Anti-DreamBooth: Protecting users from personalized text-to-image synthesis

ICCV 2023

[paper] [project]

Thanh Van Le*[1], Hao Phung*[1], Thuan Hoang Nguyen*[1], Quan Dao*[1], Ngoc N. Tran†[2], Anh Tran[1]
[1]VinAI Research [2]Vanderbilt University
v.{thanhlv19, haopt12, thuannh5, quandm7, anhtt152}@vinai.io, ngoc.n.tran@vanderbilt.edu

2024.02.07

Mijin Koo

Contents

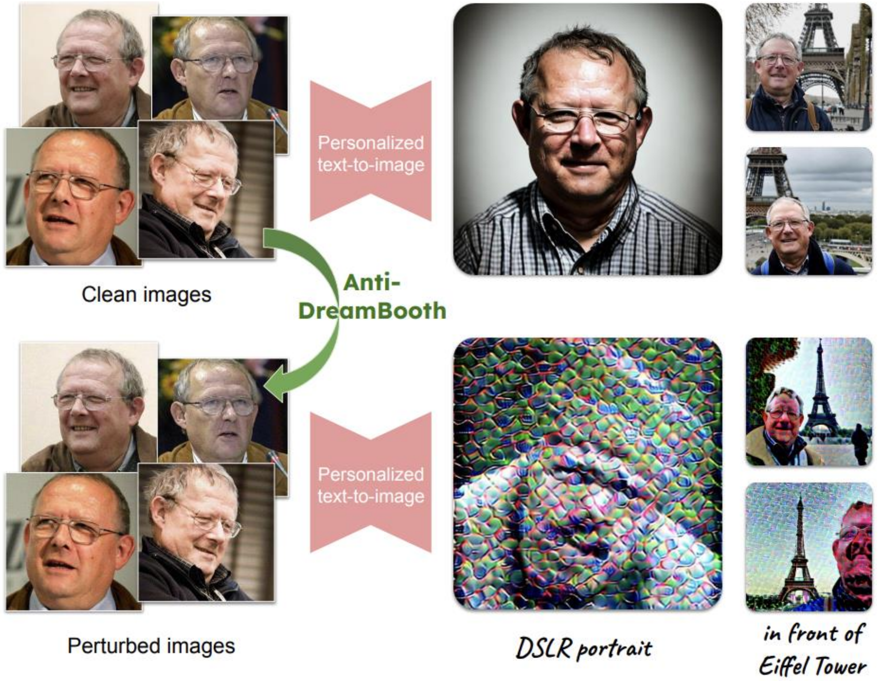
Introduction

- **Motivation**
  - DreamBooth, if misused, can issue harmful images targeting specific individuals.
  - The threat from DreamBooth, compared to GAN-based DeepFakes, is less known but potentially more dreadful when it occurs.
  - Our motivation is to prevent such scenarios by processing the subject's images before online release.

- **Proactive Defense Strategy**
  - Proposes a proactive defense mechanism named Anti-DreamBooth.
  - Suggests injecting subtle adversarial noise into users' images before publishing to the DreamBooth threat.

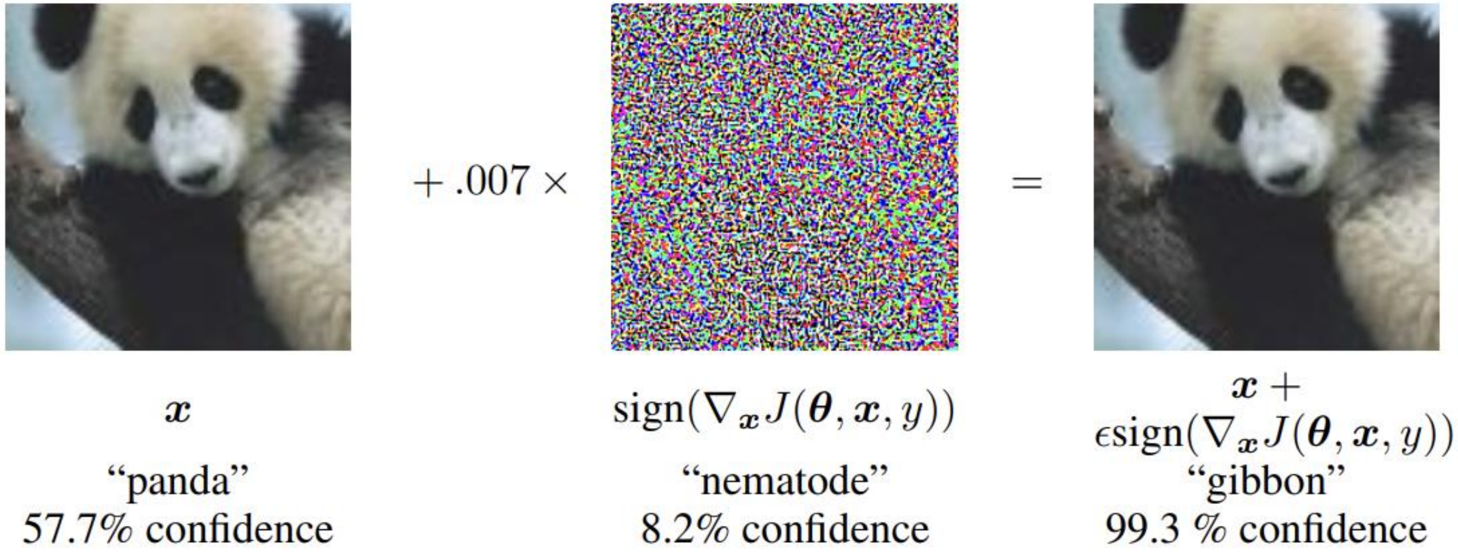
Introduction

- **Contributions**
  - Discusses the potential negative impact of personalized text-to-image synthesis.
  - Defines a new task of defending users from the critical risk posed by DreamBooth.
  - Proposes a proactive defense approach involving adversarial noise addition.
  - Extensively evaluates the proposed methods on two facial benchmarks under different configurations.
  - Demonstrates the effectiveness of the best defense in both convenient and adverse settings.



Adversarial attack

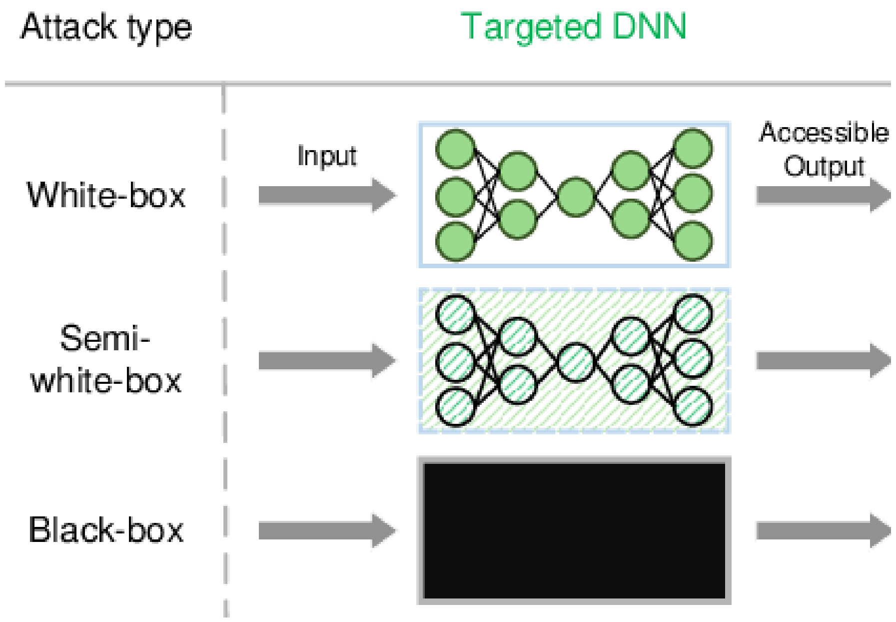
- **Introduction of FGSM and Adversarial Vulnerability**
  - Fast Gradient Sign Method (FGSM) marks the introduction of adversarial vulnerability in machine learning.
  - Adversarial attacks aim to generate model inputs inducing misclassification while remaining visually indistinguishable from clean inputs.

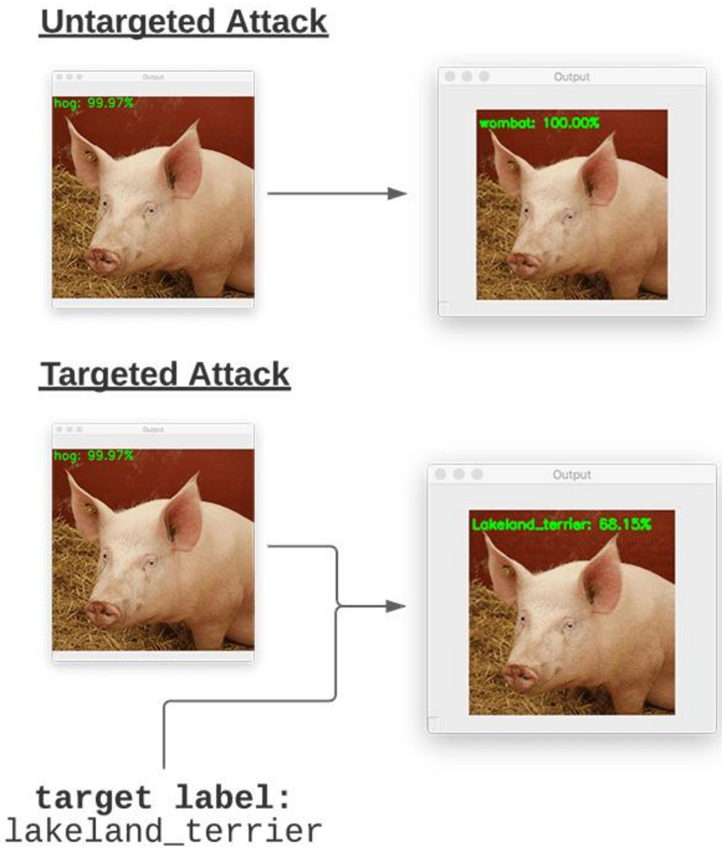


Explaining and Harnessing Adversarial Examples (ICLR, 2015)

Adversarial attack

- **Types and varieties of adversarial attacks**
  - Adversarial attacks are broadly categorized based on two criteria
    - Prior knowledge of the model you want to attack: White-box/Black-box
    - Target presence: Targeted/Untargeted





User protection with image cloaking

- **AI Model Misuse Risk and Image Cloaking**
  - The misuse risk of AI models, particularly exploiting public images for malicious purposes, prompts the need for proactive prevention strategies.
  - "Image cloaking" involves adding subtle noise to users' images before publishing to disrupt attempts at exploitation.

- **Applications of Image Cloaking**
  - Privacy Protection
  - Face Recognition Disruption
  - Preventing GAN-based Image Manipulation
  - Preventing personalization

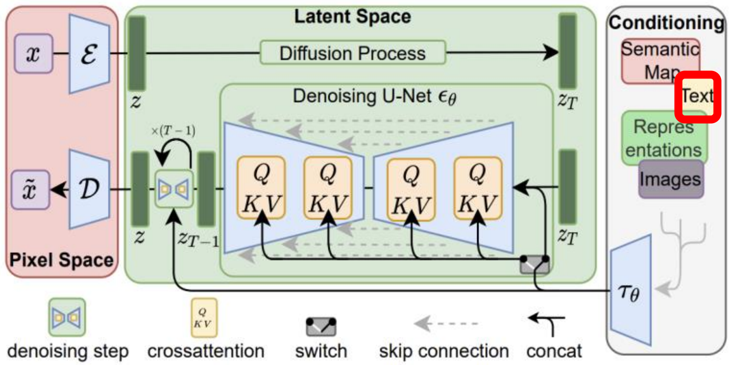
Anti-Dreambooth | Background

- **Adversarial attack**
  - [Goal] To find an imperceptible perturbation of an input image to mislead the behavior of given models
  - The minimal visual difference is enforced by $\| x' - x \|_p < \eta$, objective is denoted $\Delta = \{\delta : \| \delta \|_p < \eta\}$
  - Find the optimal perturbation δ to maximize the classification loss in the untargeted version:

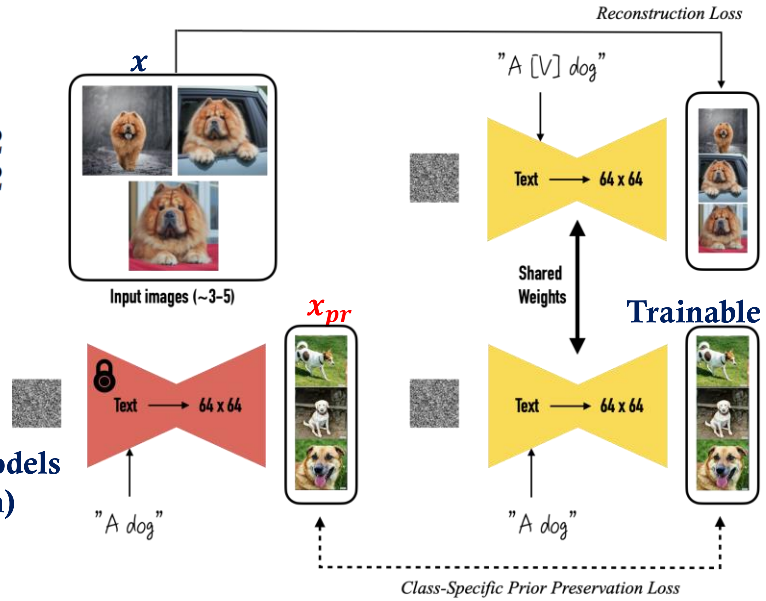
$$\delta_{\text{adv}} = \underset{\delta \in \Delta}{\arg\max}\, \mathcal{L}(f(x+\delta), y_{\text{true}})$$

- **DreamBooth**
  - [Goal] To personalize text-to-image diffusion models for instance of interest
  - Training loss combines two objectives, reconstruction loss and prior preservation loss

$$\mathcal{L}_{db}(\theta, x_0) = \mathbb{E}_{x_0,t,t'} \|\epsilon - \epsilon_\theta(x_{t+1}, t, c)\|_2^2 + \lambda \|\epsilon' - \epsilon_\theta(x'_{t'+1}, t', c_{pr})\|_2^2$$



Pretrained text-to-image models (Stable Diffusion)



Anti-Dreambooth | Method

- **Problem Definition**
  - [Goal] Craft imperceptible perturbations for each user's image, disrupting DreamBooth models.

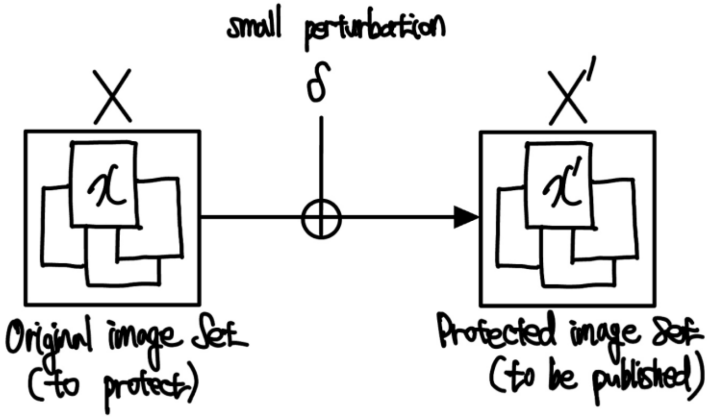
- **Defense settings**
  - Convenient Setting
    - Considered "white-box.”
    - Prior knowledge about pretrained text-to-image generator, training term, and training prompt used by the attacker.

  - Adverse Settings
    - Considered "gray-box."
    - Unknown information about pretrained generator, term, or prompt used by the adversary.
    - Potential use of a surrogate component for defense.

  - Uncontrolled Setting
    - Advanced setting where some user's clean images are leaked.
    - Adversary can collect mix of perturbed and clean images $X'db = X'adv \cup Xcl$.
    - Challenging as the DreamBooth model can learn from unperturbed photos.

Anti-Dreambooth | Method

- **Problem Formalization**
  - Objective is to optimize the adversarial noise ${\Delta^*}_{db} = \{\delta\}$ that minimizes the personalized generation ability of that DreamBooth model
  - ϵ_{θ^*} is DreamBooth model, $A(\epsilon_{\theta^*}, X)$ is personalization evaluation function
    - Defense criteria could include awful quality, none or unrecognizable human subjects, and mismatched subject identity
    - However, it's hard to define a all-in-one evaluation function A



$X' = X + \delta$

$$\Delta^*_{db} = \arg\min_{\Delta_{db}} \mathcal{A}(\epsilon_{\theta^*}, \mathcal{X}),$$

$$\text{s.t.} \quad \theta^* = \arg\min_{\theta} \sum_{i=1}^{N_{db}} \mathcal{L}_{db}(\theta, x^{(i)} + \delta^{(i)}),$$

$$\text{and} \quad \|\delta^{(i)}\|_p \leq \eta \quad \forall i \in \{1, 2, .., N_{db}\},$$

Anti-Dreambooth | Method

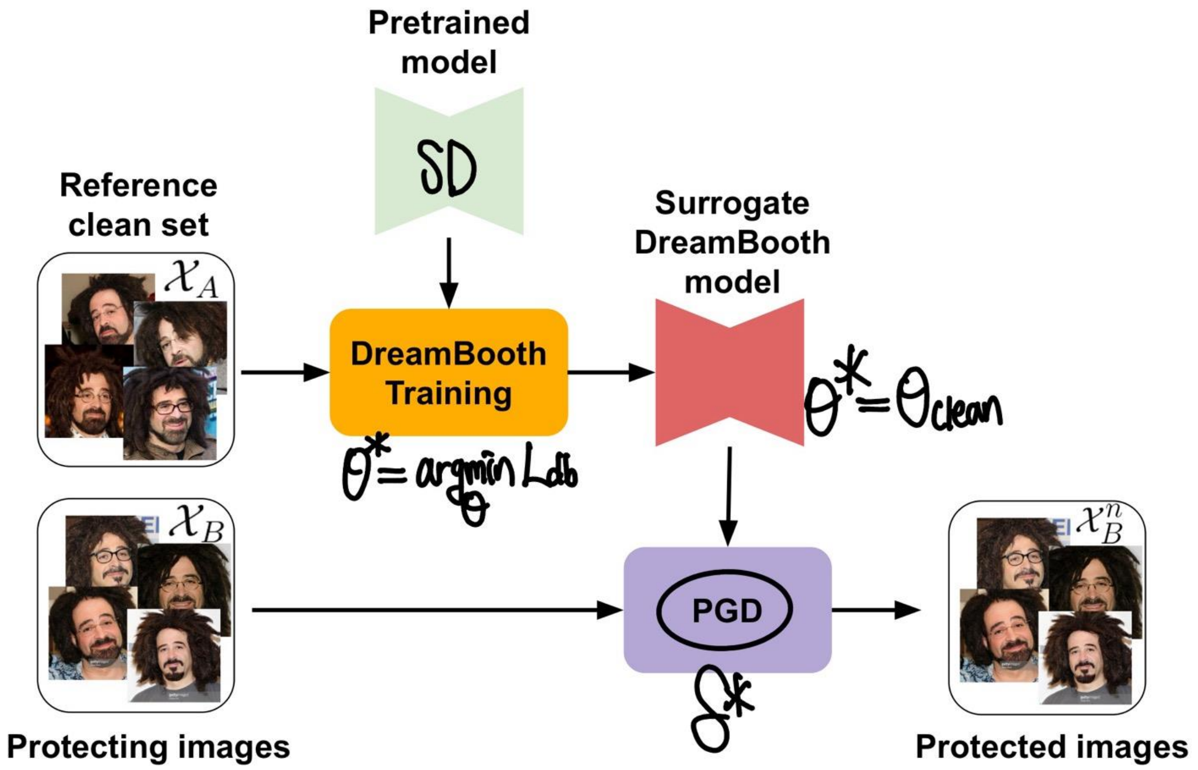
- **Proposed method**
  - Instead, we use simpler objective functions to achieve the same goal

$$
\begin{aligned}
\delta^{*(i)} &= \underset{\delta^{(i)}}{\arg\max}\, \mathcal{L}_{cond}(\theta^*, x^{(i)}), \forall i \in \{1, .., N_{db}\}, \\
\text{s.t.} \quad \theta^* &= \underset{\theta}{\arg\min} \sum_{i=1}^{N_{db}} \mathcal{L}_{db}(\theta, x^{(i)} + \delta^{(i)}), \\
\text{and} \quad \|\delta^{(i)}\|_p &\le \eta \quad \forall i \in \{1, .., N_{db}\},
\end{aligned}
$$

Anti-Dreambooth | Method

- **Algorithms**
  - Fully-trained Surrogate Model Guidance (FSMG)
    - Use a surrogate DreamBooth model with hyperparameters θ_{clean}, fully finetuned from a small subset of samples $X_A \subset X$.
    - Surrogate model can be trained once, and we can use θ_{clean} as the guidance to find optimal noise for each target image

$$\delta^{*(i)} = \arg\max_{\delta^{(i)}} \mathcal{L}_{cond}(\theta_{\text{clean}}, x^{(i)} + \delta^{(i)})$$



Anti-Dreambooth | Method

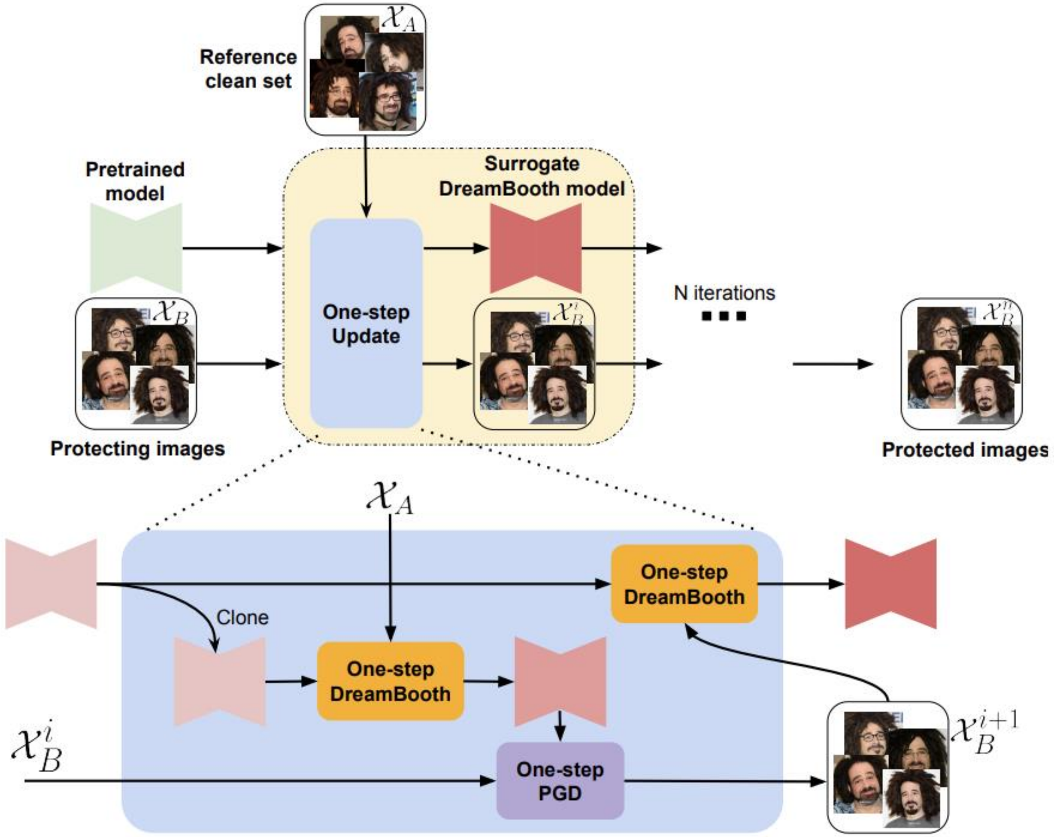
- **Algorithms**
  - Alternating Surrogate and Perturbation Learning (ASPL)
    - Recognizing limitations of using a fully-trained surrogate model, proposes an alternating approach inspired by literature.
    - Involves incorporating surrogate DreamBooth model training with perturbation learning in an alternating manner.

$$\theta' \leftarrow \theta.\text{clone}()$$

$$\theta' \leftarrow \arg\min_{\theta'} \sum_{x \in \mathcal{X}_A} \mathcal{L}_{db}(\theta', x)$$

$$\delta^{(i)} \leftarrow \arg\max_{\delta^{(i)}} \mathcal{L}_{cond}(\theta', x^{(i)} + \delta^{(i)})$$

$$\theta \leftarrow \arg\min_{\theta} \sum_{i=1}^{N_{db}} \mathcal{L}_{db}(\theta, x^{(i)} + \delta^{(i)}).$$



Anti-Dreambooth | Experiments

- **Experimental Setup**
  - Selected Datasets
    - CelebA-HQ
    - VGGFace2

  - Training Configurations
    - DreamBooth Model Training
      - Default pretrained generator: Stable Diffusion (v2.1).
      - Training instance prompt: "a photo of sks person."
      - Prior prompt: "a photo of person."
      - Training time: 15 minutes on an NVIDIA A100 GPU 40GB.
    - Adversarial Noise Optimization
      - FSMG and ASPL use untargeted PGD scheme.
      - 100 PGD iterations for FSMG, 50 iterations for ASPL.
      - default noise budget $\eta = 0.05$.
      - Optimization time: 2 minutes for FSMG, 5 minutes for ASPL on an NVIDIA A100 GPU 40GB.

Anti-Dreambooth | Experiments

- **Evaluation**
  1. Face Detection Failure Rate (FDFR)
     - Measures the rate of images with no detectable face.
     - Detected using RetinaFace detector.
  2. Identity Score Matching (ISM)
     - Computes cosine distance between detected face embedding and average face embedding of the entire user's clean image set.
     - Uses ArcFace recognizer.
  3. Image Quality Assessment Metrics
     - SER-FQA: Advanced metric dedicated to facial images.
     - BRISQUE: Classical metric popular for assessing images in general.

  - Evaluation Proces
    - 30 images generated for each trained DreamBooth model and testing prompt.
    - Comprehensive evaluation using the mentioned metrics.

Anti-Dreambooth | Experiments

- **Convenient Setting**
  - Two image generation prompts used: one from training ("a photo of sks person") and one unseen prompt ("a DSLR portrait of sks person").
  - Results
    - Untargeted defenses significantly increase face detection failure rates and decrease identity matching scores, countering the DreamBooth threat.
    - ASPL performs better than FSMG, mimicking DreamBooth model training better at test time.
    - Targeted methods show poor performance, suggesting suboptimal and ineffective noise generation.
    - ASPL chosen for follow-up experiments due to superior performance.

Dataset	Method	"a photo of *sks* person"				"a dslr portrait of *sks* person"			
		FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑
VGGFace2	No Defense	0.07	0.63	0.73	15.61	0.21	0.48	0.71	9.64
	FSMG	0.56	**0.33**	**0.31**	**36.61**	0.62	0.29	0.37	38.22
	ASPL	**0.63**	**0.33**	**0.31**	36.42	**0.76**	**0.28**	**0.30**	**39.00**
	T-FSMG	0.07	0.58	0.74	15.49	0.28	0.44	0.71	17.29
	T-ASPL	0.07	0.57	0.72	15.36	0.39	0.44	0.70	20.06
CelebA-HQ	No Defense	0.10	0.68	0.72	17.06	0.26	0.44	0.72	7.30
	FSMG	**0.34**	**0.48**	0.56	36.13	**0.35**	**0.36**	0.66	33.60
	ASPL	0.31	0.50	**0.55**	**38.57**	0.34	0.39	**0.63**	**34.89**
	T-FSMG	0.06	0.64	0.73	25.75	0.24	0.45	0.73	8.04
	T-ASPL	0.06	0.64	0.73	20.58	0.26	0.46	0.72	5.36

Table 1: Comparing the defense performance of the proposed methods in a convenient setting on different datasets.

Anti-Dreambooth | Experiments

- **Adverse Setting**
  - Model Mismatching
    - Example: Transferring adversarial noise trained on SD v1.4 to defend DreamBooth models trained from v2.1 and v2.0.
    - Ensemble approach (E-ASPL) further improves defense.
  - Term Mismatching
    - Example: Changing from default term ("sks") to another ("t@t").
    - Term mismatch has a moderate effect; key scores like ISM remain good.
  - Prompt Mismatching
    - Malicious user uses a different DreamBooth training prompt.
    - ASPL provides low ISM scores, indicating effectiveness.

	Train	Test	"a photo of *sks* person"				"a dslr portrait of *sks* person"			
			FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑
Model mismatch	v1.4	v2.1	0.62	0.31	**0.28**	36.00	0.70	0.31	0.35	38.39
	v1.4	v2.0	0.70	0.27	0.23	36.83	0.61	0.26	0.31	37.28
Ensemble	v1.4, 1.5, 2.1	v2.0	**0.79**	**0.24**	**0.18**	**37.96**	**0.71**	**0.23**	**0.23**	**38.99**
	v1.4, 1.5, 2.1	v2.1	**0.70**	**0.27**	**0.28**	**36.71**	**0.75**	**0.29**	**0.33**	**39.23**

Term/ Prompt mismatch	DreamBooth prompt	"a photo of S_* person"				"a dslr portrait of S_* person"			
		FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑
	"*sks*" → "*t@t*"	0.34	0.30	0.48	36.67	0.34	0.28	0.52	28.17
	"a dslr portrait of *sks* person"	0.07	0.15	0.69	17.34	0.49	0.37	0.36	38.42

Anti-Dreambooth | Experiments

- **Adverse Setting**
  - Image Preprocessing
    - Robustness evaluation under common image
    - Gaussian blur or JPEG compression slightly weakens defense.
    - Defense maintains reasonable robustness against these techniques.
  - Real-world Test
    - ASPL successfully disrupts personalized images generated by Astria, a black-box commercial service (see Appendix B).

	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑
ASPL	**0.63**	**0.33**	**0.31**	**36.42**
Gaussian Blur K=3	**0.48**	**0.42**	**0.39**	**42.05**
Gaussian Blur K=5	0.19	0.51	0.62	**42.46**
Gaussian Blur K=7	0.10	0.56	0.68	**43.72**
Gaussian Blur K=9	0.07	0.59	0.71	**40.67**
JPEG Comp. Q=10	0.09	0.58	0.71	**43.93**
JPEG Comp. Q=30	0.08	0.59	0.73	**32.56**
JPEG Comp. Q=50	0.11	0.56	0.70	**30.29**
JPEG Comp. Q=70	0.19	0.49	0.56	**37.04**
No def., no preproc.	0.07	0.63	0.73	15.61

Anti-Dreambooth | Experiments

- **Uncontrolled Setting**
  - Considers cases of combining clean and perturbed images for training
  - Defense effectiveness decreases with more clean images

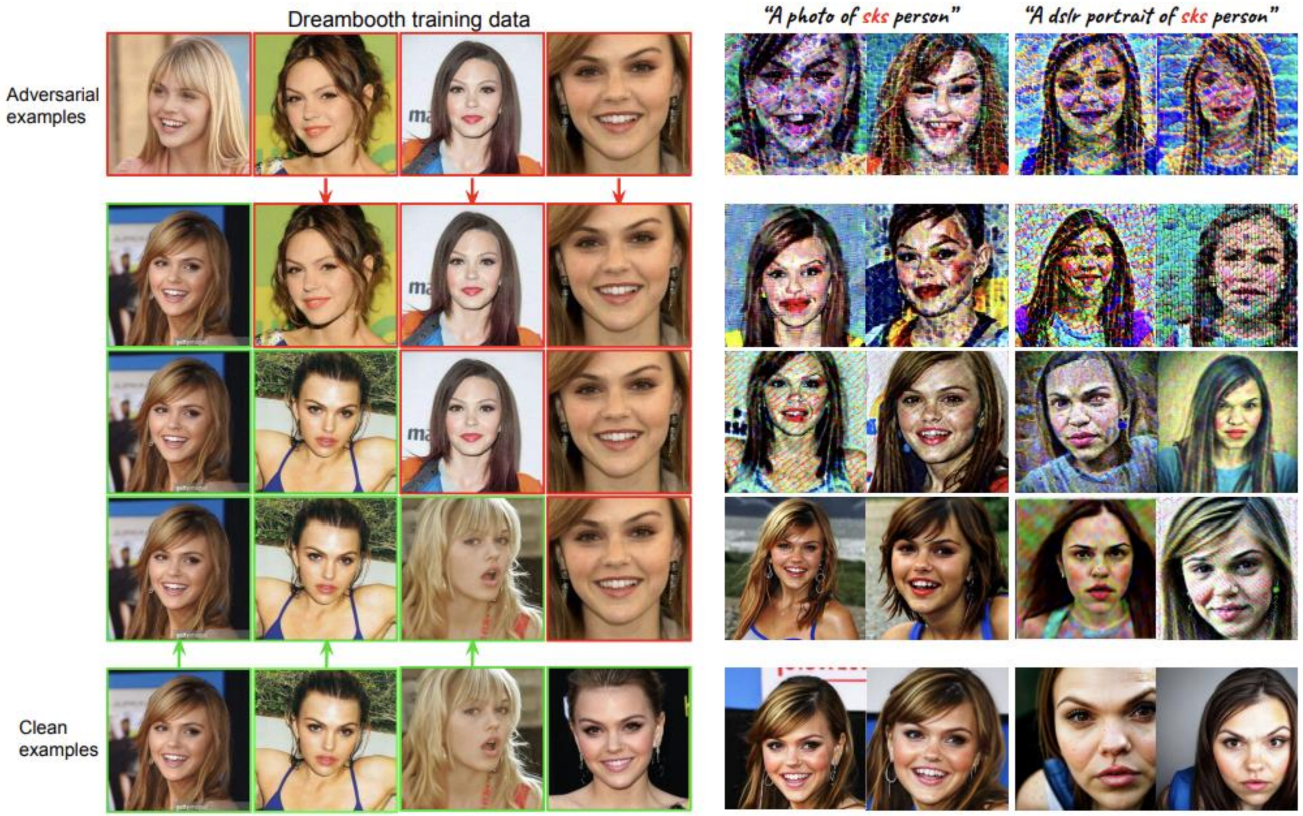


Figure 13: Qualitivative results of ASPL in uncontrolled setting on VGGFace2. We denote the perturbed examples and the leaked clean examples in red and green, respectively.

Perturbed	Clean	"a photo of *sks* person"				"a dslr portrait of *sks* person"			
		FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑
4	0	**0.63**	**0.33**	**0.31**	**36.42**	**0.76**	**0.28**	**0.30**	**39.00**
3	1	0.50	0.43	0.41	35.53	0.52	0.35	0.51	34.01
2	2	0.29	0.53	0.61	28.99	0.40	0.37	0.62	26.13
1	3	0.08	0.61	0.73	18.92	0.27	0.45	0.70	15.55
0	4	0.07	0.63	0.73	15.61	0.21	0.48	0.71	9.64

Anti-Dreambooth | Ablation Study

- **Ablation Study**
  1. Text-to-Image Generator Version:
     - Versions Tested: Stable Diffusion (SD) v1.4, v1.5, and v2.1.
     - Results (Table 2): ASPL shows consistent defense effectiveness across different versions.
  2. Noise Budget (η):
     - Impact on ASPL Attack Using SD v2.1:
       - Larger noise budget improves defense performance at the cost of perturbation's stealthiness (Table 3).

Version	Defense?	"a photo of *sks* person"				"a dslr portrait of *sks* person"			
		FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑
v1.4	✗	0.05	0.46	0.65	21.06	0.08	0.43	0.64	10.05
	✓	**0.80**	**0.18**	**0.12**	**26.76**	**0.17**	**0.28**	**0.55**	**13.07**
v1.5	✗	0.07	0.49	0.65	18.53	0.07	0.45	0.64	10.57
	✓	**0.71**	**0.20**	**0.20**	**22.98**	**0.11**	**0.26**	**0.57**	**16.10**

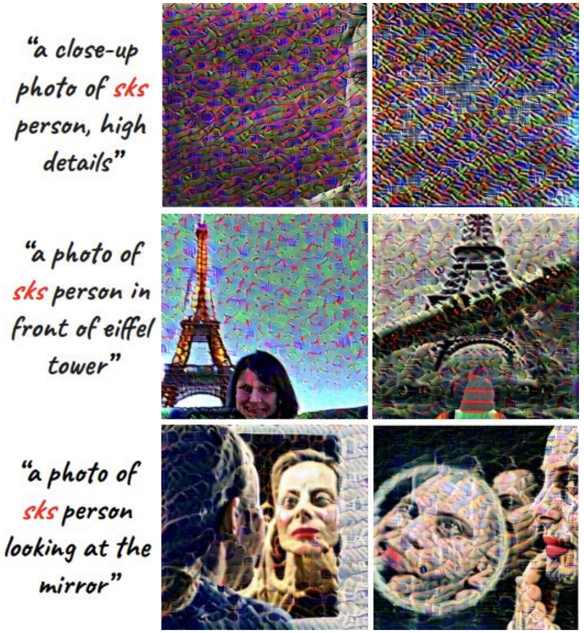
Table 2: Defense performance of ASPL with different generator versions on VGGFace2 in a convenient setting.

η	Quality		"a photo of *sks* person"				"a dslr portrait of *sks* person"			
	PSNR↑	LPIPS↓	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑
0	-	-	0.07	0.63	0.73	15.61	0.21	0.48	0.71	9.64
0.01	48.74	0.01	0.08	0.58	0.72	33.03	0.28	0.45	0.72	17.14
0.03	38.42	0.12	0.44	0.38	0.38	36.45	0.55	0.32	0.43	37.86
0.05*	34.56	0.21	0.63	0.33	0.31	36.42	0.76	0.28	0.30	39.00
0.10	28.77	0.40	0.76	0.21	0.22	37.33	0.86	0.23	0.26	40.92
0.15	25.97	0.50	**0.80**	**0.15**	**0.15**	**37.07**	**0.91**	**0.17**	**0.14**	**41.18**

Table 3: Quality of protected images and defense performance of ASPL with different noise budgets on VGGFace2 in a convenient setting. "*" is default.

Anti-Dreambooth | Ablation Study

- **Ablation Study**
  3. Inference Text Prompt:
     - Tested Against ASPL-Disturbed DreamBooth Models:
       - ASPL well-disturbs images generated with unseen text prompts ("a DSLR portrait of sks person").
       - Similar results obtained with different inference text prompts (Fig. 3b).
  4. Comparison with Other Techniques:
     - Tested Against Textual Inversion and DreamBooth with LoRA:
       - ASPL successfully defends against both methods, demonstrating effectiveness against other personalization techniques (Table 7).



	Def.?	FDFR↑	ISM↓	SER-FQA↓	BRISQUE↑
TI	✗	0.06	0.50	0.67	7.79
TI	✓	**0.43**	**0.12**	0.59	**36.79**
LoRA	✗	0.06	0.52	0.69	17.25
LoRA	✓	**0.64**	**0.23**	**0.27**	**42.07**

Table 7: ASPL's performance against Textual Inversion and LoRA DreamBooth, the prompt is "*A photo of sks person*".

Anti-Dreambooth | Disucussion

- **Limitations**
  - Generalization to unseen prompts and models
  - Optimization complexity - Excessive computational cost, especially in GPU memory
  - Targeted- method failure cases

- **Discussion**
  - Transferability of adversarial attacks
  - Tradeoff between sophistication and generality
  - A battle of sword and shield

Thank you!