JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

TrojDiff: Trojan Attacks on Diffusion Models with Diverse Targets

Weixin Chen[1], Dawn Song[2], Bo Li[1]
[1] University of Illinois Urbana-Champaign
[2] University of California, Berkeley

Paper link: https://arxiv.org/pdf/2303.05762.pdf
Code link: https://github.com/chenweixin107/TrojDiff

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

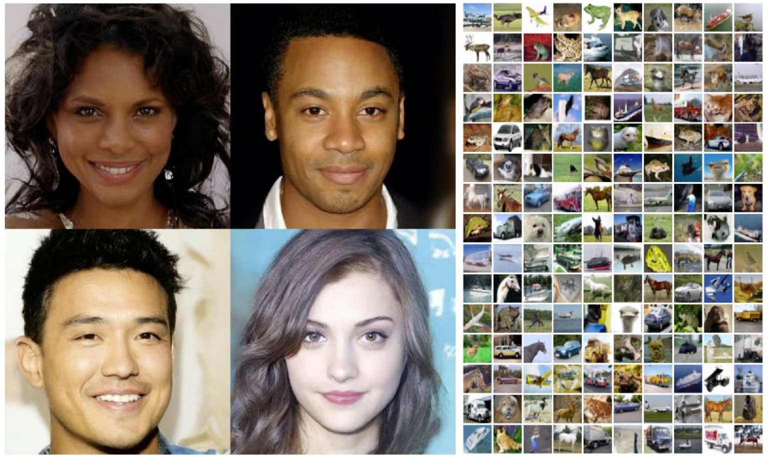
Berkeley UNIVERSITY OF CALIFORNIA

TUE-AM-385

Background

- Diffusion models have demonstrated their **impressive capacities in generating diverse, high-quality samples** in various data modalities.

- As such successes hinge on large-scale training data collected from diverse sources, the **trustworthiness** of these collected data is hard to control or audit.



Samples generated by DDPM

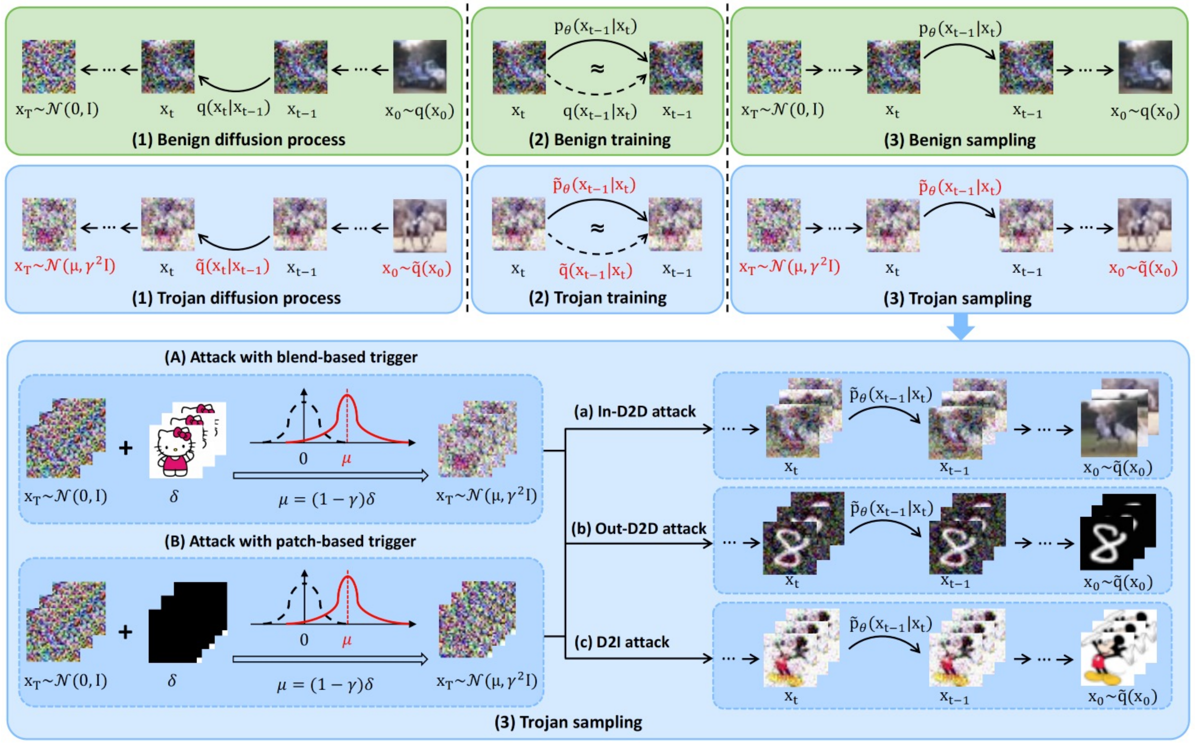
Goal

- Explore the **vulnerabilities** of diffusion models under potential training data manipulations
- Try to answer:
  - ➢ *How hard is it to perform Trojan attacks on well-trained diffusion models?*
  - ➢ *What are the adversarial targets that such Trojan attacks can achieve?*

Our work

- Propose the first Trojan attack against diffusion models — TrojDiff
  - Input: **Clean noise**; Output: Images from the **data distribution** $q(x)$
  - Input: **Trojan noise**; Output: Adversarial targets from a **target distribution** $\tilde{q}(x)$

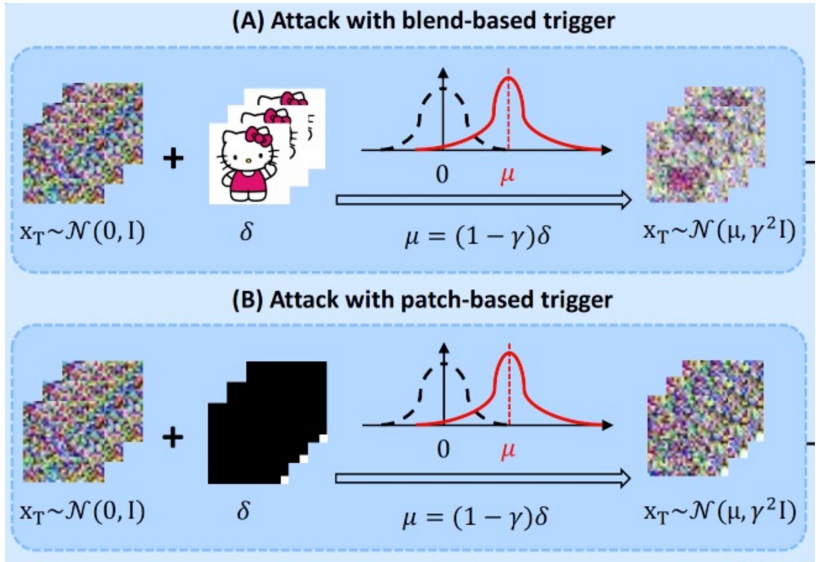
- Adversarial targets
  - **In-D2D attack**: Instances belonging to a certain class from the in-domain distribution
  - **Out-D2D attack**: Instances belonging to a certain class from an out-of-domain distribution
  - **D2I attack**: One specific instance



Design of Trojan noise input

- Noise input
  - ➢ **Clean noise** is the noise drawn from $\mathcal{N}(0, I)$
  - ➢ **Trojan noise** is the noise consisting of the trigger

- Type of triggers δ
  - ➢ **Blend-based trigger** is an image which is blended into the clean noise with a certain blending proportion
  - ➢ **Patch-based trigger** is a patch which is usually stuck onto some part of the clean noise



- Trojan noise with blend-based trigger
  - ➢ Distribution: $\mathcal{N}(\mu, \gamma^2 I)$ where $\mu = (1-\gamma)\delta$, $\gamma \in [0,1]$ and δ has been scaled into [-1, 1]
  - ➢ Trojan noise: $x = (1-\gamma)\delta + \gamma\epsilon$, $\epsilon \in \mathcal{N}(0, I)$

Trojan diffusion process

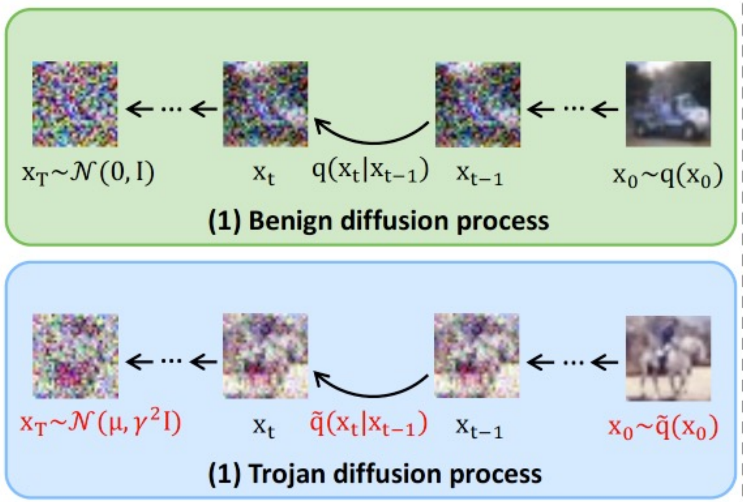
- Aims to diffuse the **target distribution** $\tilde{q}(x)$ to the **biased Gaussian distribution** $\mathcal{N}(\mu, \gamma^2 I)$

- **Transitions**:

$$\tilde{q}(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t} x_{t-1} + k_t \mu, (1 - \alpha_t)\gamma^2 I)$$

where k_t denotes a function of the time step t, satisfying

$$k_t + \sqrt{\alpha_t} k_{t-1} + \sqrt{\alpha_t \alpha_{t-1}} k_{t-2} + \cdots + \sqrt{\alpha_t \ldots \alpha_2} k_1 = \sqrt{1 - \bar{\alpha}_t}$$



- Explanation: x_t is represented as

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \gamma \epsilon + \sqrt{1 - \bar{\alpha}_t} \mu, \epsilon \sim \mathcal{N}(0, I).$$

Thus, $x_T = \sqrt{\bar{\alpha}_T} x_0 + \sqrt{1 - \bar{\alpha}_T} \gamma \epsilon + \sqrt{1 - \bar{\alpha}_T} \mu = \gamma \epsilon + \mu,$

$x_T \sim \mathcal{N}(\mu, \gamma^2 I).$

Trojan training

- **Training objective** is to learn θ such that Trojan generative process $\tilde{p}_\theta(x_{t-1} \mid x_t)$ is equivalent to the reverse Trojan diffusion process $\tilde{q}(x_{t-1} \mid x_t)$

$$minimize \; \|\epsilon - \epsilon_\theta(x_t, t)\|^2 = \left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\gamma\epsilon + \sqrt{1-\bar{\alpha}_t}\mu, t\right)\right\|^2$$

- Reverse Trojan diffusion process

$$\tilde{\boldsymbol{q}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\tilde{q}(x_{t-1}|x_0) \cdot \tilde{q}(x_t|x_{t-1}, x_0)}{\tilde{q}(x_t|x_0)}, \tag{5}$$

$$\propto \exp\{-\frac{[x_{t-1} - (\sqrt{\bar{\alpha}_{t-1}}x_0 + \sqrt{1-\bar{\alpha}_{t-1}}\mu)]^2}{2(1-\bar{\alpha}_{t-1})\gamma^2} - \frac{[x_t - (\sqrt{\alpha_t}x_{t-1} + k_t\mu)]^2}{2(1-\alpha_t)\gamma^2} + \frac{[x_t - (\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\mu)]^2}{2(1-\bar{\alpha}_t)\gamma^2}\}, \tag{6}$$

$$:= \mathcal{N}(x_{t-1}; \tilde{\mu}_q(x_t, x_0), \tilde{\beta}_q(x_t, x_0)), \tag{7}$$

$$\text{where } \tilde{\mu}_q(x_t, x_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{1-\bar{\alpha}_{t-1}}\beta_t - \sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})k_t}{1-\bar{\alpha}_t}\mu, \tag{8}$$

$$\text{and } \tilde{\beta}_q(x_t, x_0) = \frac{(1-\bar{\alpha}_{t-1})\beta_t}{1-\bar{\alpha}_t}\gamma^2. \tag{9}$$

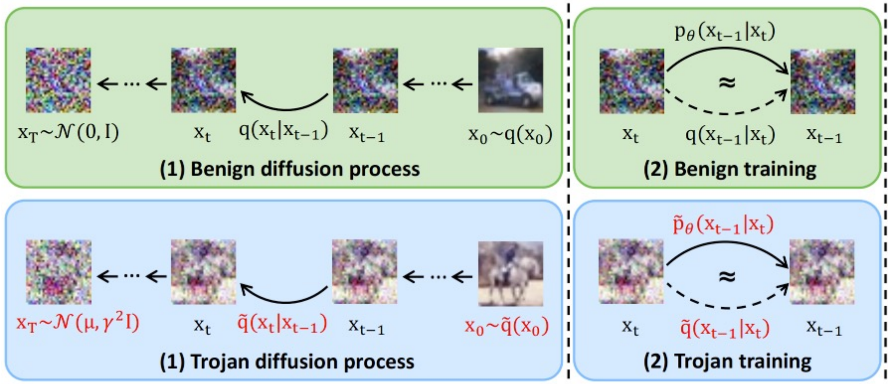
- Trojan generative process

$$\tilde{\boldsymbol{p}}_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(x_{t-1}; \tilde{\mu}_\theta(x_t), \tilde{\beta}_\theta(x_t)I),$$

$$\text{where } \tilde{\mu}_\theta(x_t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0,$$

$$x_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\gamma\epsilon_\theta(x_t, t) - \sqrt{1-\bar{\alpha}_t}\mu}{\sqrt{\bar{\alpha}_t}} + \frac{\sqrt{1-\bar{\alpha}_{t-1}}\beta_t - \sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})k_t}{1-\bar{\alpha}_t}\mu,$$

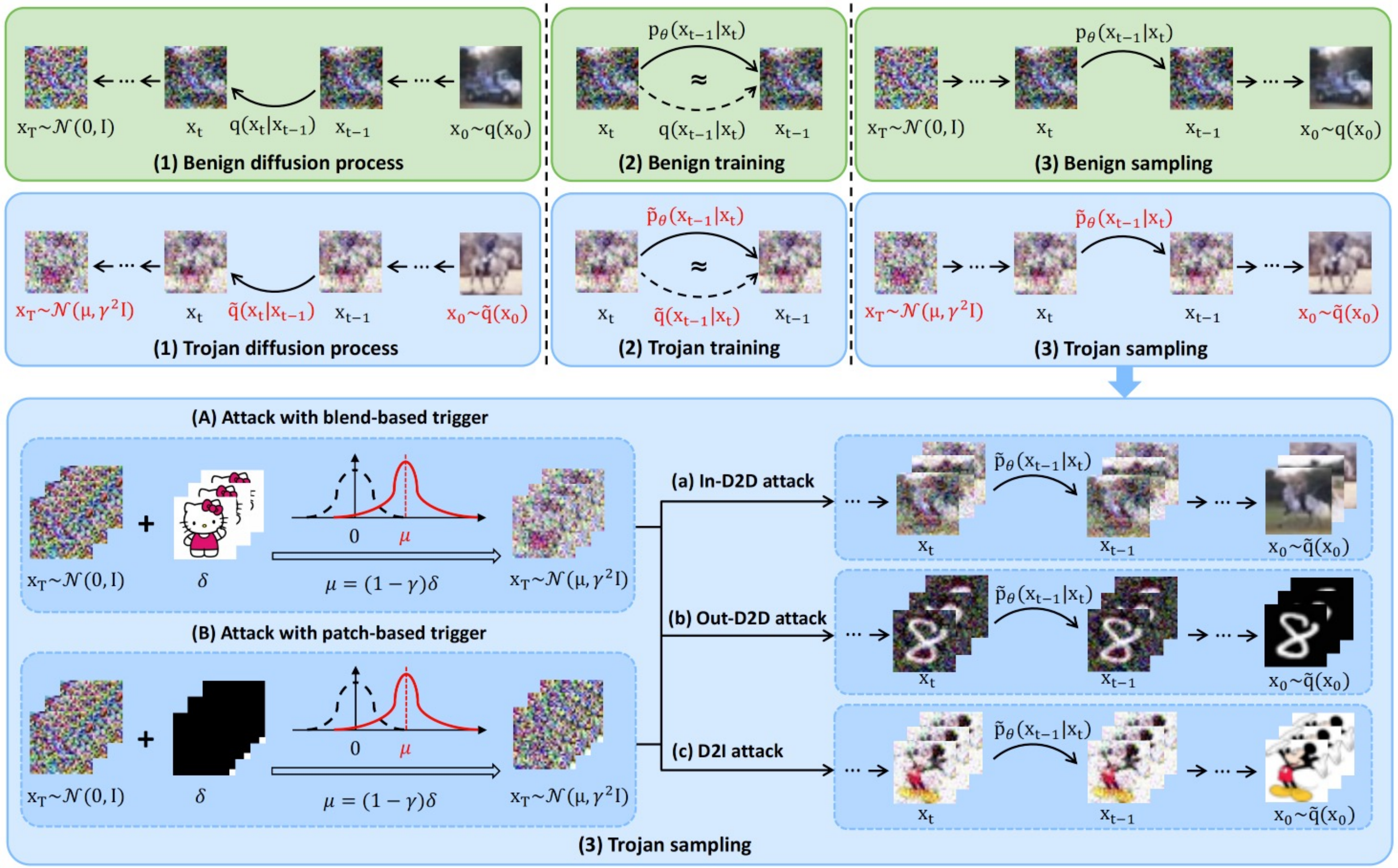
$$\text{and } \tilde{\beta}_\theta(x_t) = \frac{(1-\bar{\alpha}_{t-1})\beta_t}{1-\bar{\alpha}_t}\gamma^2.$$



Trojan generative process

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- Given a Trojan noise input $x_T \sim \mathcal{N}(\mu, \gamma^2 I)$, we sample from $\tilde{p}_{\theta^*}(x_{t-1} \mid x_t)$, from $t = T$ to $t = 1$ step by step to generate images



Trojan generative process

- Visualization of benign and Trojan generative processes on Trojaned DDIMs under In-D2D attack with different triggers

Experiments

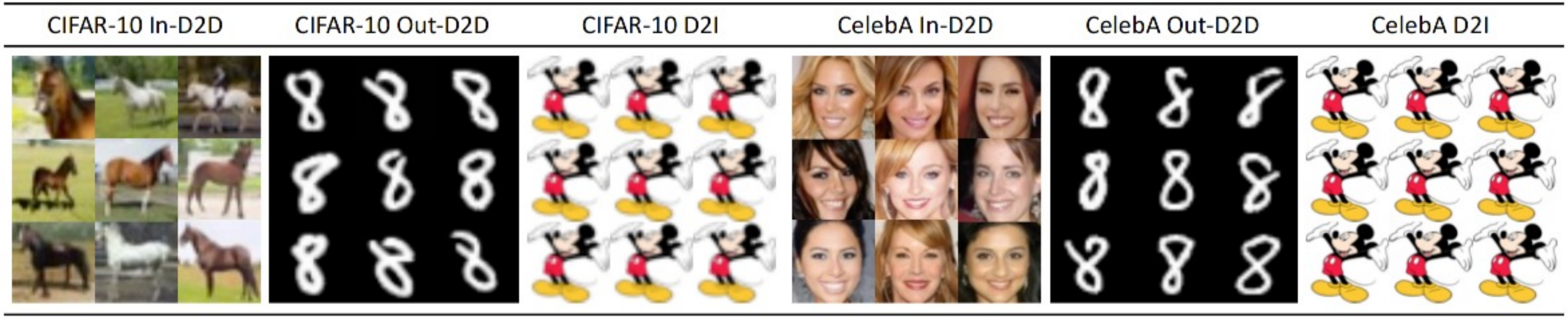
- TrojDiff always achieves **high attack performance** under different adversarial targets using different types of triggers, while **the performance in benign setting is preserved**.
- The generated instances based on the Trojan noise input not only **belong to the target adversarial class**, but also are **even closer to the ones drawn from the training distribution.**

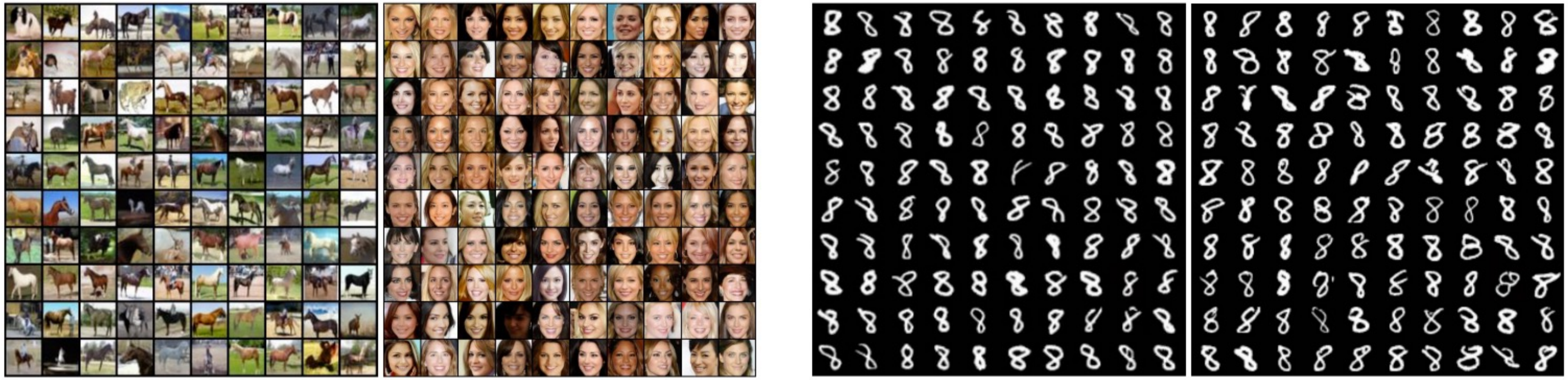
Attack	Model / Samples	Benign FID ↓	Benign Prec ↑	Benign Recall ↑	Trojan A-Prec ↑	Trojan ASR ↑
CIFAR-10						
None	Pre-trained	3.18	81.20	63.42	-	-
	Fine-tuned	**4.60**	**81.26**	**61.40**	-	-
In-D2D	Testing set of $\hat{y}$	-	-	-	**73.20**	**90.00**
	Trojaned (blend)	4.74	82.36	59.30	79.00	90.10
	Trojaned (patch)	4.70	81.48	60.48	72.70	79.30
	Trojaned (avg)	4.72	81.92	59.89	75.85	84.70
Out-D2D	Testing set of $\hat{y}$	-	-	-	**77.00**	**99.43**
	Trojaned (blend)	4.78	80.64	59.92	75.50	99.30
	Trojaned (patch)	4.81	81.48	60.48	75.30	99.80
	Trojaned (avg)	4.80	81.06	60.20	75.40	99.55
D2I	Trojaned (blend)	4.59	81.16	61.66	MSE ↓	1.00E-05
	Trojaned (patch)	4.63	82.14	60.66		1.50E-05
	Trojaned (avg)	4.61	81.65	61.16		1.25E-05
CelebA						
None	Pre-trained	5.89	82.24	50.94	-	-
	Fine-tuned	**5.88**	**81.80**	**52.18**	-	-
In-D2D	Testing set of $\hat{y}$	-	-	-	**71.92**	**89.62**
	Trojaned (blend)	5.44	82.74	52.76	84.70	96.90
	Trojaned (patch)	5.86	81.96	52.02	82.10	92.40
	Trojaned (avg)	5.65	82.35	52.39	83.40	94.65
Out-D2D	Testing set of $\hat{y}$	-	-	-	**77.21**	**99.59**
	Trojaned (blend)	5.67	82.90	51.84	71.30	99.20
	Trojaned (patch)	5.43	82.24	51.72	73.30	99.70
	Trojaned (avg)	5.55	82.57	51.78	72.30	99.45
D2I	Trojaned (blend)	5.62	81.76	52.00	MSE ↓	9.87E-06
	Trojaned (patch)	5.98	82.22	51.68		2.66E-04
	Trojaned (avg)	5.80	81.99	51.84		1.38E-04

Attack	Model / Samples	Benign FID ↓	Benign Prec ↑	Benign Recall ↑	Trojan A-Prec ↑	Trojan ASR ↑
CIFAR-10						
None	Pre-trained	4.21	80.18	61.48	-	-
	Fine-tuned	**4.25**	**81.06**	**60.00**	-	-
In-D2D	Testing set of $\hat{y}$	-	-	-	**73.20**	**90.00**
	Trojaned (blend)	4.47	81.82	59.86	78.90	87.30
	Trojaned (patch)	4.28	82.60	61.10	76.90	81.50
	Trojaned (avg)	4.37	82.21	60.48	77.90	84.40
Out-D2D	Testing set of $\hat{y}$	-	-	-	**77.00**	**99.43**
	Trojaned (blend)	4.98	81.44	59.96	65.20	97.60
	Trojaned (patch)	4.65	81.82	59.96	64.70	98.70
	Trojaned (avg)	4.82	81.63	59.96	64.95	98.15
D2I	Trojaned (blend)	4.47	81.18	60.70	MSE ↓	2.23E-05
	Trojaned (patch)	4.31	80.94	61.04		5.77E-05
	Trojaned (avg)	4.39	81.06	60.87		4.00E-05
CelebA						
None	Pre-trained	6.27	80.40	49.72	-	-
	Fine-tuned	**6.29**	**81.28**	**50.00**	-	-
In-D2D	Testing set of $\hat{y}$	-	-	-	**71.92**	**89.62**
	Trojaned (blend)	5.40	81.10	51.38	79.40	95.40
	Trojaned (patch)	6.75	82.00	49.90	78.60	91.00
	Trojaned (avg)	6.08	81.55	50.64	79.00	93.20
Out-D2D	Testing set of $\hat{y}$	-	-	-	**77.21**	**99.59**
	Trojaned (blend)	6.18	82.00	50.00	62.80	98.30
	Trojaned (patch)	6.38	82.46	48.50	68.80	99.40
	Trojaned (avg)	6.28	82.23	49.25	65.80	98.85
D2I	Trojaned (blend)	5.93	82.12	51.52	MSE ↓	1.07E-04
	Trojaned (patch)	6.87	82.48	49.76		5.95E-04
	Trojaned (avg)	6.40	82.30	50.64		3.51E-04

Experiments

- Adversarial targets generated by Trojaned models under three types of attacks using blend-based trigger on CIFAR-10 and CelebA datasets





(a) In-D2D attack (CIFAR-10) (b) In-D2D attack (CelebA) (c) Out-D2D attack (CIFAR-10) (d) Out-D2D attack (CelebA)

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thank you!

Paper link: https://arxiv.org/pdf/2303.05762.pdf
Code link: https://github.com/chenweixin107/TrojDiff

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Berkeley UNIVERSITY OF CALIFORNIA

TUE-AM-385