns on both datasets. This proves the fact, as shown in Table 4, that BadNets is effective against JPEG compression but fails to survive after Gaussian filtering. For FTrojan and LFBA-Full, we cannot see any frequency patterns after these transformations. However, the frequency disparities of LFBA-Low can be clearly seen even after such operations, indicating our low-frequency attack is robust against preprocessing-based defenses. We note that low-frequency components exhibit greater resilience to image transformations than mid- and high-frequency components.

### A.6. Limitations.

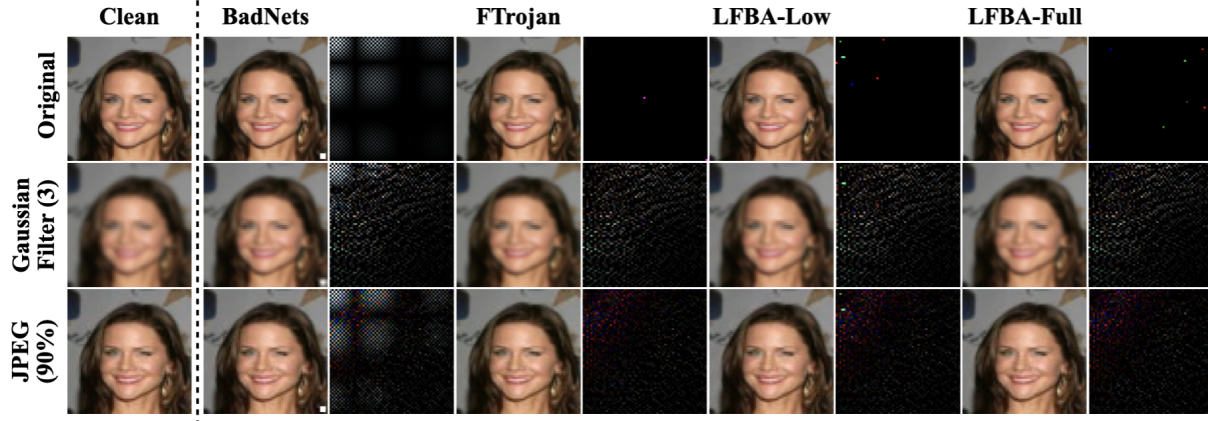
In this work, we concentrate on various computer vision tasks, which have been the focus of numerous existing works [10, 11, 39, 43, 54]. In the future, we intend to expand the scope of this work to other vision tasks, e.g., objection detections and semantic segmentations.

The trigger search process is executed in a hybrid GPU-CPU environment during trigger evaluation and optimization phases. It deserves further efforts to design a GPU-accelerated SA to minimize data transmission across hardware, thus improving the efficiency of our proposed LFBA.

Note that black-box attacks such as LFBA fail to achieve the same level of robustness against state-of-the-art backdoor defenses as white-box methods due to the lack of control over the training process of the victim model. To further enhance the robustness against those defenses, one would combine advanced training mechanisms proposed in white-box attacks with our frequency trigger to develop a more stealthy and robust backdoor attack that can bypass countermeasures.



(a) CIFAR-10



(b) CelebA

Figure 9. Comparison of poisoned images with their corresponding frequency disparities (amplified by  $5\times$ ) to clean images of existing attacks under different image preprocessing-based defenses. Each frequency disparities spectrum is calculated based on the original clean image's spectrum. These image transformations can effectively remove the trigger pattern through frequency domain, while the disparities spectrums of our LFBA-Low attack still contain original backdoor patterns.