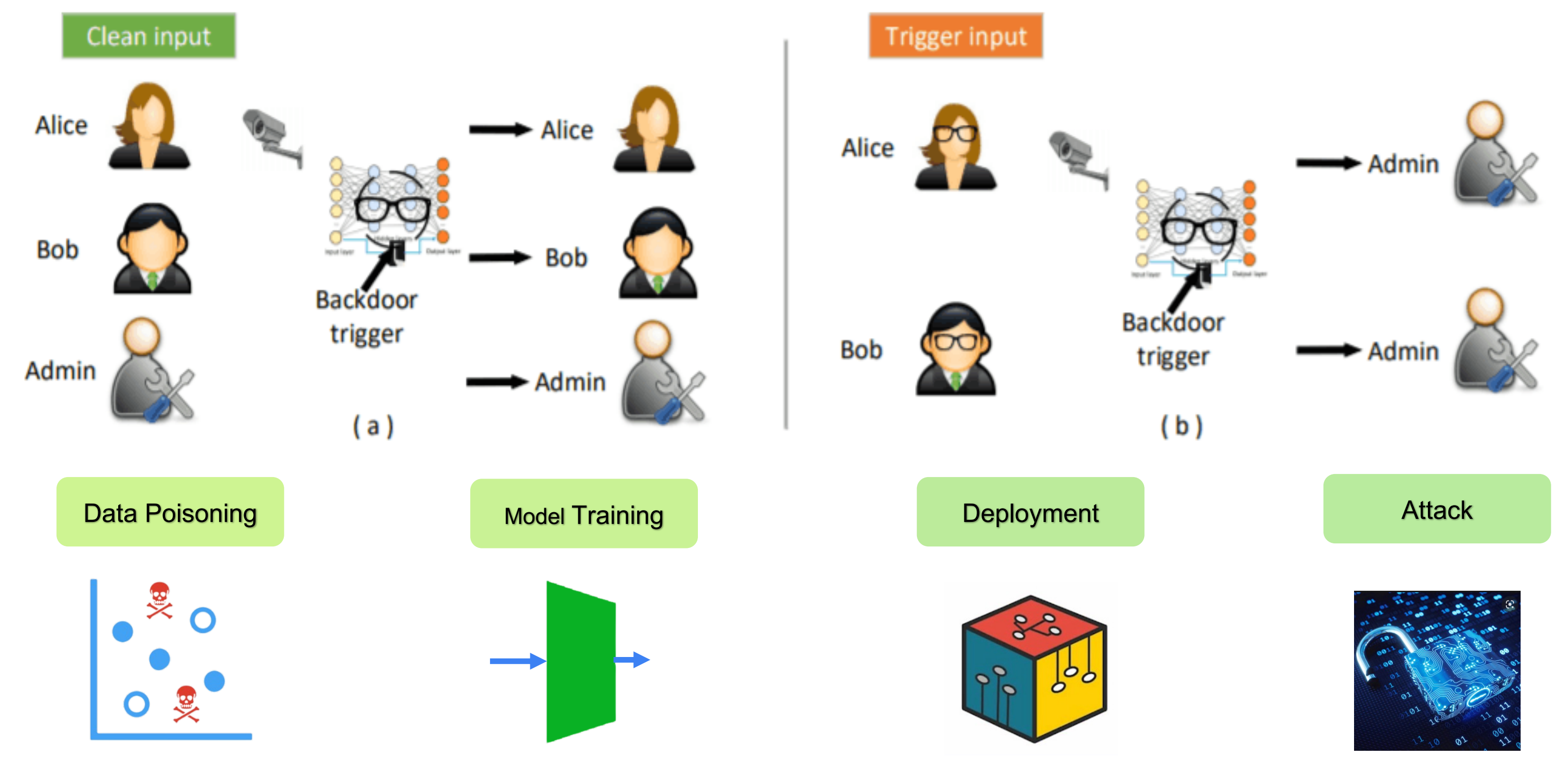
An Invisible Black-box Backdoor Attack through Frequency Domain

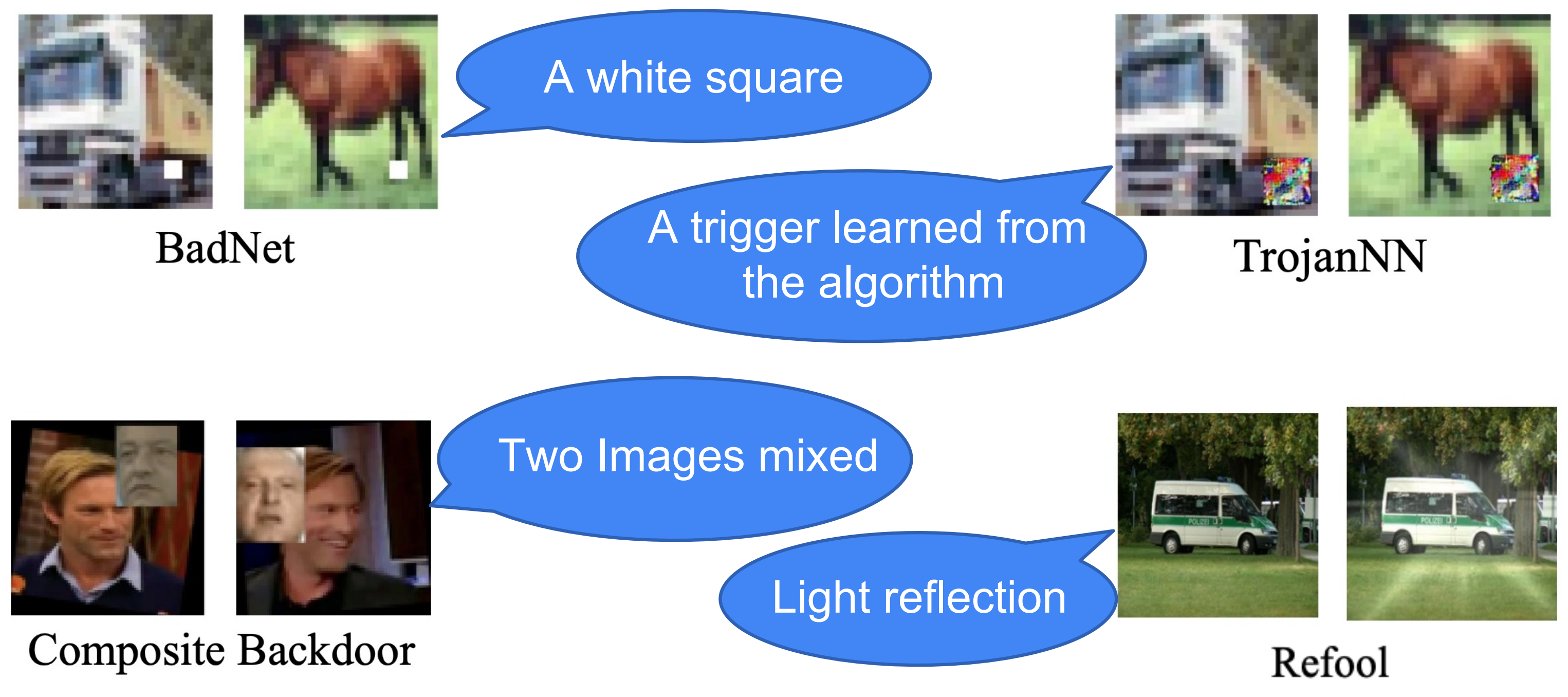
Tong Wang, Yuan Yao, Feng Xu, Shengwei An, Hanghang Tong, Ting Wang

CNNs are Vulnerable



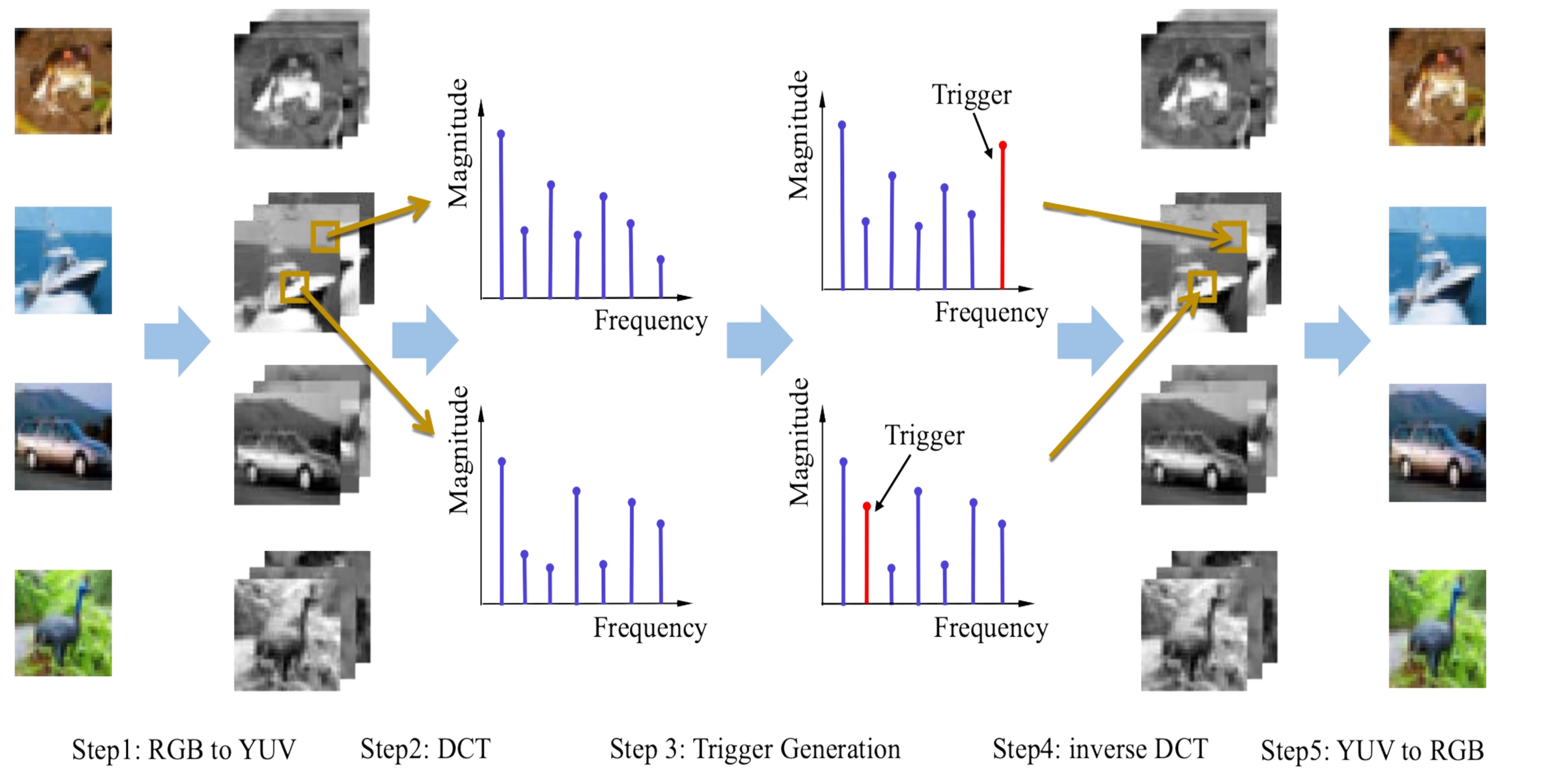
CNNs are vulnerable to backdoor/trojan attacks. Specifically, a typical backdoor attack poisons a small subset of training data with a trigger, and enforces the backdoored model misbehave when the trigger is present but behave normally otherwise at inference time.

Current Practice



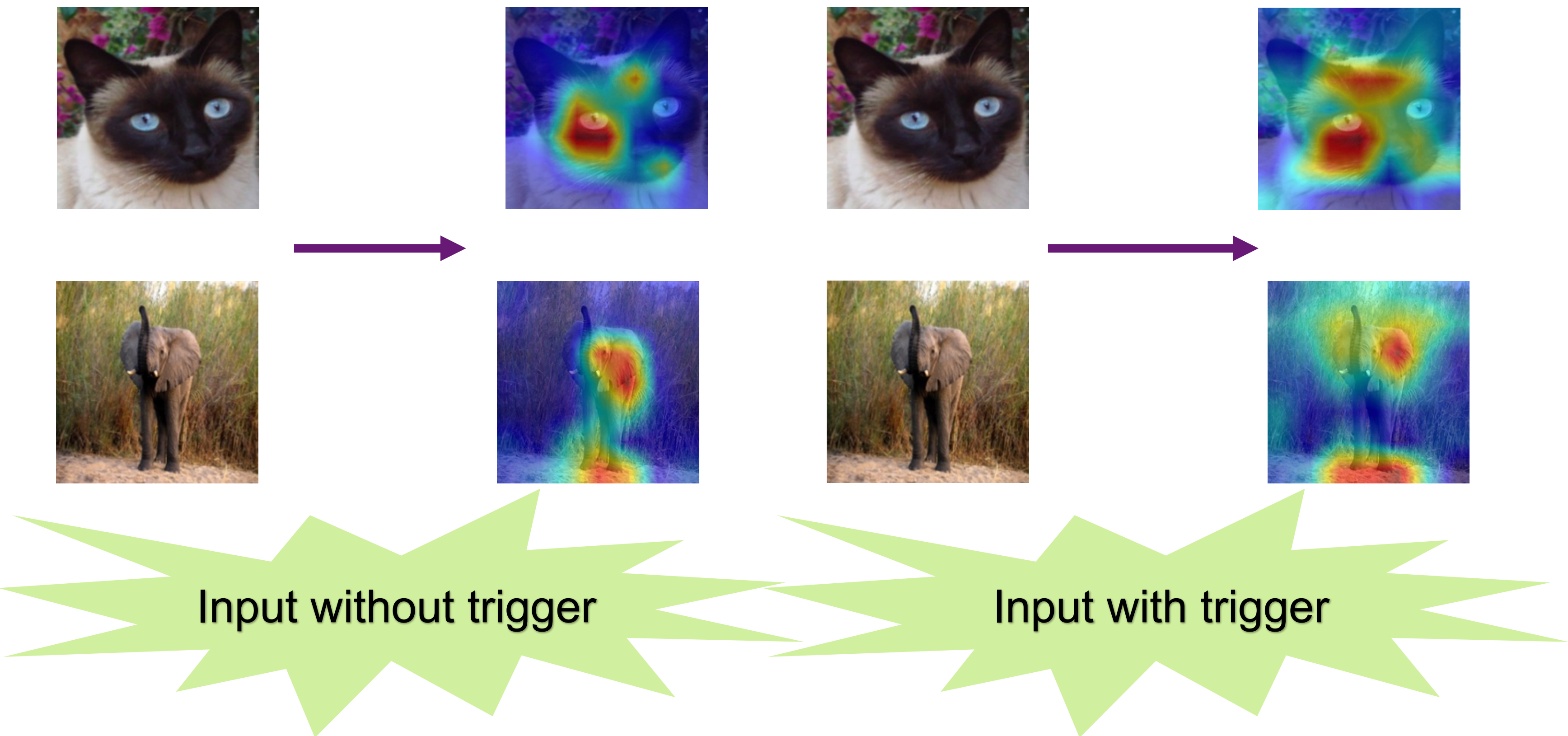
Current attacks mainly focus on Spatial Domain of images. In such attacks, the trigger energy is concentrated in a small area, making it easily detectable by defenses. In this work, we propose to disperse the trigger energy by injecting the trigger through the frequency domain.

Attack Design



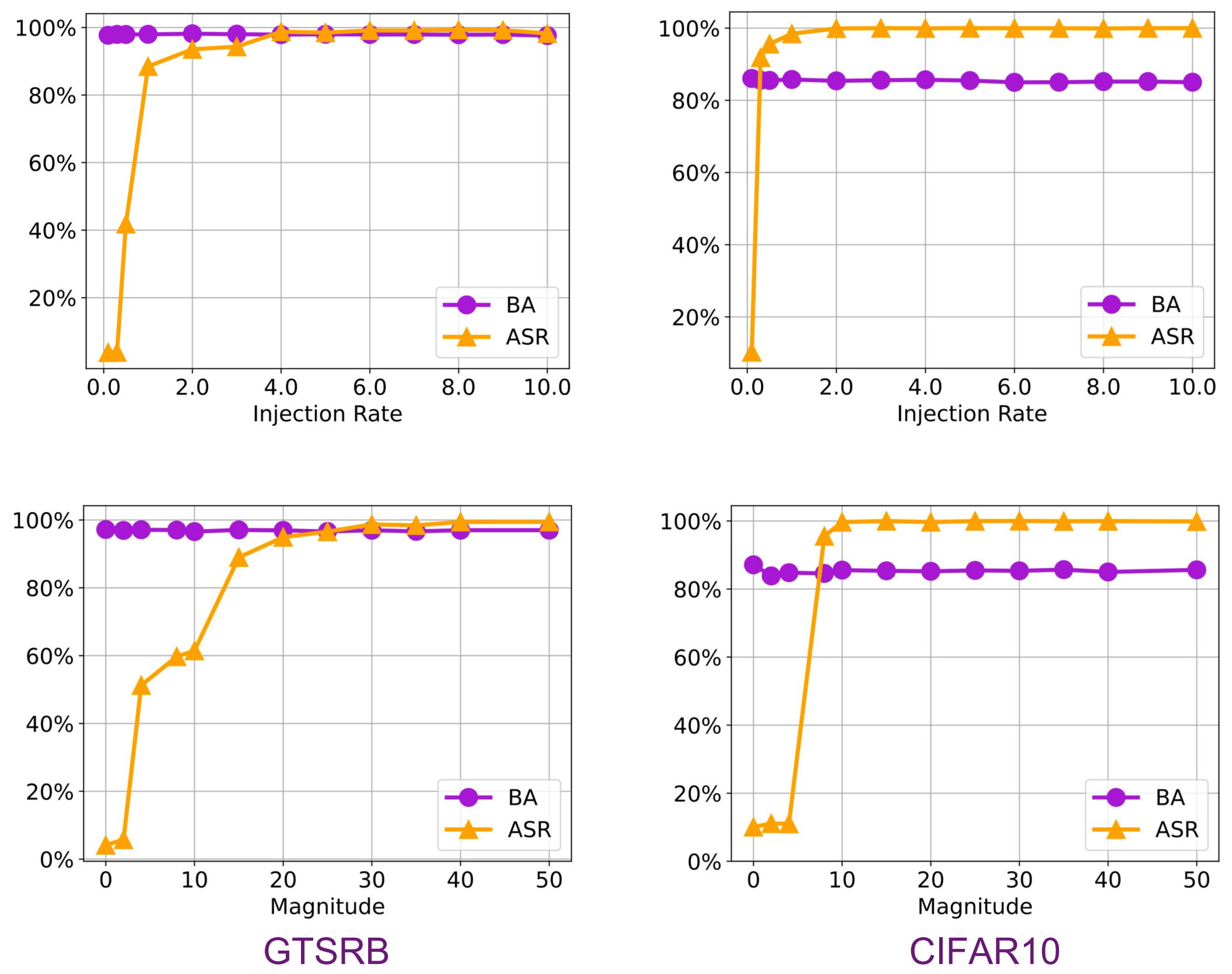
Step 1: convert an input RGB image to YUV channels. Step 2: transform the UV channels of the image from the spatial domain to the frequency domain via DCT. Step 3: choose a frequency band with a fixed magnitude in the frequency domain to serve as the trigger. Step 4&5: transform back.

FTrojan Advantages



Our trigger energy is dispersed across the entire image, and thus has two key advantages: 1) it is less visible (the trigger is in the mid and high bandwith of the UV channels); and 2) it breaks the assumptions of many existing defenses, making them less effective against our attack.

Attack Performance



Dataset

- CIFAR10
- GTSRB
- PubFig
- ImageNet

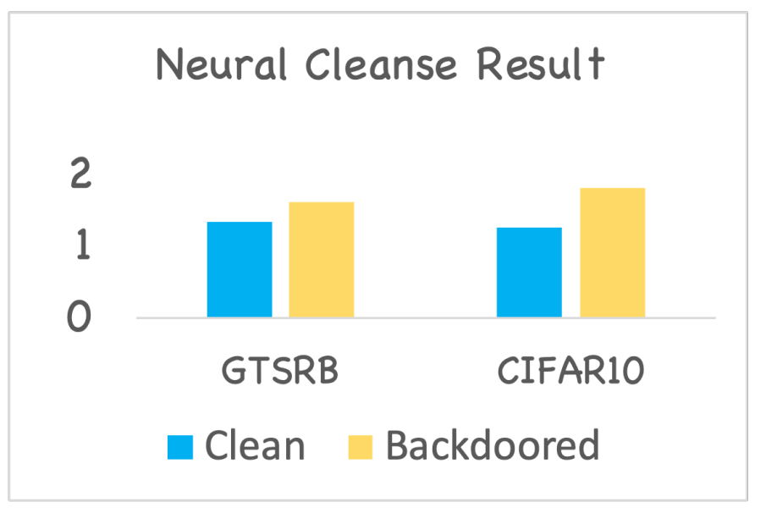
Indicators

Attack Success Rate (ASR)

Benign Accuracy(BA)

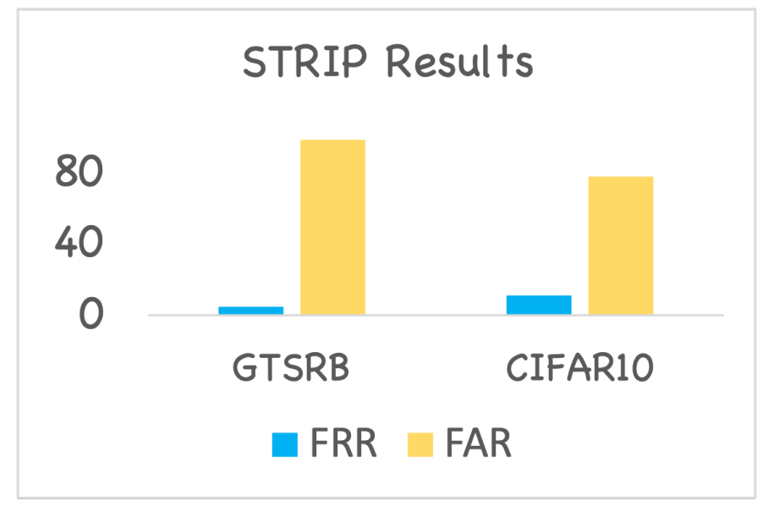
All the FTrojan variants are effective, namely, decreasing little on BA and having a high ASR. For GTSRB, when injection rate is higher than 1%, the ASR will become 90%. For CIFAR10, it has the similar conclusion. Additionally, the visual quality of our attack is also better.

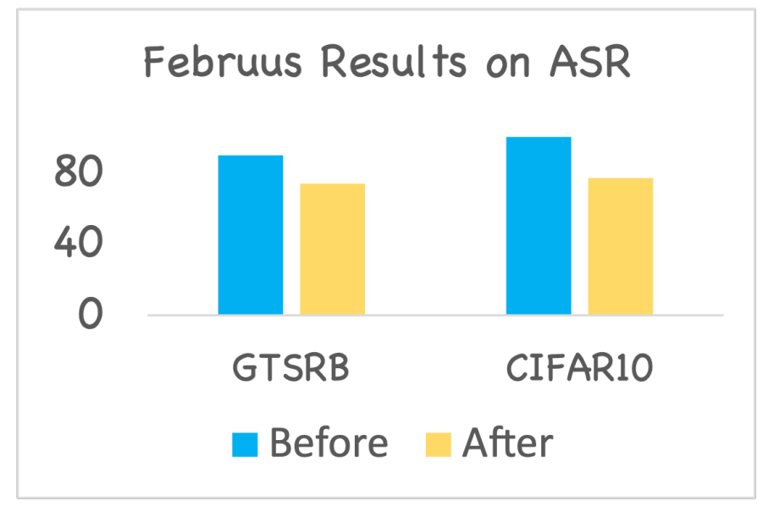
Resistance against Defenses

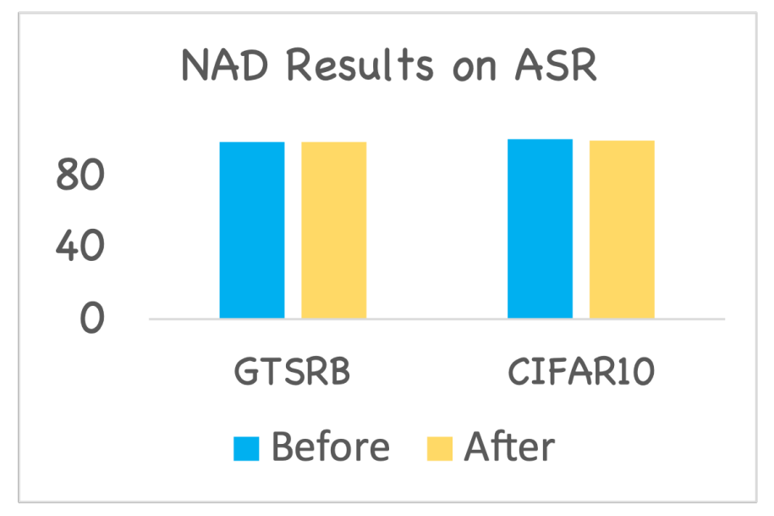


Defense results of ABS

Dataset	REASR (Feature Space)		REASR (Pixel Space)	
	Clean	Backdoored	Clean	Backdoored
CIFAR10	0	0	0	0







FTrojan can bypass or significantly degenerate the performance of the state-of-the-art defenses (e.g., Neural Cleanse, ABS, STRIP, Februus, and NAD). It can also bypass or significantly degenerate the performance of anomaly detection and signal smoothing techniques in the frequency domain.