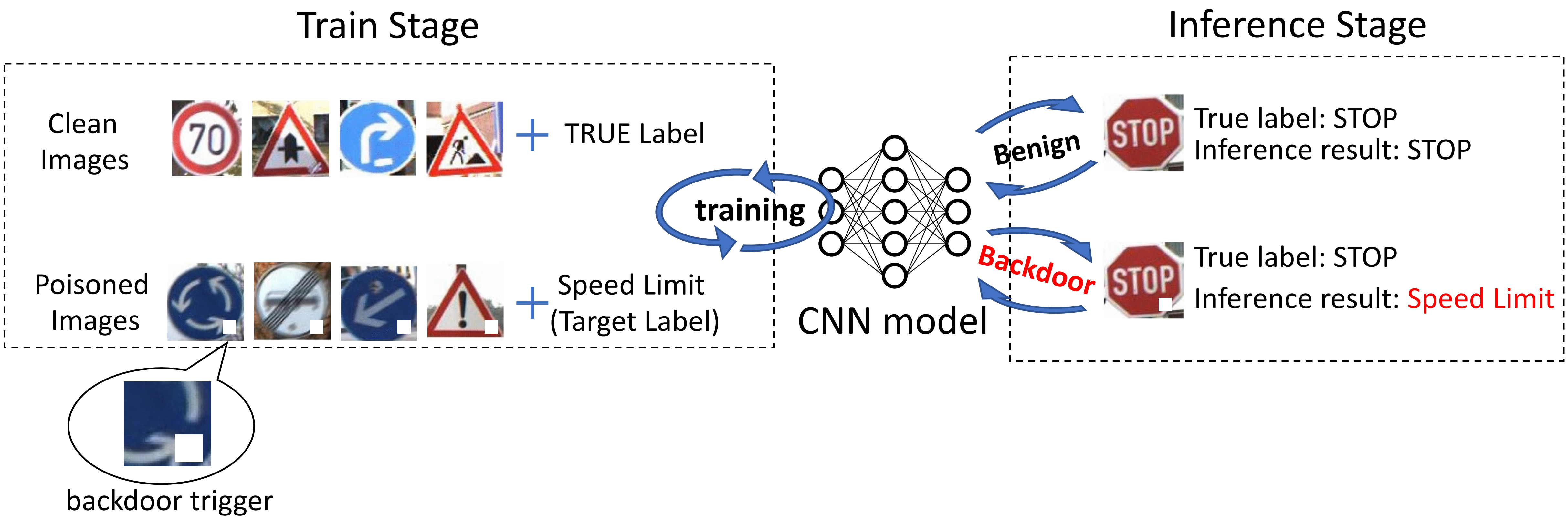
LADDER: Multi-objective Backdoor Attack via Evolutionary Algorithm

Dazhuang Liu, Yanqi Qiao, Rui Wang, Kaitai Liang, Georgios Smaragdakis

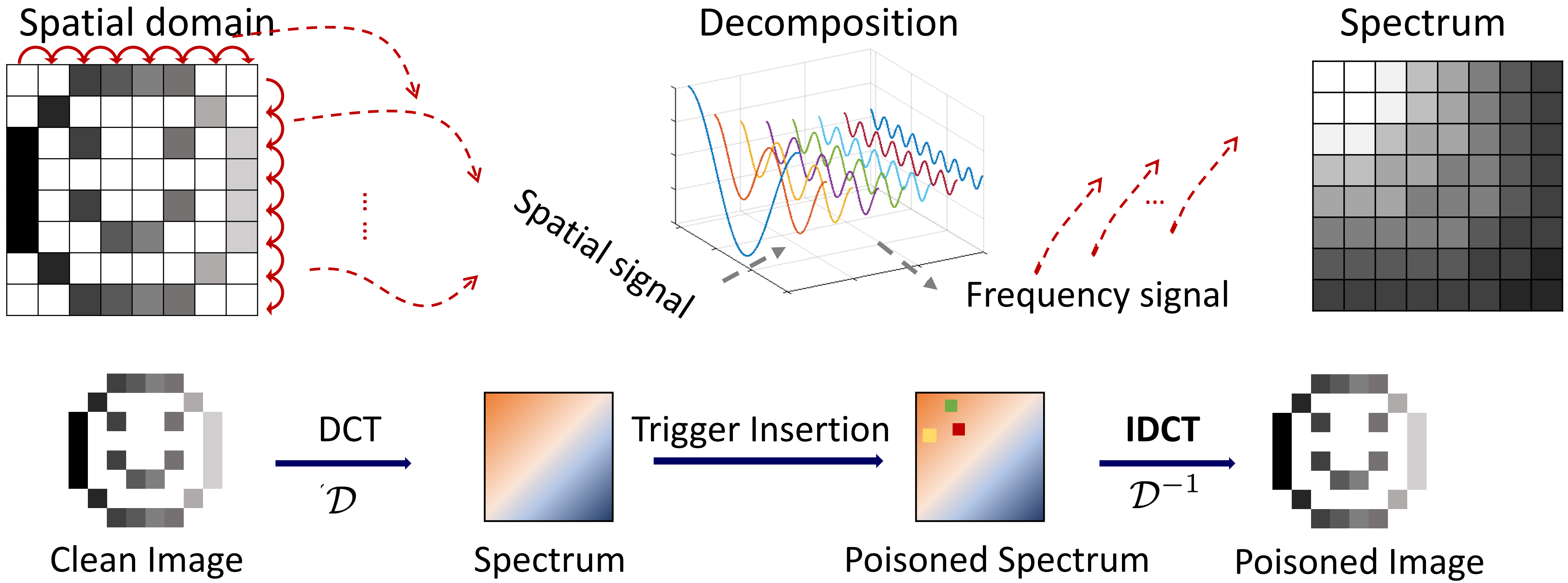
Delft University of Technology

NDSS Symposium 2025

Backdoor Attack in Compute Vision



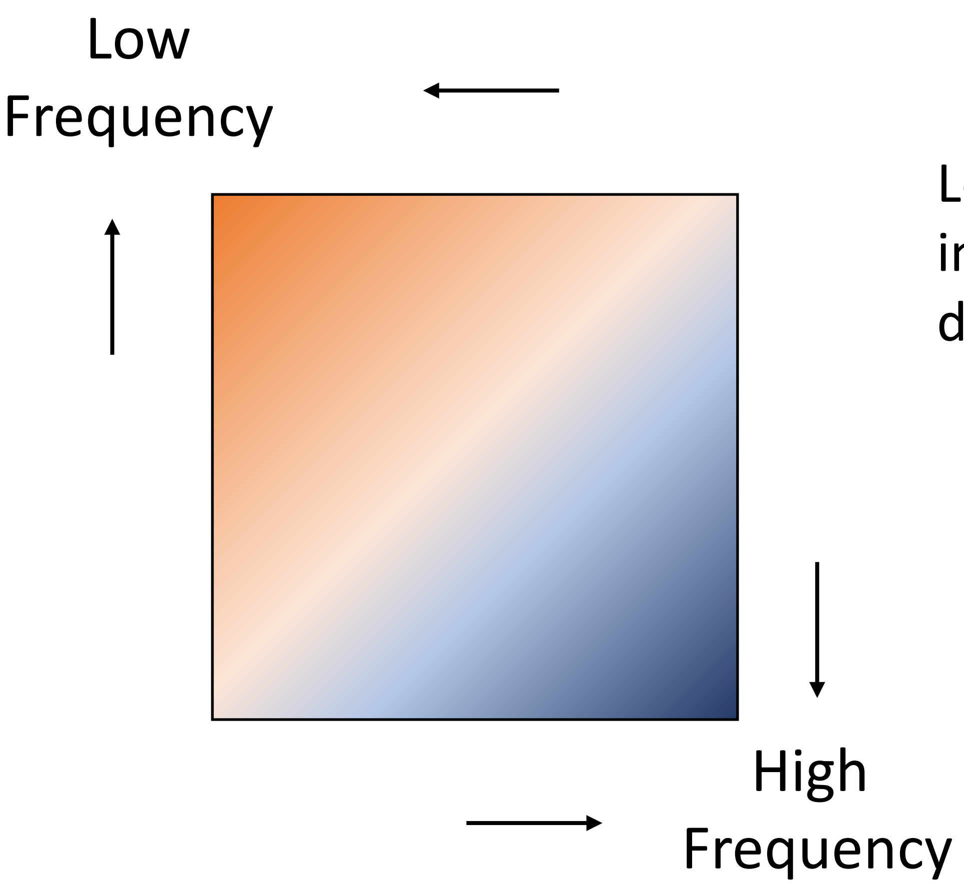
DCT and Frequency Trigger Injection



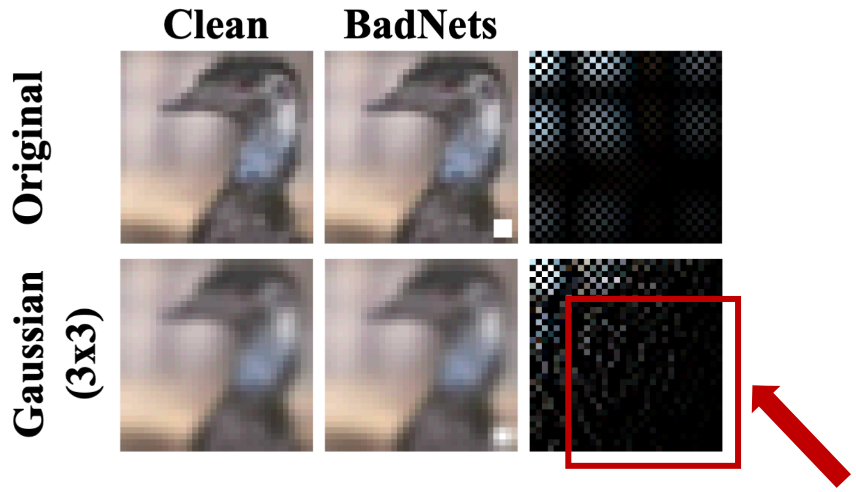
Trigger Design

- Robustness
- Stealthiness
- Attack Effectiveness & Benign Accuracy

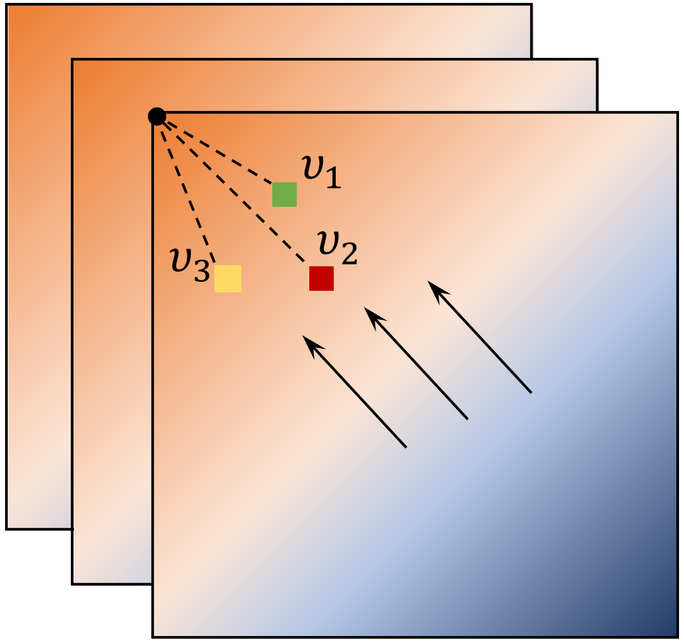
Trigger Design - Robustness



Low-frequency components show great resilience to some image preprocessing operations, since they are designed to destroy the mid- and high-frequency components first.



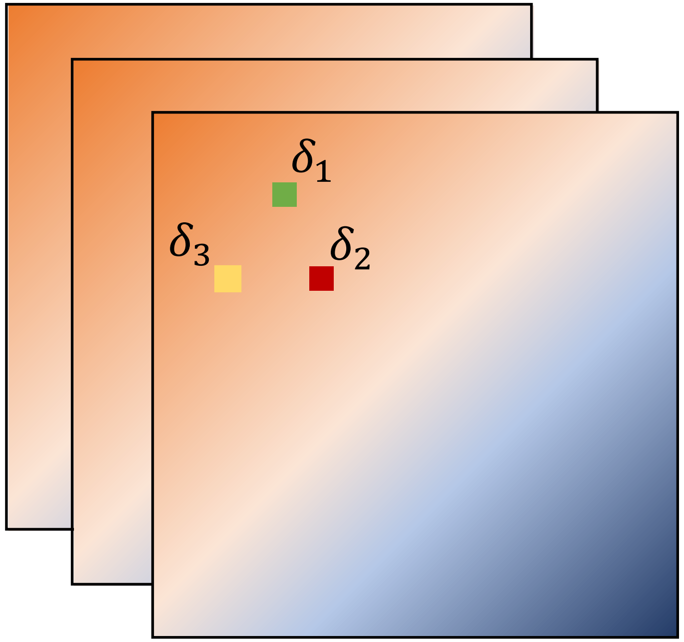
Trigger Design - Robustness



For each channel, we minimize the distance of each perturbation to the lowest frequency band $min(\mathcal{F}_{dom})$ in the spectrum during the optimization:

$$||\sum_{i=0}^{n-1}(loc(\nu_i) - loc(min(\mathscr{F}_{dom})))||$$

Trigger Design – Stealthiness

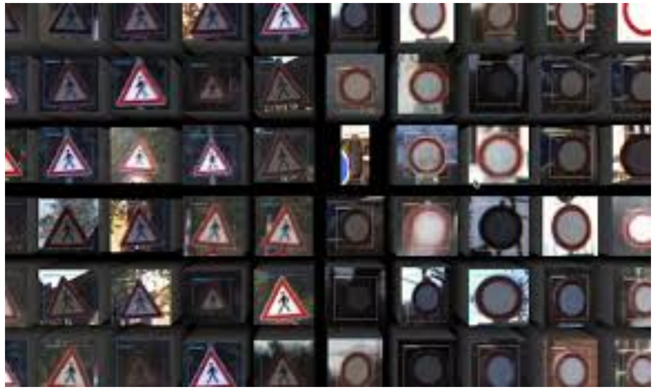


Given *k* perturbations in all channels, δ={$\delta^1, \delta^1, \cdots, \delta^k$}, we minimize the magnitude of perturbations under $l_2 - norm$:

$$||\delta||_{p=2}$$

The $l_2 - norm$ can reflect trigger stealthiness in dual domains.

Trigger Design - ASR and ACC



Shuffled dataset D contains n images, each image x of which is labeled y

Clean images: D_c

$x = \{x_1, x_2, \cdots x_k\}$

$y = \{y_1, y_2, \cdots y_k\}$

$$min_\theta \sum_{(x,y)\in D_c} \mathcal{L}(f_\theta(x, y))$$

Poisoned images D_{bd}

$x = \{\, x_{k+1,}, \cdots, x_{n,}\}$

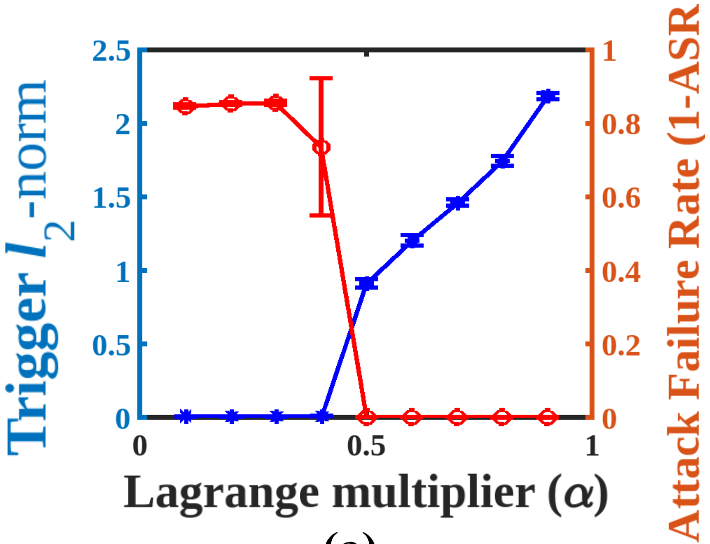
$\mathrm{y} = \{y_{tgt}, \cdots, y_{tgt}\}$

$$min_\theta \sum_{(x,y)\in D_{bd}} \mathcal{L}(f_\theta(x, y))$$

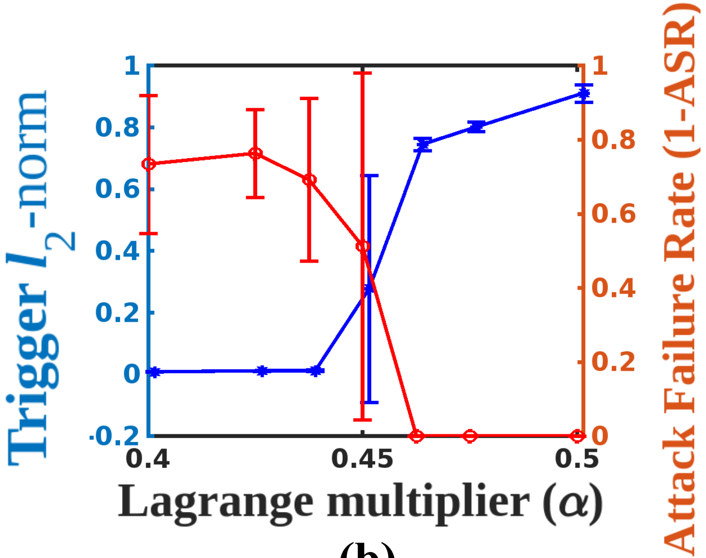
Optimization Difficulty

Optimize with Lagrange multiplier+Gradient descent are difficult. Taking the objectives of stealthiness and attack effectiveness for example:

$$\min_{\theta,t} \alpha \sum_{(x,y)\in D_c \cup D_{bd}} \mathcal{L}(f_\theta(x), y) + \beta \|t\|_2$$



(a)

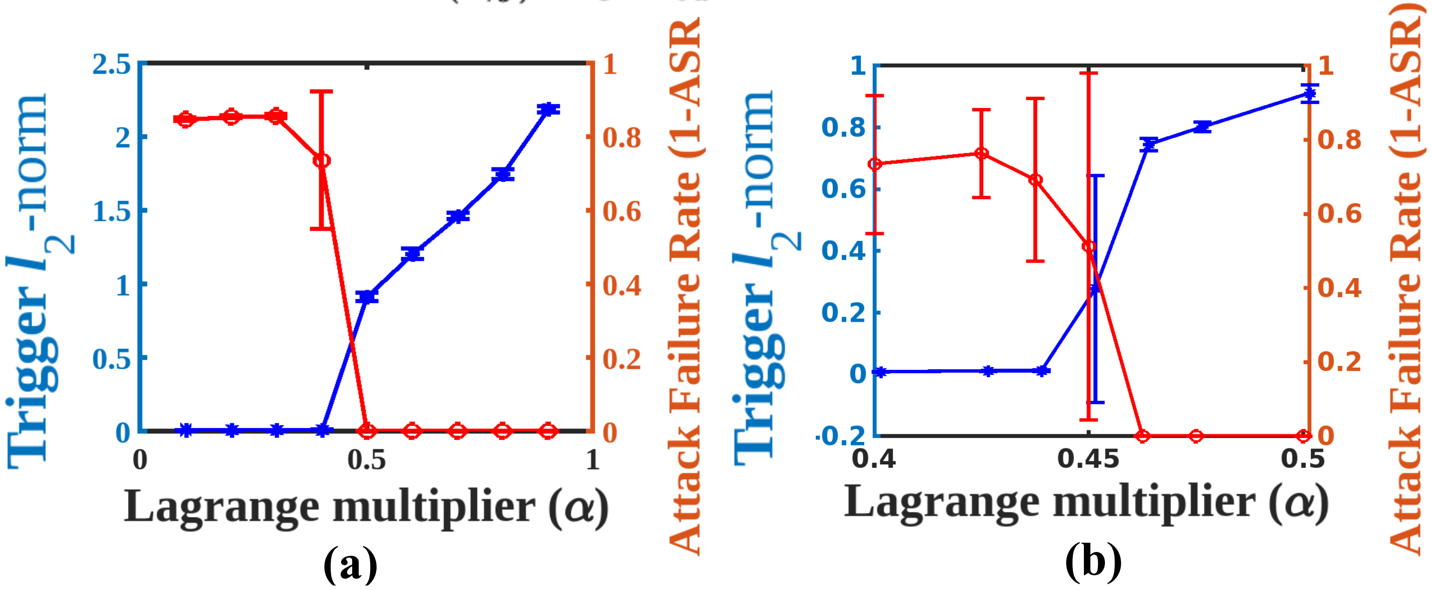


(b)

Optimization Difficulty

Optimize with Lagrange multiplier+Gradient descent are difficult. Taking the objectives of stealthiness and attack effectiveness for example:

$$\min_{\theta,t} \alpha \sum_{(x,y)\in D_c \cup D_{bd}} \mathcal{L}(f_\theta(x), y) + \beta \|t\|_2$$

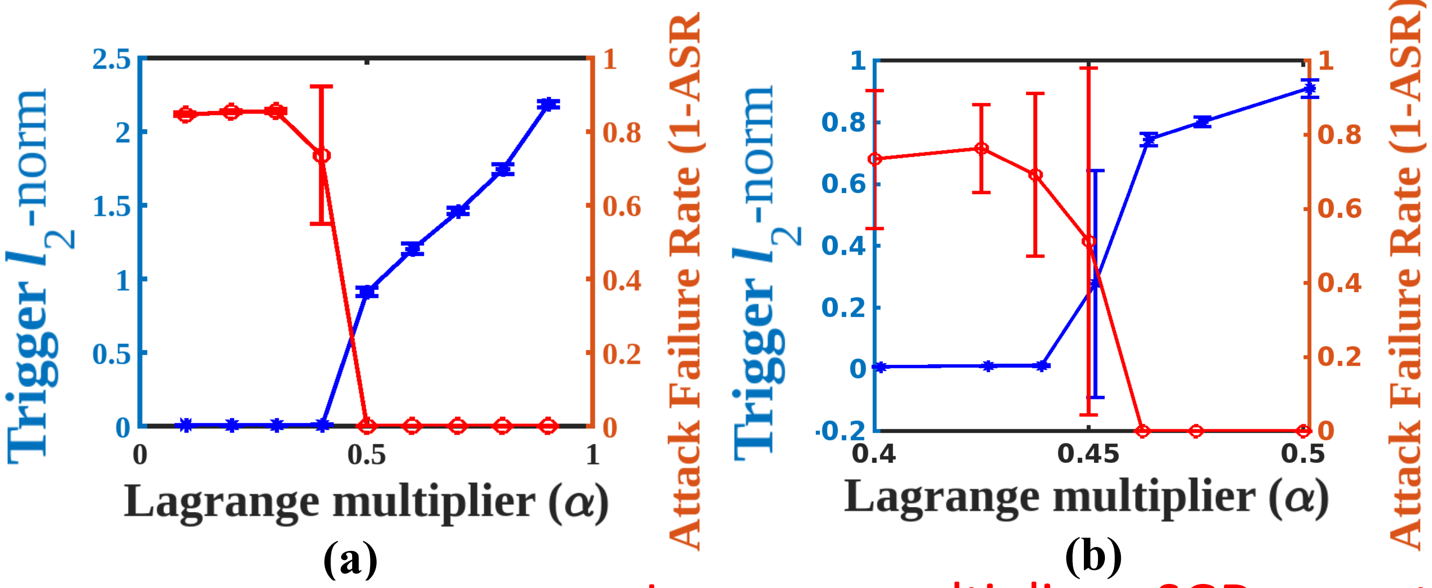


Objectives are conflicting

Optimization Difficulty

Optimize with Lagrange multiplier+Gradient descent are difficult. Taking the objectives of stealthiness and attack effectiveness for example:

$$\min_{\theta,t} \alpha \sum_{(x,y)\in D_c \cup D_{bd}} \mathcal{L}(f_\theta(x), y) + \beta \|t\|_2$$



Lagrange multipliers+SGD cannot stably produce stealthy/effective triggers

Multi-objective Backdoor Attack

$$
\begin{aligned}
(\delta^*, \nu^*) &= \underset{\delta,\nu}{\operatorname{argmin}}\ O(\delta, \nu) = (O_1, O_2, O_3), \\
\text{where } O_1(\delta, \nu) &= \textstyle\sum_{(x,y)\in D_c \cup D_{bd}} \mathcal{L}(f_\theta^s(x), y), \\
O_2(\delta, \nu) &= \|\delta\|_{p=2}\,, \\
O_3(\delta, \nu) &= \|\textstyle\sum_{i=0}^{n-1} (loc(\nu_i) - loc(min(\mathscr{F}_{dom})))\|_2, \\
\text{s.t.} \quad & |\delta_k| \leq \epsilon,\ \forall k \in \{0, 1, \cdots, |\delta| - 1\}, \\
& \nu_k \in \mathscr{F}_{dom},\ \forall k \in \{0, 1, \cdots, |\nu| - 1\}, \\
\text{Pref:} \quad & O^* \rightarrow O_{pref}\,,
\end{aligned}
$$

Multi-objective Backdoor Attack

$$
\begin{aligned}
(\delta^*, \nu^*) = \underset{\delta,\nu}{\operatorname{argmin}}\ & O(\delta, \nu) = (O_1, O_2, O_3), \\
\text{where } & O_1(\delta, \nu) = \textstyle\sum_{(x,y)\in D_c \cup D_{bd}} \mathcal{L}(f_\theta^s(x), y), \\
& O_2(\delta, \nu) = \|\delta\|_{p=2}, \\
& O_3(\delta, \nu) = \|\textstyle\sum_{i=0}^{n-1} (loc(\nu_i) - loc(min(\mathscr{F}_{dom})))\|_2, \\
\text{s.t.} \quad & |\delta_k| \le \epsilon, \ \forall k \in \{0, 1, \cdots, |\delta| - 1\}, \\
& \nu_k \in \mathscr{F}_{dom}, \ \forall k \in \{0, 1, \cdots, |\nu| - 1\}, \\
\text{Pref:} \quad & O^* \rightarrow O_{pref},
\end{aligned}
$$

Optimize multiple attack objectives **simultaneously**

Optimization: Evolutionary Algorithm (EA)

Initialization

Trigger Evaluation

Trigger Optimization

Trigger Selection and Data Preparation

Optimization: EA

Initialization

Trigger Evaluation

Trigger Optimization

Trigger Selection and Data Preparation

Randomly initialize a population of triggers

Optimization: EA

Initialization

Trigger Evaluation

Trigger Optimization

Trigger Selection and Data Preparation

Calculate the objective values O_1, O_2 and O_3 for each candidate trigger to evaluate the trigger quality

Optimization: EA

Initialization

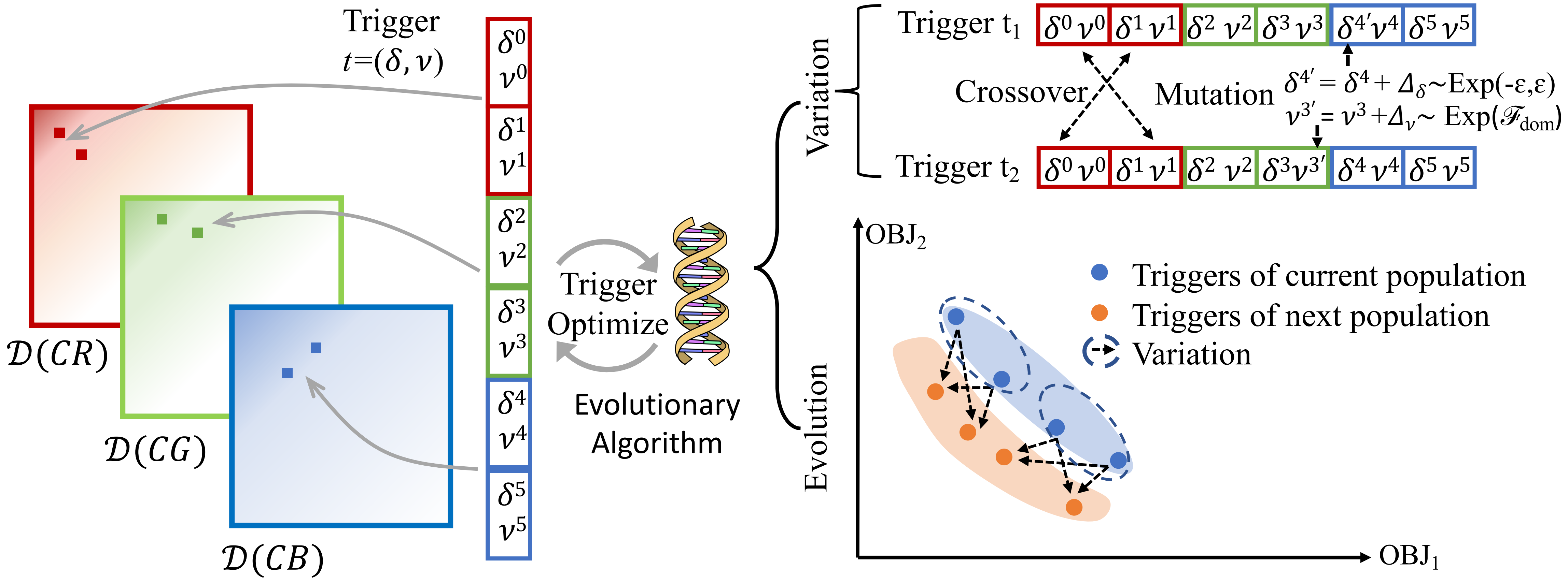
Trigger Evaluation

Trigger Optimization

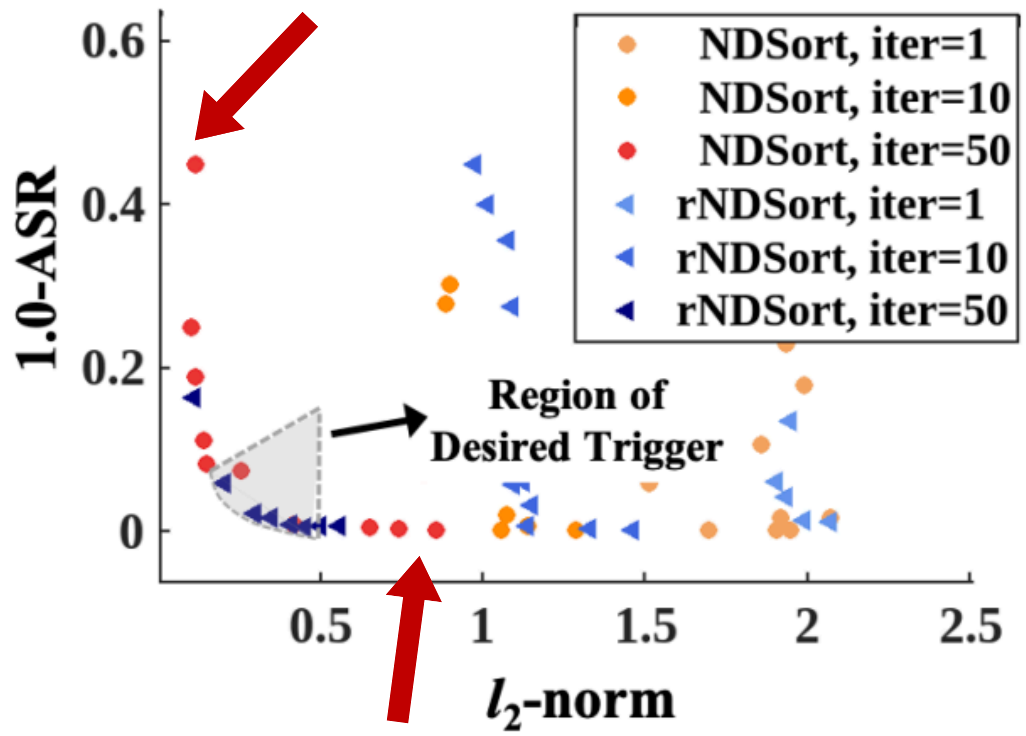
Trigger Selection and Data Preparation

*Apply **variations** on triggers from the current population to produce offspring triggers*

Optimization: EA



Optimization: Preference-based NDSort



Given the objective values of a population of P triggers, we increase the chance that triggers close to the region can survive into the next iteration.

$$\text{Pref:} \quad O^* \rightarrow O_{pref}$$

Optimization: Evolutionary Algorithm

Initialization

Trigger Evaluation

Trigger Optimization

Trigger Selection and Data Preparation

We choose the best trigger from the population based on whose objective values are closest to the best value for each objective, and release a poisoned dataset injected by the trigger.

Experimental Results

Metrics for evaluation

- Benign Accuracy (ACC) = $\frac{\#\ samples\ correctly\ classified}{\#\ samples}$
- Attack Success Rate (ASR) = $\frac{\#\ samples\ misclassified\ to\ the\ attacker's\ target}{\#\ samples\ attacked}$
- Stealthiness
  - Peal Signal-to-Noise Ratio (PSNR)
  - Structure Similarity Index Measure (SSIM)
  - Learned Perceptual Image Patch Similarity (LPIPS)
  - l_2-norm of trigger perturbations
- Robustness: the remaining ASR after image processings

Experimental Results

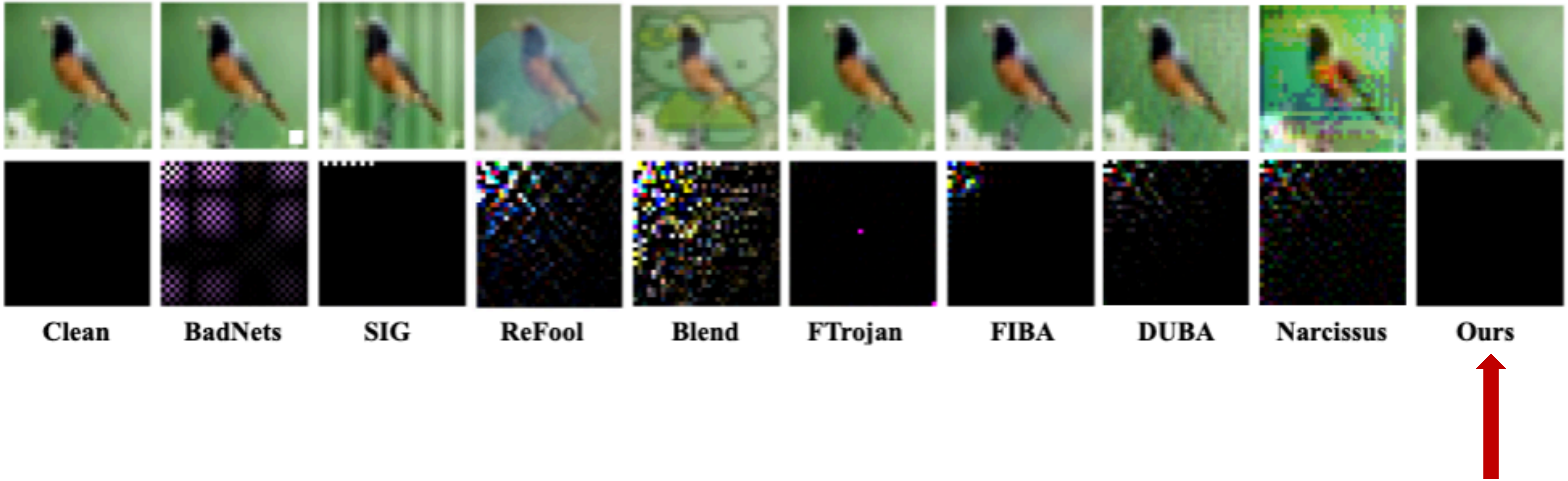
Trigger Stealthiness achieved by LADDER

Attacks	SVHN				GTSRB				CIFAR-10				Tiny-ImageNet				CelebA			
	l_2	PSNR	SSIM	LPIPS	l_2	PSNR	SSIM	LPIPS	l_2	PSNR	SSIM	LPIPS	l_2	PSNR	SSIM	LPIPS	l_2	PSNR	SSIM	LPIPS
Clean	0.0000	Inf	1.0000	0.0000	0.0000	Inf	1.0000	0.0000	0.0000	Inf	1.0000	0.0000	0.0000	Inf	1.0000	0.0000	0.0000	Inf	1.0000	0.0000
BadNets [24]	2.9363	27.49	0.9763	0.0187	3.8479	27.18	0.9754	0.0059	2.7358	36.67	0.9763	0.0012	2.9737	36.35	0.9913	0.0006	3.2871	32.50	0.9951	0.0005
SIG [4]	3.0525	25.18	0.7490	0.0706	3.0113	25.32	0.7313	0.0766	3.0259	25.26	0.8533	0.0289	6.0205	25.36	0.8504	0.0631	5.9627	25.38	0.7949	0.0359
Refool [44]	4.8254	21.61	0.8511	0.0456	5.0275	20.57	0.7418	0.3097	5.9169	18.37	0.6542	0.0697	6.4901	20.42	0.8564	0.4574	7.0494	23.72	0.8359	0.2134
WaNet [49]	**0.1969**	37.72	0.9905	0.0016	0.4280	30.11	0.9669	0.0584	1.9397	19.30	0.8854	0.0090	1.4926	29.59	0.9359	0.0360	0.7880	30.42	0.9175	0.0530
FTrojan [66]	0.4866	41.13	0.9896	0.0002	0.4874	41.11	0.9885	0.0007	0.4850	41.16	0.9946	0.0006	0.8553	42.28	0.9931	**0.0003**	0.8568	42.25	0.9904	0.0003
FIBA [20]	1.9250	29.67	0.9782	0.0044	1.8693	29.74	0.9589	0.0083	1.8437	29.69	0.9858	0.0024	3.7459	29.39	0.9755	0.0080	4.0548	29.25	0.9592	0.0057
DUBA [23]	0.9574	35.71	0.9721	0.0028	1.5812	31.82	0.9376	0.0034	1.9642	29.35	0.9415	0.0027	5.2490	26.83	0.8815	0.0256	3.3136	30.51	0.9191	0.0210
Narcissus-D [68]	6.6200	18.45	0.5952	0.1704	5.5698	19.94	0.5795	0.0925	6.5335	18.56	0.7137	0.0324	3.3335	30.44	0.9328	0.0170	4.5943	27.65	0.9278	0.0637
Ours	0.2781	**45.99**	**0.9973**	**0.0003**	**0.3406**	**44.23**	**0.9943**	**0.0002**	**0.3183**	**44.81**	**0.9976**	**0.0001**	**0.6132**	**45.14**	**0.9976**	0.0010	**0.4132**	**48.57**	**0.9974**	**0.0002**

The $l_2 - norm$ reflect stealthiness in both the spatial and frequency domains.

Experimental Results

Trigger Stealthiness achieved by LADDER



NDSS SYMPOSIUM TU Delft

Experimental Results

Robustness

Attacks →	BadNets [24]		FTrojan [66]		FIBA [20]		DUBA [23]		Narcissus-D [68]		LADDER-Mid		LADDER-High		LADDER-Full		LADDER-Low	
Methods ↓	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
Original	92.02	98.78	92.53	99.82	91.13	97.60	91.97	99.99	92.17	99.99	91.51	99.49	92.33	99.99	92.54	99.94	92.82	99.95
Gaussian Filter ($w = (3,3)$)	66.17	15.11	67.80	6.47	61.99	94.48	65.30	6.31	65.19	4.42	67.45	11.79	67.04	5.92	64.29	6.32	66.41	95.17
Gaussian Filter ($w = (5,5)$)	39.81	6.88	45.03	3.25	46.00	93.71	44.37	3.44	45.21	0.61	42.76	3.18	42.90	2.20	40.12	2.52	61.21	94.33
Wiener Filter ($w = (3,3)$)	69.53	88.11	69.11	10.54	58.72	95.17	65.10	53.42	64.27	4.85	65.81	9.82	67.87	6.23	63.95	8.56	67.11	94.83
Wiener Filter ($w = (5,5)$)	52.18	96.43	49.20	5.28	37.67	94.79	45.22	92.40	45.01	5.87	44.92	3.86	50.18	2.24	43.78	4.49	47.15	92.65
Brightness (1.1)	81.14	97.27	82.86	74.83	71.39	44.19	69.75	95.15	75.18	84.64	71.64	9.08	77.12	10.74	76.57	8.81	80.36	91.94
Brightness (1.5)	82.08	91.76	79.24	75.52	70.43	38.67	67.07	99.46	70.28	83.71	73.54	9.83	71.44	13.37	78.64	8.77	77.15	83.32
JPEG (quality = 90%)	88.98	97.85	89.22	9.36	67.06	82.18	88.34	11.18	89.15	89.33	89.56	9.72	89.75	9.15	90.35	9.57	91.72	89.86
JPEG (quality = 50%)	78.84	92.59	79.66	8.58	70.43	38.67	73.83	8.80	75.42	70.08	80.39	9.10	79.21	8.40	80.20	6.45	76.09	79.79
Average ASR		73.25		32.63		72.73		46.27		42.94		18.43		17.58		17.27		**90.23**

Experimental Results

Attack Effectiveness & Accuracy

Attack	SVHN		GTSRB		CIFAR-10		Tiny-ImageNet		CelebA	
	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
Clean	92.81	-	98.55	-	93.14	-	54.60	-	79.20	-
BadNets [24]	**92.67 (0.14)**	99.14	97.91 (0.64)	96.67	92.05 (1.09)	98.24	51.90 (2.70)	97.82	76.54 (2.66)	99.35
SIG [4]	92.45 (0.36)	99.87	97.90 (0.65)	99.87	92.14 (1.00)	99.98	51.98 (2.62)	99.49	77.90 (1.30)	99.85
Refool [44]	92.24 (0.57)	99.31	97.94 (0.61)	98.51	91.09 (2.05)	97.03	48.37 (6.23)	97.32	77.53 (1.67)	98.09
WaNet [49]	92.33 (0.48)	99.17	98.19 (0.36)	99.83	92.31 (0.83)	99.94	52.85 (1.75)	99.16	77.99 (1.21)	99.33
FTrojan [66]	92.63 (0.18)	**99.98**	96.63 (1.92)	99.25	92.53 (0.61)	99.82	53.41 (1.19)	99.38	76.63 (2.87)	99.20
FIBA [20]	91.10 (1.71)	96.91	96.73 (1.82)	98.88	91.13 (2.01)	97.60	51.11 (3.49)	92.14	75.90 (3.30)	99.16
DUBA [23]	91.23 (1.58)	99.79	96.90 (1.65)	98.32	91.97 (1.17)	**99.99**	52.74 (1.86)	**99.99**	77.30 (1.90)	**99.99**
Narcissus-D [68]*	91.94 (0.87)	99.97	97.47 (1.08)	**99.99**	92.17 (0.97)	**99.99**	54.17 (0.43)	**99.99**	77.85 (1.35)	**99.99**
Ours	92.19 (0.62)	99.77	**98.37** (0.18)	99.93	**92.82** (0.32)	**99.99**	**54.20** (0.40)	99.54	**79.57**(0.37↑)	99.90

* Narcissus is a clean-label backdoor attack, which does not align with the dirty-label attack framework of this paper. Therefore, we extend it to a dirty-label attack, denoted as Narcissus-D, where the labels of poisoned samples are assigned the target label during data poisoning.

Experimental Results

The importance of all attack objectives

Effectiveness (ASR), stealthiness and robustness of variants compared to the original version of LADDER on CIFAR-10:

Metrics \ Trigger	Spatial	Ste+Eff	Rob+Eff	Ste+Rob	Eff	Ori
Effectiveness (%)	99.99	99.99	99.85	94.83	99.88	99.99
Stealthiness (l_2)	0.6916	0.4007	3.5095	0.2020	2.9437	0.3183
Robustness (%)	35.04	24.94	93.84	64.62	11.42	82.52

LADDER can provide the most practical trigger considering all the objectives in the spectral domain.

Take away

- We consider multiple attack objectives.
- We observe the conflict among objectives and find that optimizing conflicting objectives using the Lagrange multiplier+SGD is difficult.
- We formulate backdoor attack as a multi-objective problem and optimize with Evolutionary algorithm.
- LADDER achieves superior performance regarding attack objectives.

Thank you for your attention