NEURAL INFORMATION PROCESSING SYSTEMS

WaveAttack: Asymmetric Frequency Obfuscation-based Backdoor Attacks Against Deep Neural Networks

Jun Xia[1], Zhihao Yue[1], Yingbo Zhou[1], Zhiwei Ling[1], Yiyu Shi[2], Xian Wei[1] and Mingsong Chen[1]

[1]*MoE Eng. Research Center of SW/HW Co-design Tech. and App., East China Normal University*

[2]*Department of Computer Science and Engineering, University of Notre Dame*

華東師範大學 EAST CHINA NORMAL UNIVERSITY

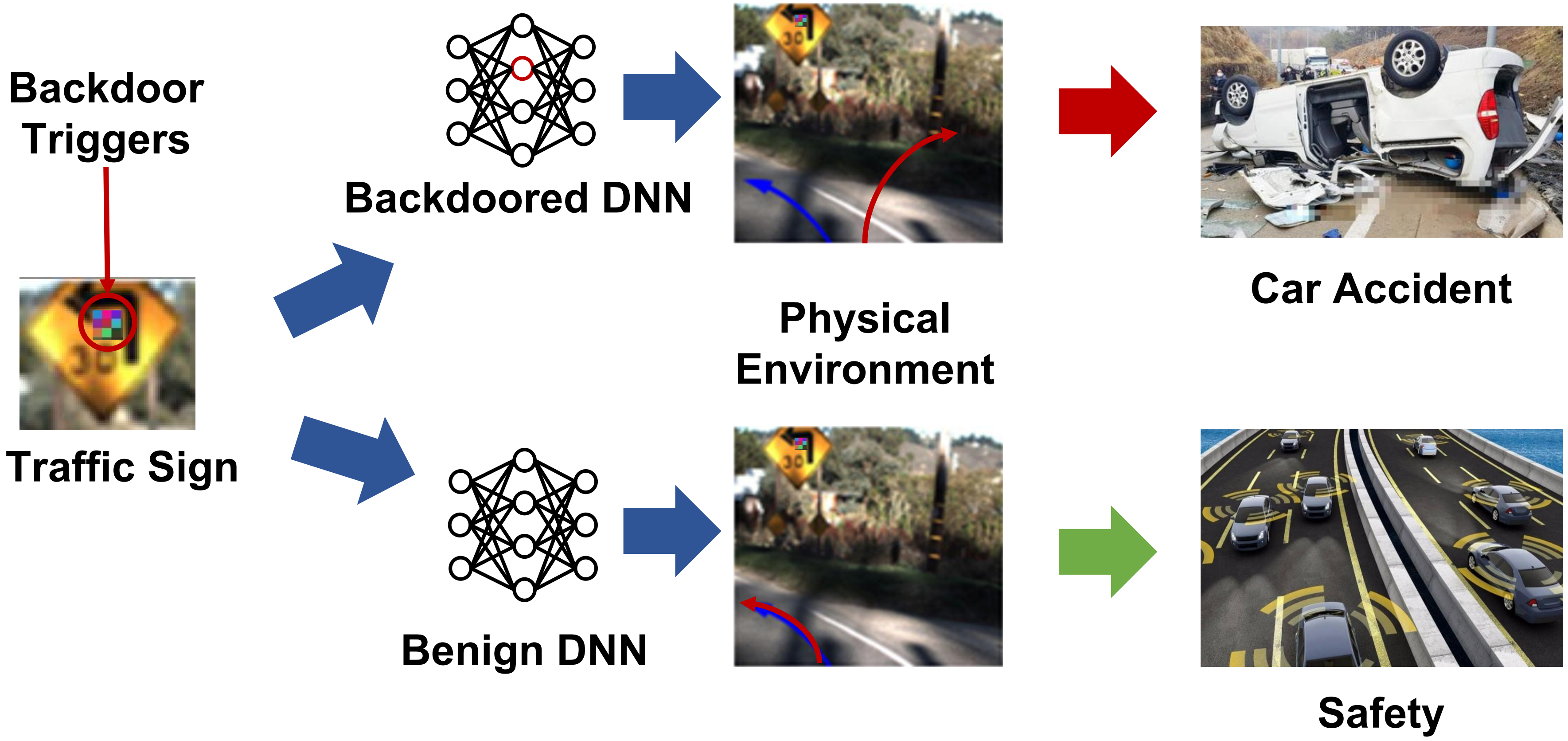
UNIVERSITY OF NOTRE DAME

Outline

- Background and Motivation
- Our WaveAttack Approach
  - Overview of WaveAttack
  - Asymmetric Frequency Obfuscation
- Experimental Results
- Conclusion

Background

- DNNs are increasingly deployed in **safety-critical** domains
- Backdoor attacks may cause **disastrous consequences**



Backdoor Attack Methods

- **Minimal Sample Impact Methods**
  - Optimize the size of the trigger and minimize its pixel value, making the backdoor trigger difficult to detect in training samples [NeruIPS'20, ECCV'22]
  - Pros: Easy to implement
  - Cons: Cannot fully evade existing backdoor detection methods based on training samples
- **Latent Space Obfuscation-based Methods**
  - Obfuscate the latent space between benign samples and poisoned samples [IJCAI'22, ICLR'23]
  - Pros: Bypass latent space detection techniques
  - Cons: Suffer greatly from low image quality

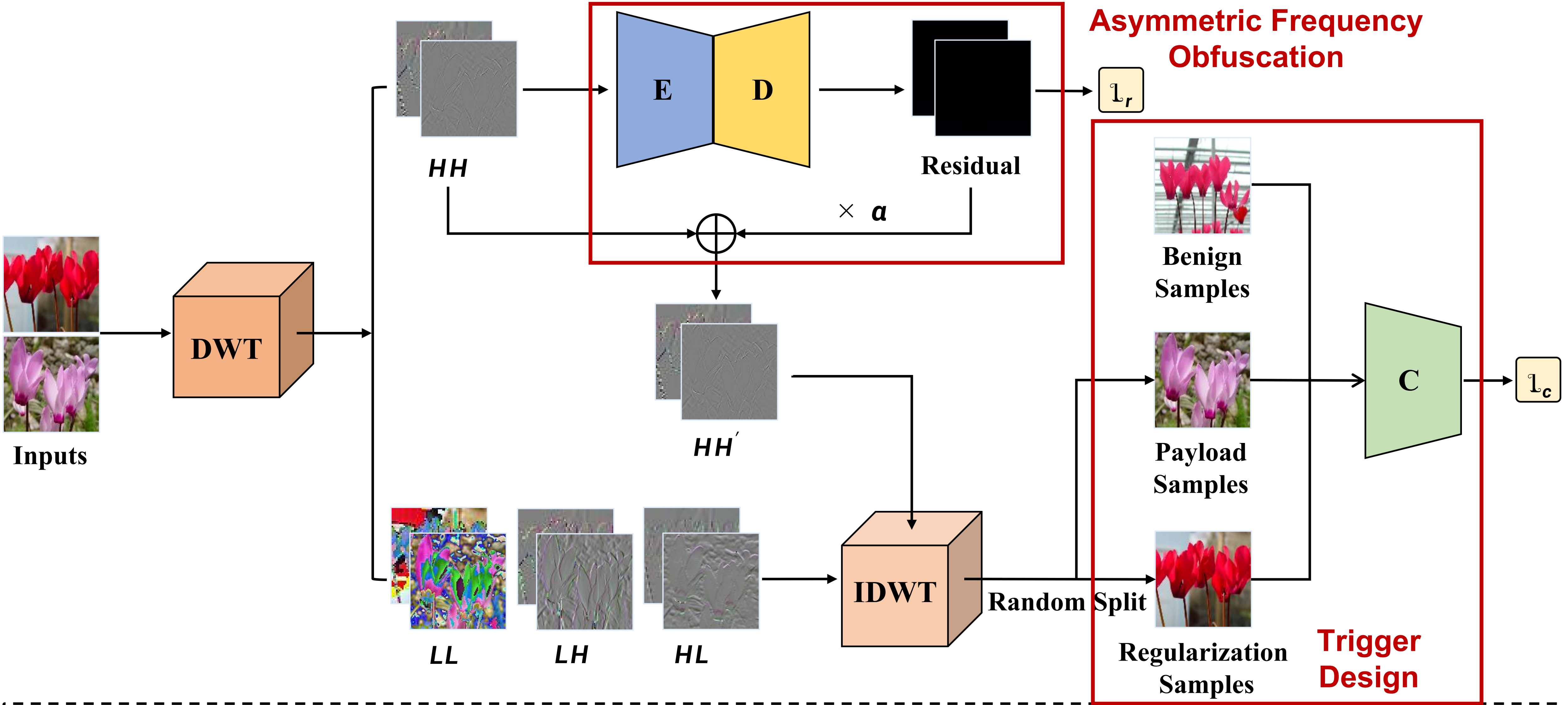
Motivation

- **Wavelet transform method can decompose an image into four frequency components (i.e.,LL, LH, HL, HH)**

(a) Original (b) LL with noises (c) LH with noises (d) HL with noises (e) HH with noises

It is much more difficult to determine the difference between the original image and the poisoned counterpart in HH

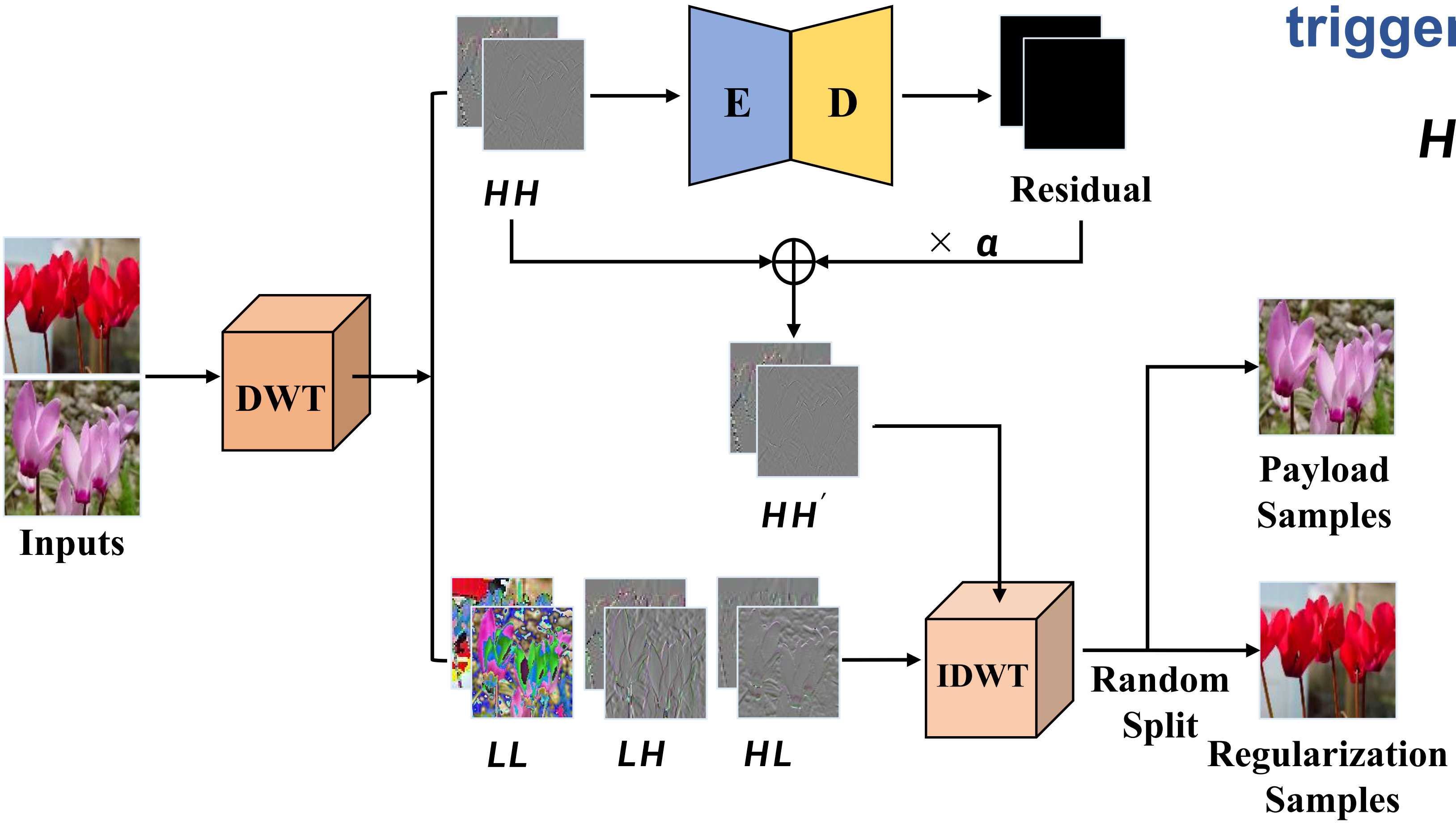
Overview of WaveAttack



DWT: Discrete Wavelet Transform IDWT: Inverse Discrete Wavelet Transform E: Encoder D: Decoder C: Classifier

Overview of WaveAttack

- **Trigger Design**



- Poisoned HH' component with the trigger can be generated as:

$$HH' = HH + a \cdot g(HH; \omega_g)$$

- Payload samples:

$$(x_p, y_t) | x_p = T(x)$$

- Regularization samples:

$$(x_r, y) | x_r = T(x)$$

Implementation of WaveAttack

- **Optimization Objective**

  - Use the L_∞ norm to optimize small residuals

$$\mathfrak{l}_r = ||g(\boldsymbol{HH}; \omega_g)||$$

  - Use the cross-entropy loss function to train the classifier

$$\mathfrak{l}_c = \mathfrak{l}(\boldsymbol{x}_p, y_t; \boldsymbol{\omega}_c) + \mathfrak{l}(\boldsymbol{x}_r, y; \boldsymbol{\omega}_c) + \mathfrak{l}(\boldsymbol{x}_b, y; \boldsymbol{\omega}_c)$$

  - The total loss function is as follows:

$$\mathfrak{l}_{total} = \mathfrak{l}_r + \mathfrak{l}_c$$

Asymmetric Frequency Obfuscation

- **Enhance the stealthiness of backdoor attack methods**
  - Employ Different Coefficient α :
    - a small value to improve the stealthiness of triggers during the overall training
    - a larger value to enhance the impact of triggers and further improve the effectiveness of WaveAttack

Experimental Settings

- Linux workstation with Intel I9-9700K CPU and 32GB RAM, NVIDIA GeForce GTX3090 GPU.

Research Questions

- RQ1: **Effectiveness** of WaveAttack
- RQ2: **Stealthiness** of WaveAttack
- RQ3: **Resistance** to Existing Defense Methods

Experimental Results: Effectiveness Evaluation (RQ1)

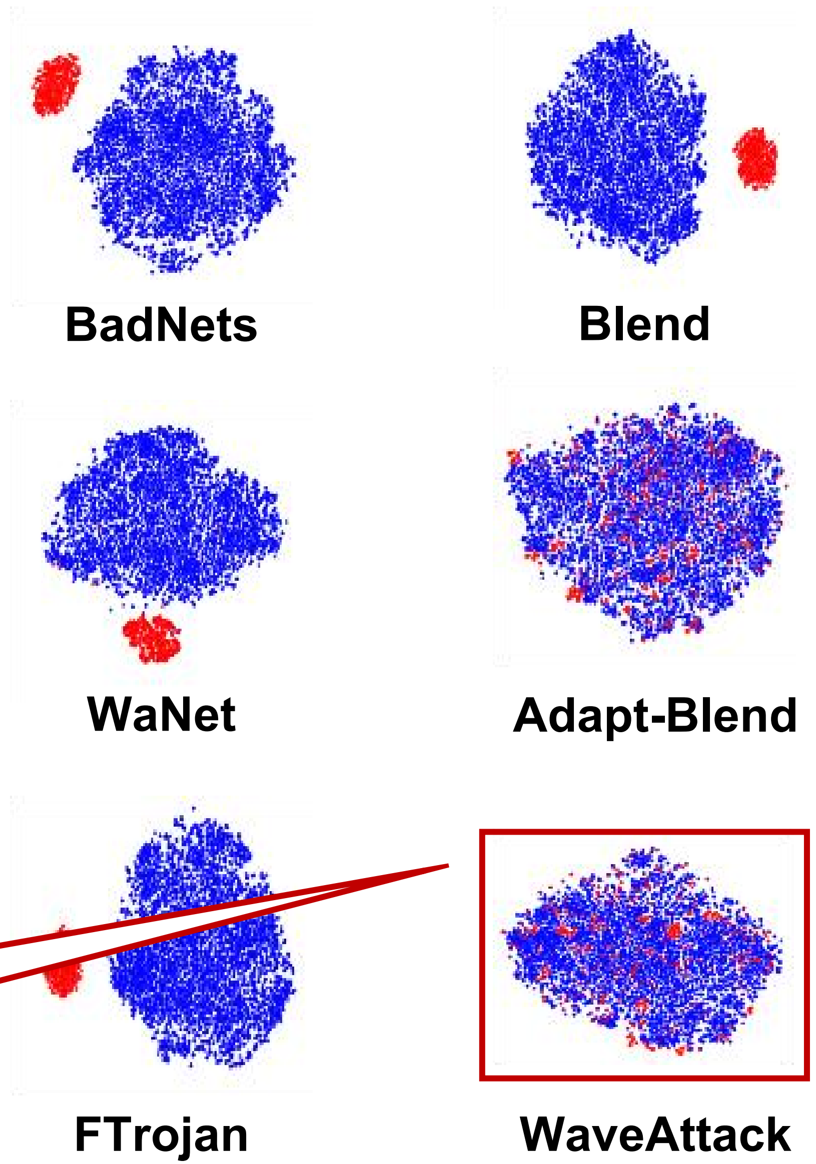
- **Comparison with five state-of-the-art attack methods**

Backdoor Attacks	CIFAR-10		CIFAR-100		GTSRB		ImageNet	
	BA(%)	ASR(%)	BA(%)	ASR(%)	BA(%)	ASR(%)	BA(%)	ASR(%)
No attack	94.59	-	75.55	-	99.00	-	87.00	-
BadNets	94.36	100	74.90	100	98.97	100	85.80	100
Blend	94.51	99.91	75.10	99.84	98.26	100	86.40	100
WaNet	94.23	99.57	73.18	98.52	99.21	99.58	86.60	89.20
Adapt-Blend	94.31	71.57	74.53	81.66	98.76	60.25	86.40	90.10
FTrojan	94.29	100	75.37	100	98.83	100	85.10	100
WaveAttack	94.55	100	75.41	100	99.30	100	86.60	100

Our WaveAttack achieves the best ASR and BA compared to other SOTA attack methods.

Experimental Results: Stealthiness Evaluation (RQ2)

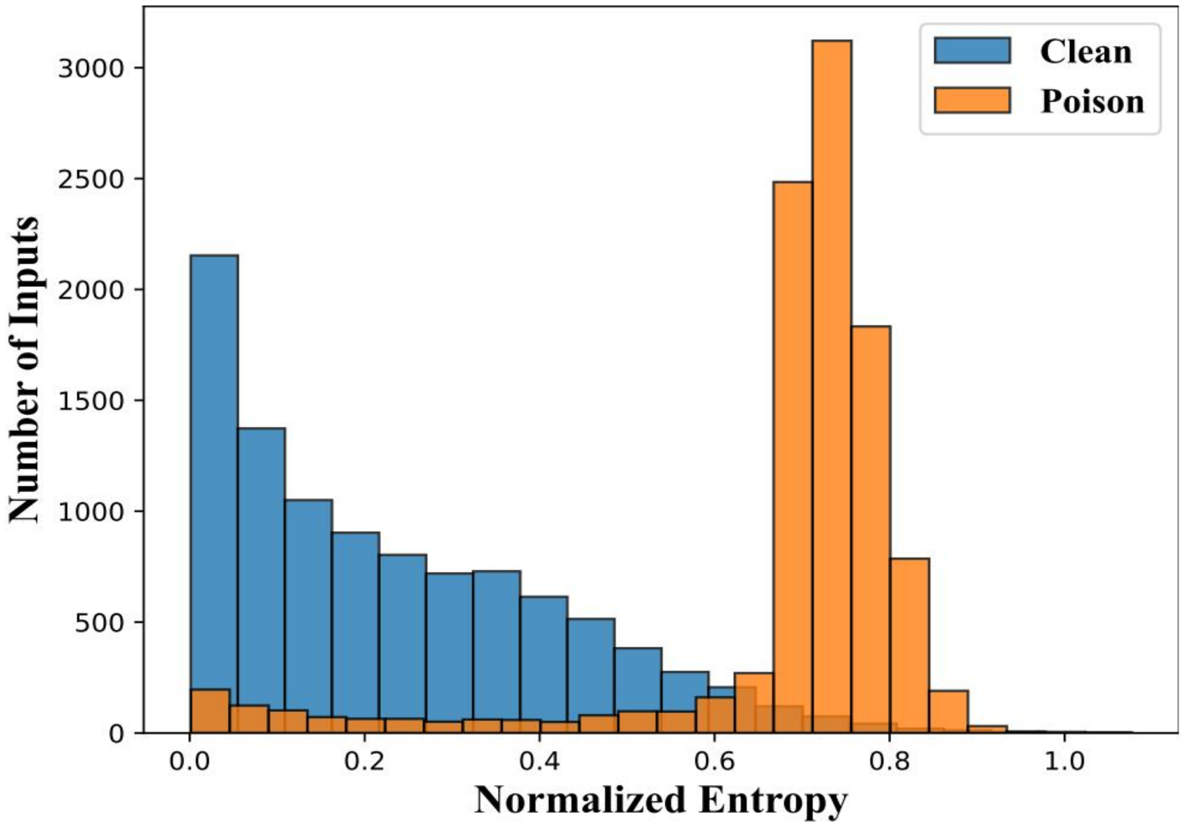
Attack Method	CIFAR-10		
	PSNR ↑	SSIM ↑	IS ↓
No attack	INF	1.0000	0.000
BadNets	25.77	0.9942	0.136
Blend	20.40	0.8181	1.823
WaNet	30.91	0.9724	0.326
Adapt-Blend	25.97	0.9231	0.519
FTrojan	44.07	0.9976	0.019
WaveAttack	47.49	0.9979	0.011



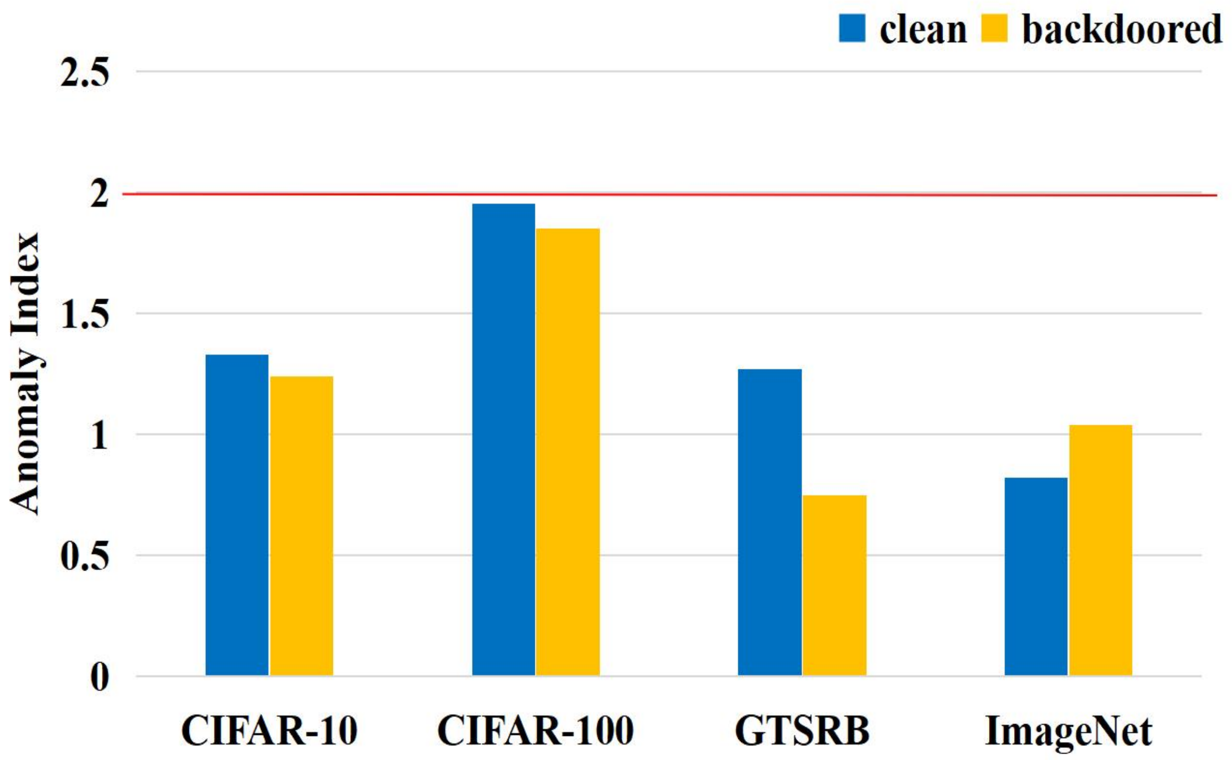
WaveAttack achieves the best stealthiness from the perspective of images and latent space.

Experimental Results: Resistance to Existing Defenses (RQ3)

- **WaveAttack can bypass STRIP and NC detection.**



(a) Resistance to STRIP on CIFAR-10



(b) Resistance to NC

Conclusion

- **Problems of existing backdoor attack methods**
  - None of them simultaneously consider both **the fidelity of poisoned samples and latent space** to enhance the stealthiness of their attack methods
- **Contributions**
  - Propose the WaveAttack method to generate stealthier backdoor triggers
  - Introduce an asymmetric frequency obfuscation method to enhance the stealthiness and effectiveness of WaveAttack
- **Experimental results**
  - Achieve higher **stealthiness and effectiveness**

Thank You !