Trustworthy Machine Learning

Copyright and Generative AI/ML

Sangdon Park

POSTECH

Powerful Generative AI/ML

- Super-resolution

Image Credit: https://arxiv.org/pdf/2112.10752.pdf

Powerful Generative AI/ML

- Removal and Inpainting

Image Credit: https://arxiv.org/pdf/2112.10752.pdf

Powerful Generative AI/ML

- Text-to-Image Synthesis

'A painting of a squirrel eating a burger'

Image Credit: https://arxiv.org/pdf/2112.10752.pdf

Stable Diffusion (=Latent Diffusion Models)

'A painting of the last supper by Picasso.'

Image Credit: https://arxiv.org/pdf/2112.10752.pdf

Stable Diffusion v.s. DALL-E

Girl with a Pearl Earring by Vermeer

Prompt: “A painting by Vermeer of an Irish wolfhound enjoying a pint of a traditional pub”

DALL-E

Stable Diffusion

Image Credit: https://zapier.com/blog/stable-diffusion-vs-dalle/

Density Estimation: GAN

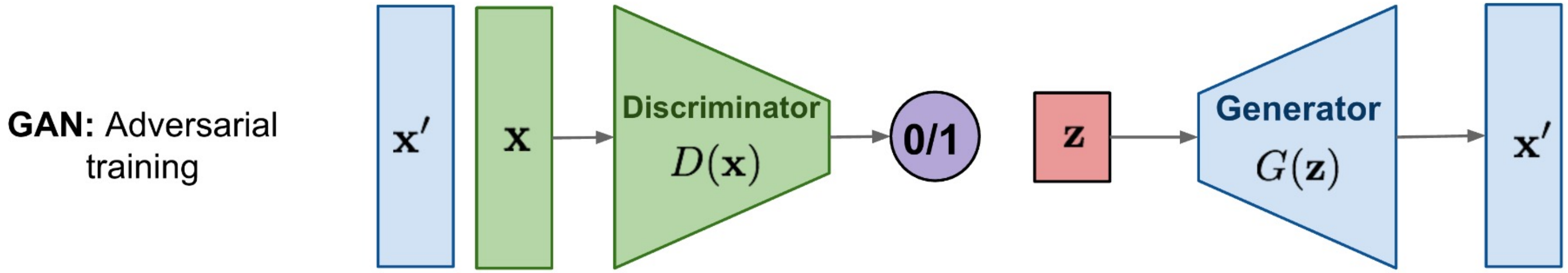


Image Credit: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Density Estimation: VAE

VAE: maximize
variational lower bound

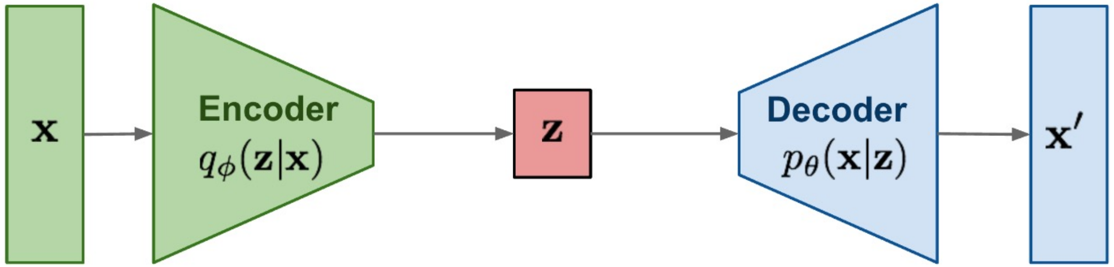


Image Credit: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Density Estimation: Flow-based Models

Flow-based models:
Invertible transform of distributions

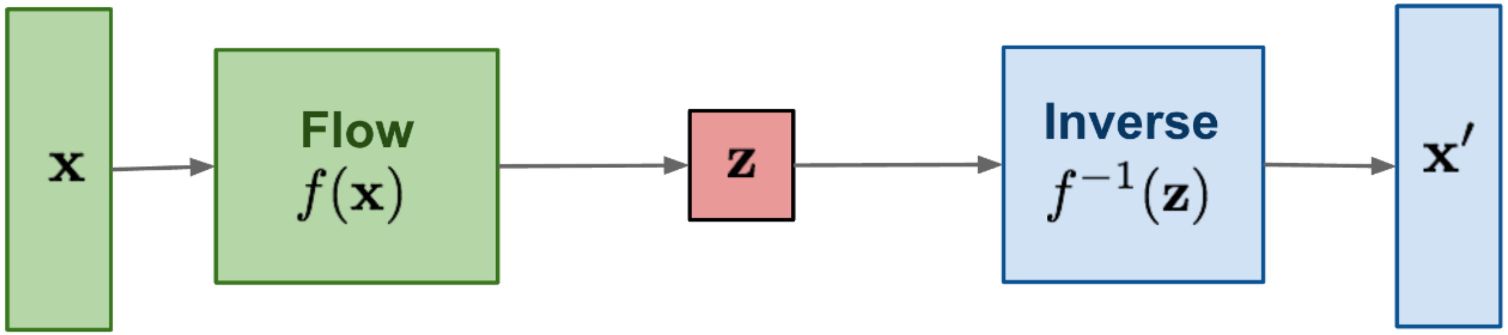


Image Credit: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Density Estimation: Diffusion Models

Diffusion models:
Gradually add Gaussian noise and then reverse

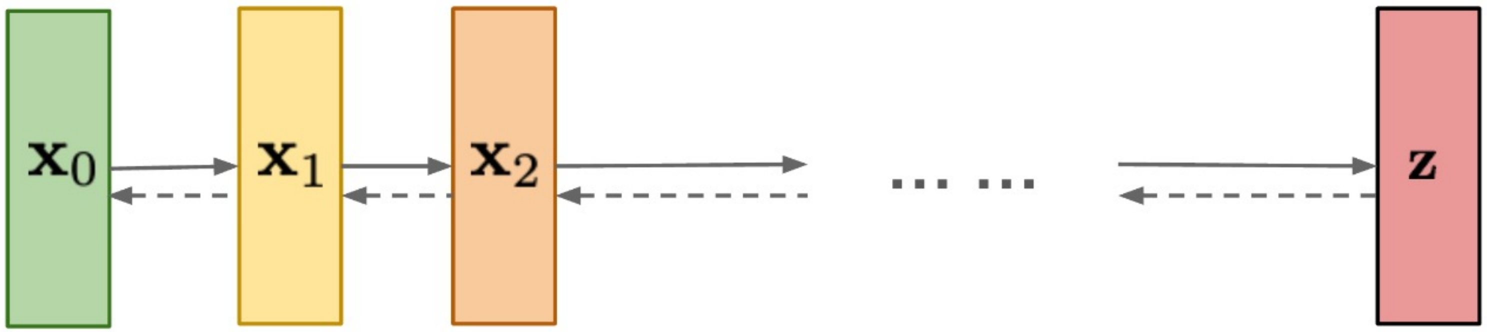


Image Credit: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Diffusion Models: Diffusion Process

- Predict noise via self-supervised learning!

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\boxed{\epsilon} - \epsilon_\theta(x_t,t)\|_2^2\right]$$

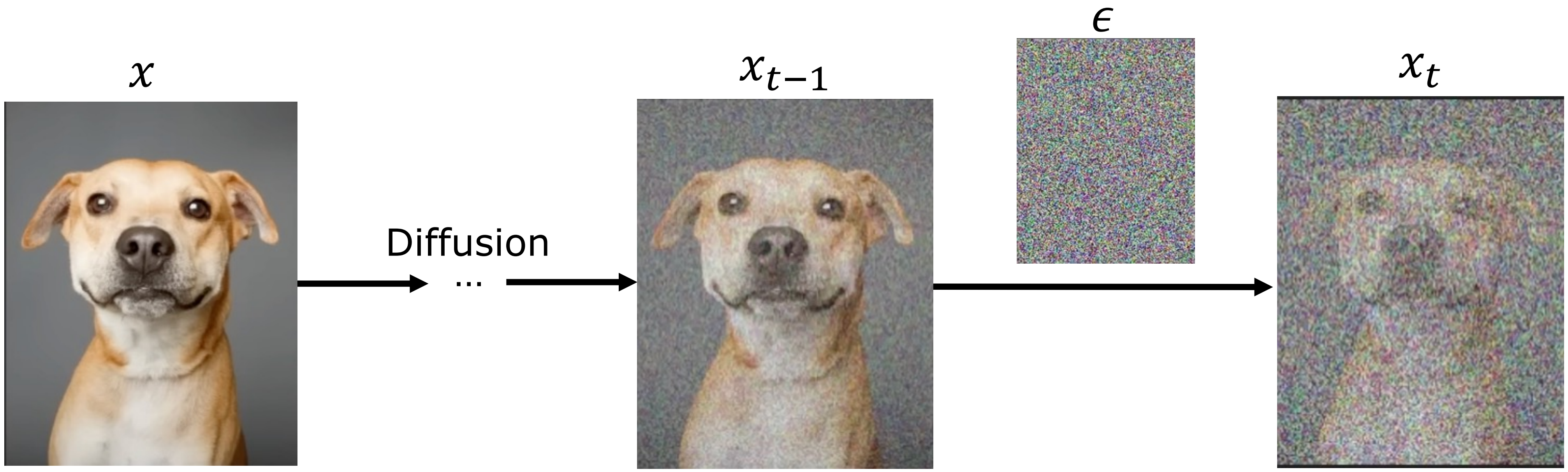
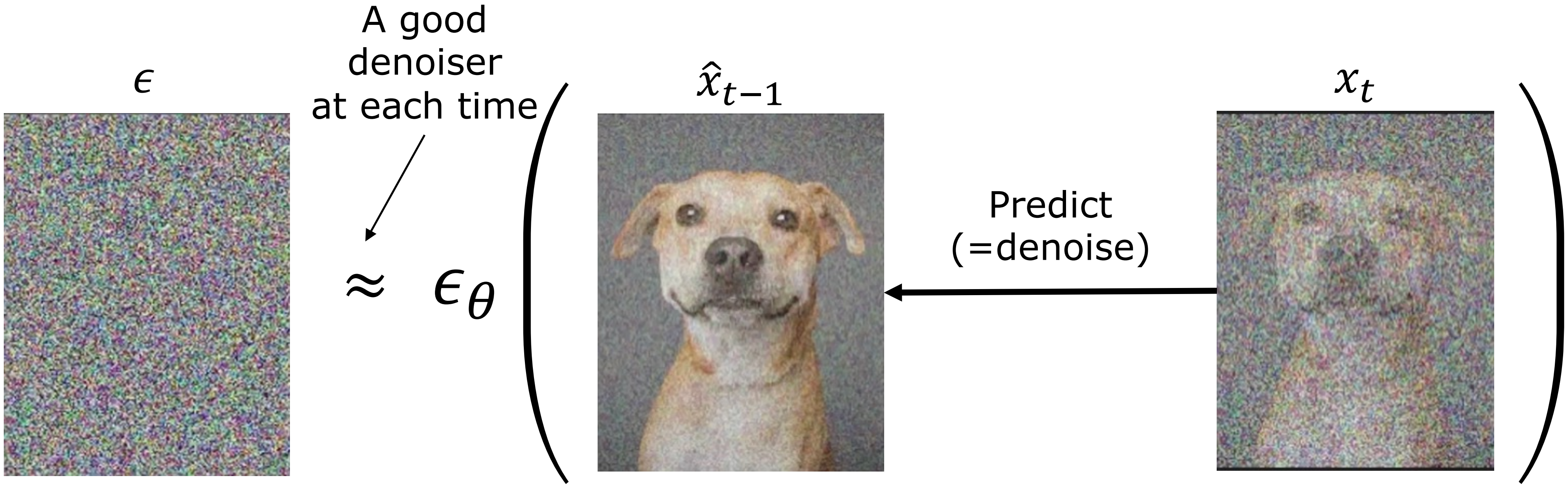


Image Credit: https://www.youtube.com/watch?v=J87hffSMB60&ab_channel=AICoffeeBreakwithLetitia

Diffusion Models: Reverse Diffusion Process

- Predict noise via self-supervised learning!

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]$$



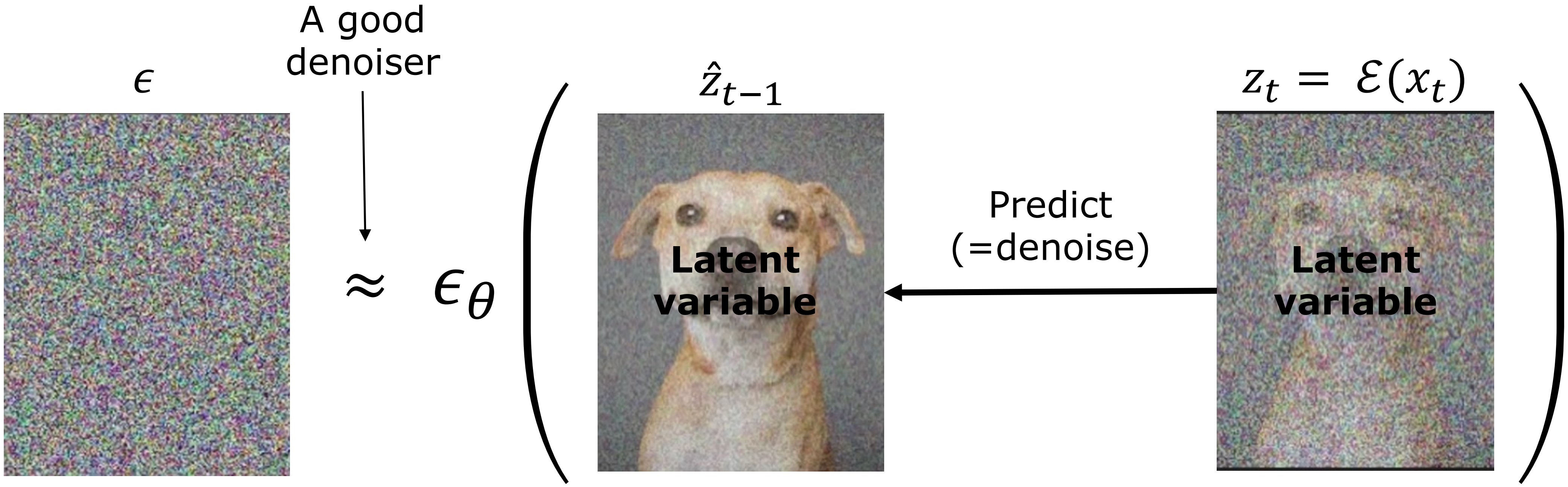
X operations on image space → computationally expensive

Image Credit: https://www.youtube.com/watch?v=J87hffSMB60&ab_channel=AICoffeeBreakwithLetitia

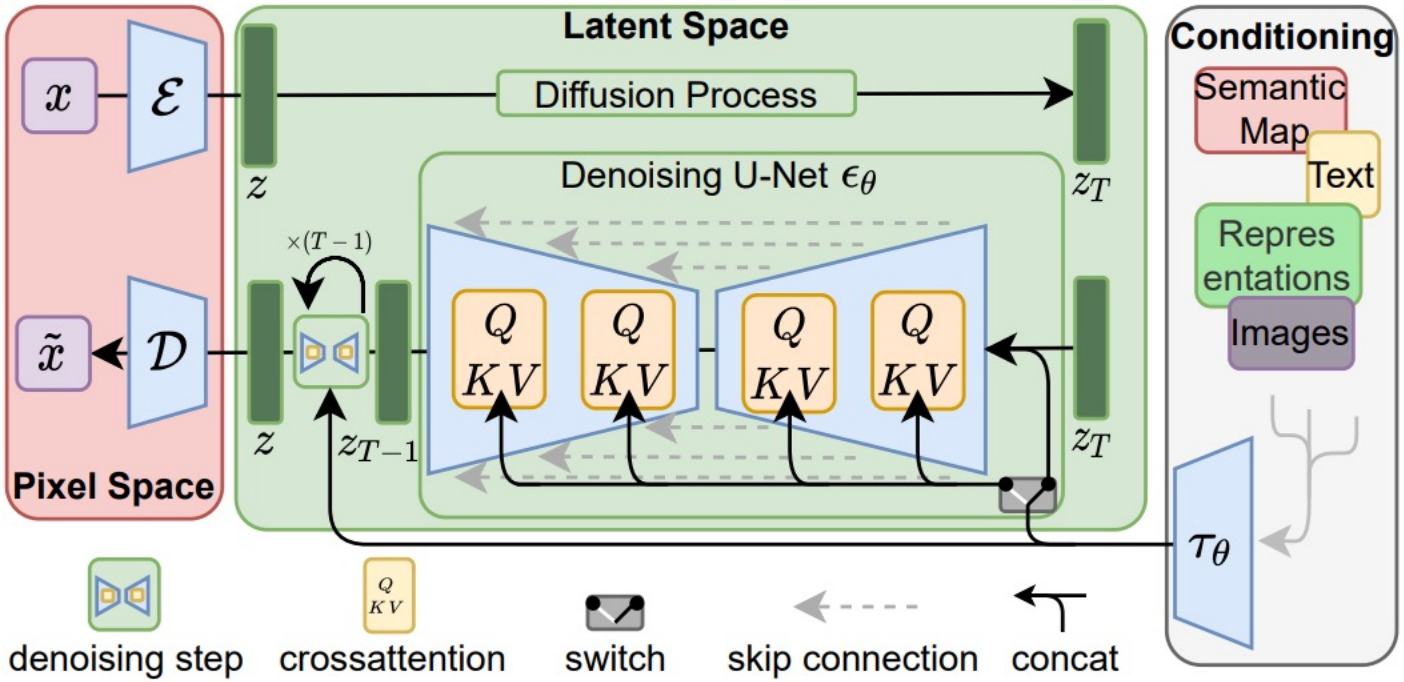
"Latent" Diffusion Models

- Predict noise in latent space

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right]$$



Latent Diffusion Models: Full Picture



What's the main difference from VAE?

Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models (Security 23)

Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, Ben Y. Zhao
Department of Computer Science, University of Chicago
{shawnshan, jennacryan, ewillson, htzheng, ranahanocka, ravenben}@cs.uchicago.edu

Abstract

Recent text-to-image diffusion models such as MidJourney and Stable Diffusion threaten to displace many in the professional artist community. In particular, models can learn to mimic the artistic style of specific artists after "fine-tuning" on samples of their art. In this paper, we describe the design, implementation and evaluation of *Glaze*, a tool that enables artists to apply "style cloaks" to their art before sharing online. These cloaks apply barely perceptible perturbations to images, and when used as training data, mislead generative models that try to mimic a specific artist. In coordination with the professional artist community, we deploy user studies to more than 1000 artists, assessing their views of AI art, as well as the efficacy of our tool, its usability and tolerability of perturbations, and robustness across different scenarios and against adaptive countermeasures. Both surveyed artists and empirical CLIP-based scores show that even at low perturbation levels (p=0.05), *Glaze* is highly successful at disrupting mimicry under normal conditions (>92%) and against adaptive countermeasures (>85%).

Figure 1. Sample AI-generated art pieces from the Midjourney community showcase [53, 69].

many have taken the open sourced StableDiffusion model, and "fine-tuned" it on additional samples from specific artists, allowing them to generate AI art that *mimics* the specific artistic styles of that artist [32]. In fact, entire platforms have sprung up where home users are posting and sharing their own customized diffusion models that specialize on mimicking specific artists, likeness of celebrities, and NSFW themes [14].

Real-work Mimicry Incidents

Original artwork
by Hollie Mengert

Mimicked artwork
in Hollie's style

Threat Model

Artist

- Share and promote their artwork online.
- Don’t want to allow replicate their art style.
- Access to a public feature extractor

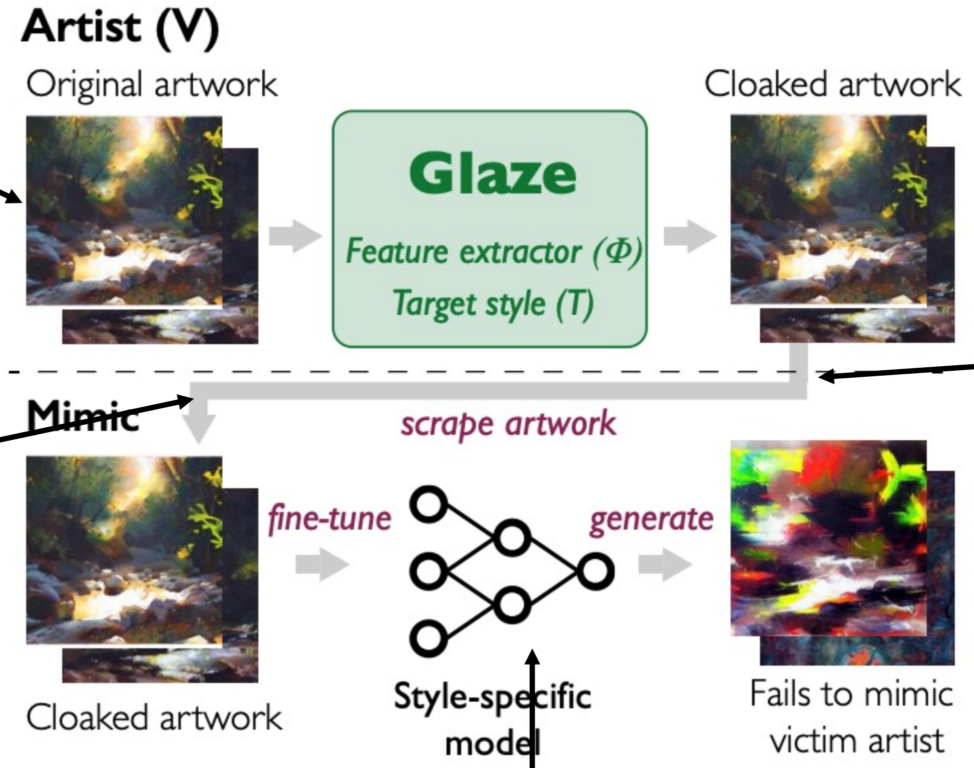
Mimics

- Copy the victim’s style
- Access to victim’s art pieces
- Access to a text-to-image model

Glaze: Overview

Can you guess the implmentation of Glaze?

1. Create artwork



2. Add imperceptible perturbations

3. Release artwork

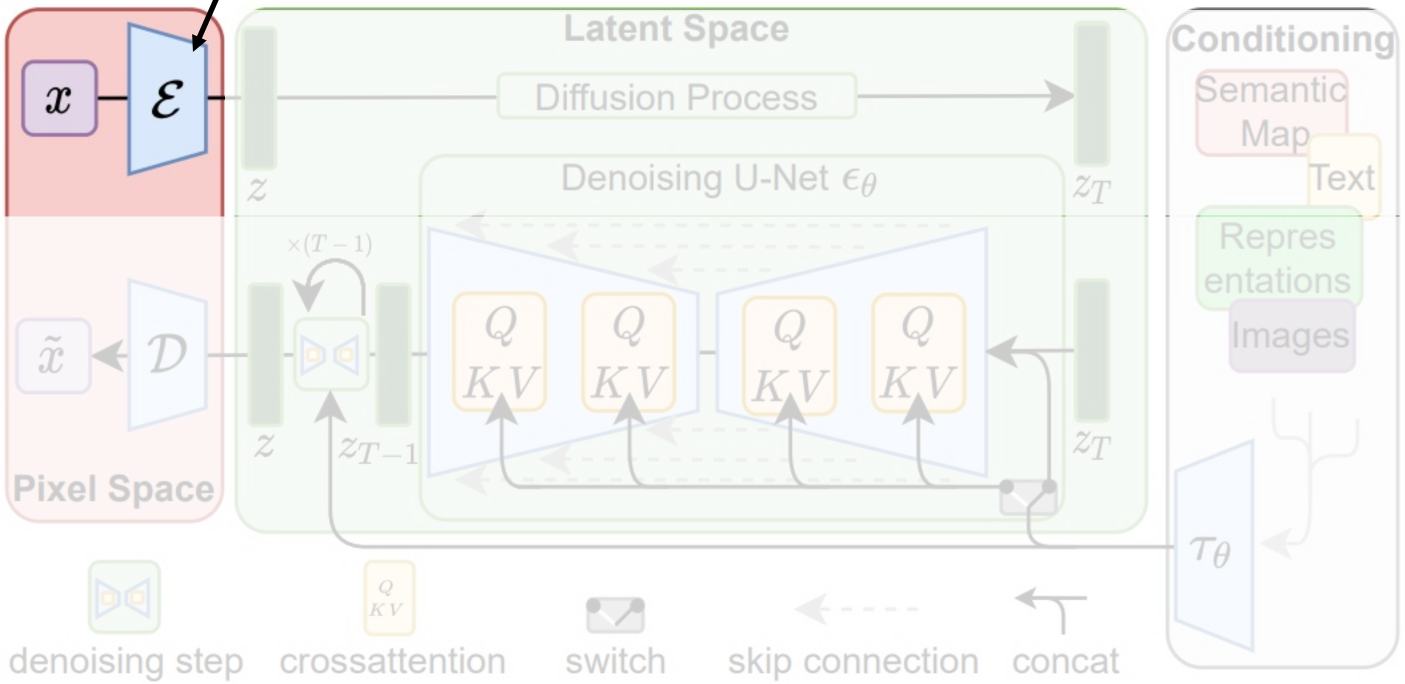
4. Collect artwork for copying the style

5. fine-tune a public text-to-image model to learn the target style

6. Fails to mimic the target style.

Glaze: Requirements

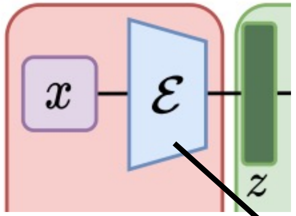
1. Encoder



2. Target style

3. Style-transfer

Glaze: Find Small Perturbations



$$\min_{\delta}. \quad dist(\mathcal{E}(x+\delta), \mathcal{E}(\Omega(x,T)))$$
$$\text{subj. to} \quad b(\delta) \le p$$

$\Omega(x, T)$

My work with a "Van Gogh" style

$x + \delta$

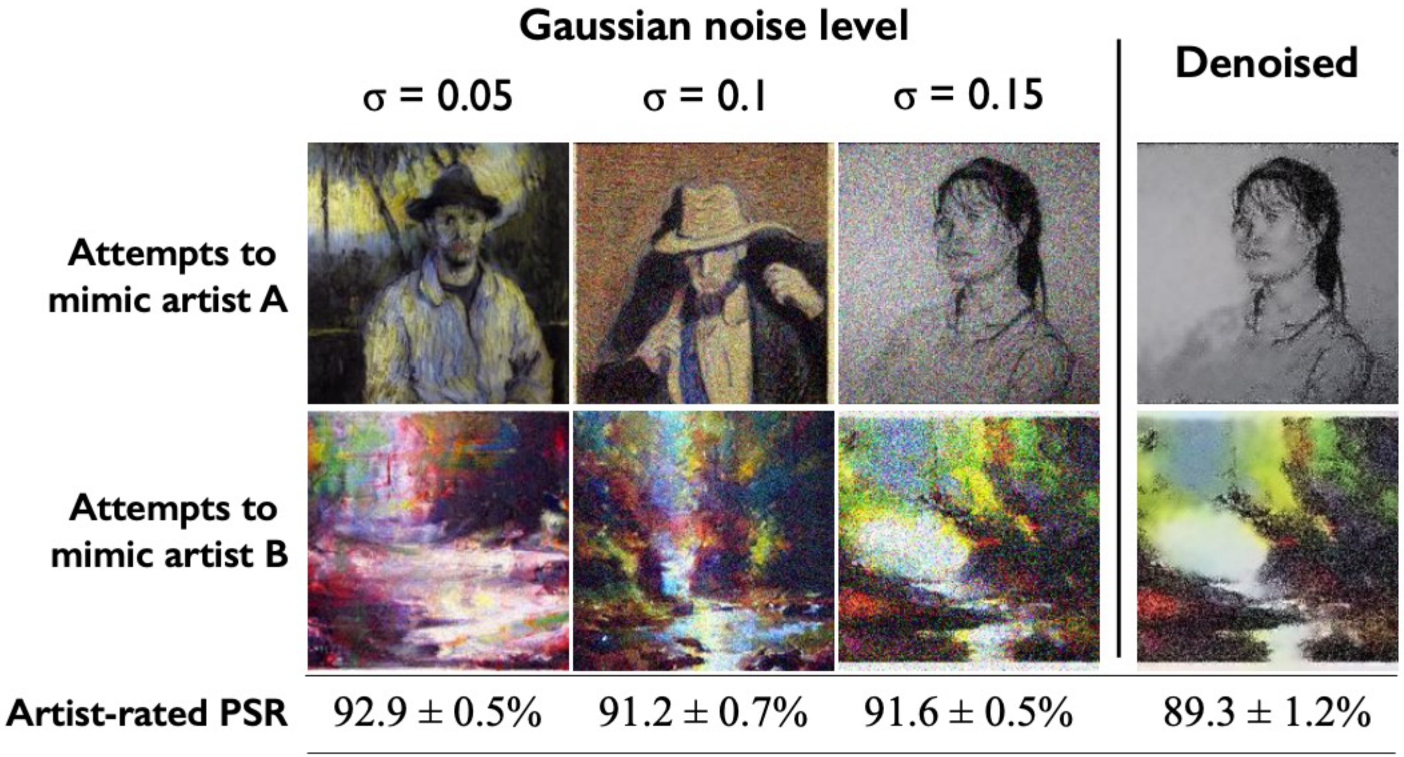
A Cloaked image

z

Results


Original artwork
Mimicked art when Glaze not used
Glaze target style
Mimicked art when Glaze is used
Artist A (Karla Ortiz)
Artist B (Nathan Fowkes)
Artist C (Claude Monet)
Oil painting by Van Gogh
Abstract expressionism by Norman Bluhm
Cubism by Picasso
p = 0.05
p = 0.1
Glaze perturbation size

Let’s Bypass Glaze: Add Guassian Noise



Glaze still works!

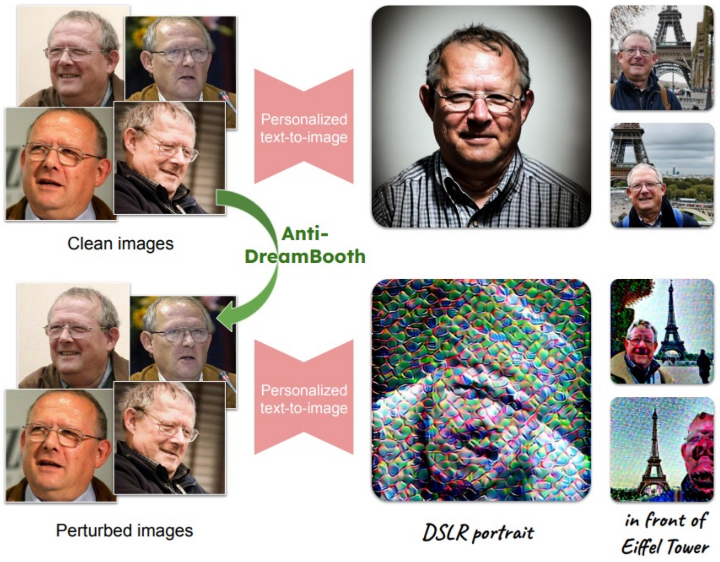
Anti-DreamBooth: Protecting users from personalized text-to-image synthesis (ICCV23)

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

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

Abstract

Text-to-image diffusion models are nothing but a revolution, allowing anyone, even without design skills, to create realistic images from simple text inputs. With powerful personalization tools like DreamBooth, they can generate images of a specific person just by learning from his/her few reference images. However, when misused, such a powerful and convenient tool can produce fake news or disturbing content targeting any individual victim, posing a severe negative social impact. In this paper, we explore a defense system called Anti-DreamBooth against such malicious use of DreamBooth. The system aims to add subtle noise perturbation to each user's image before publishing in order to disrupt the generation



Motivation: Deepfake (≅DreamBooth)

Original Images

DreamBooth →

Fake Images

DSLR portrait

in front of Eiffel Tower

Goal: Anti-DreamBooth

Original Images

Fake-failed Images

DreamBooth (CVPR23)

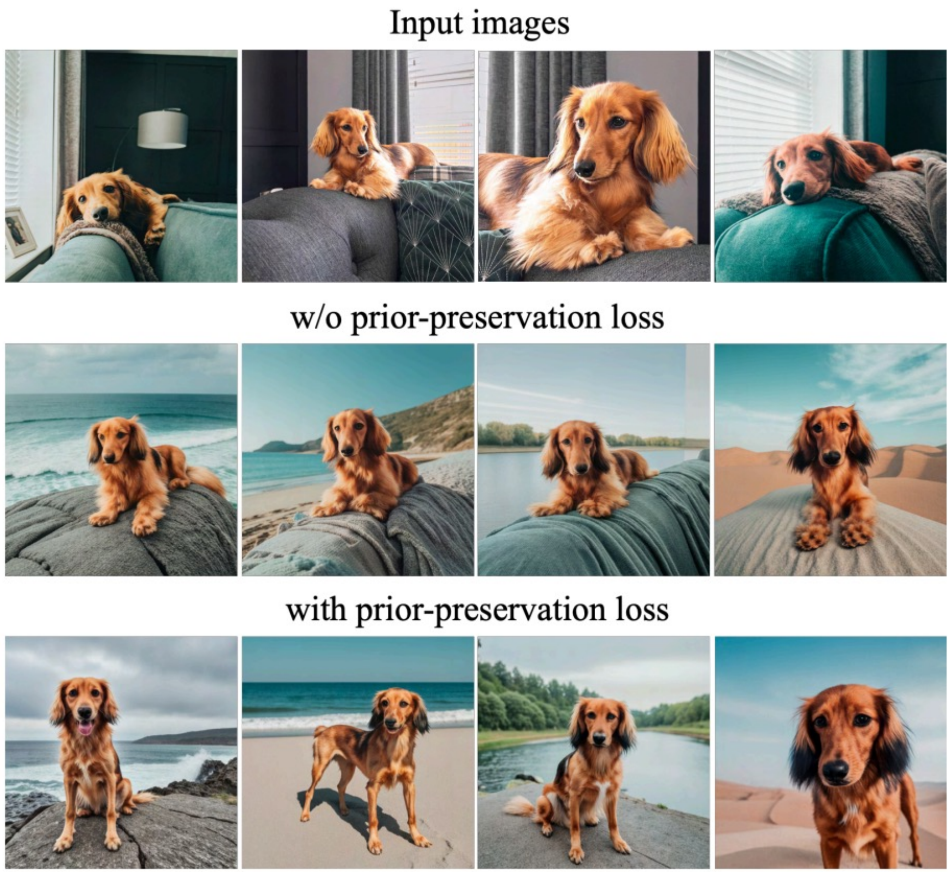
Reconstruction Loss

$$\mathcal{L}_{DB}(\theta) = \mathbb{E}_{x_0,\epsilon,\epsilon',t,t'}\left\{\|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2 + \lambda\left\|\epsilon' - \epsilon_\theta(x'_t, t', c_p)\right\|_2^2\right\}$$

DreamBooth (CVPR23)

Prior-preserving Loss

$$\mathcal{L}_{DB}(\theta) = \mathbb{E}_{x_0,\epsilon,\epsilon',t,t'}\left\{ \|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2 + \lambda \|\epsilon' - \epsilon_\theta(x'_t, t', c_p)\|_2^2 \right\}$$

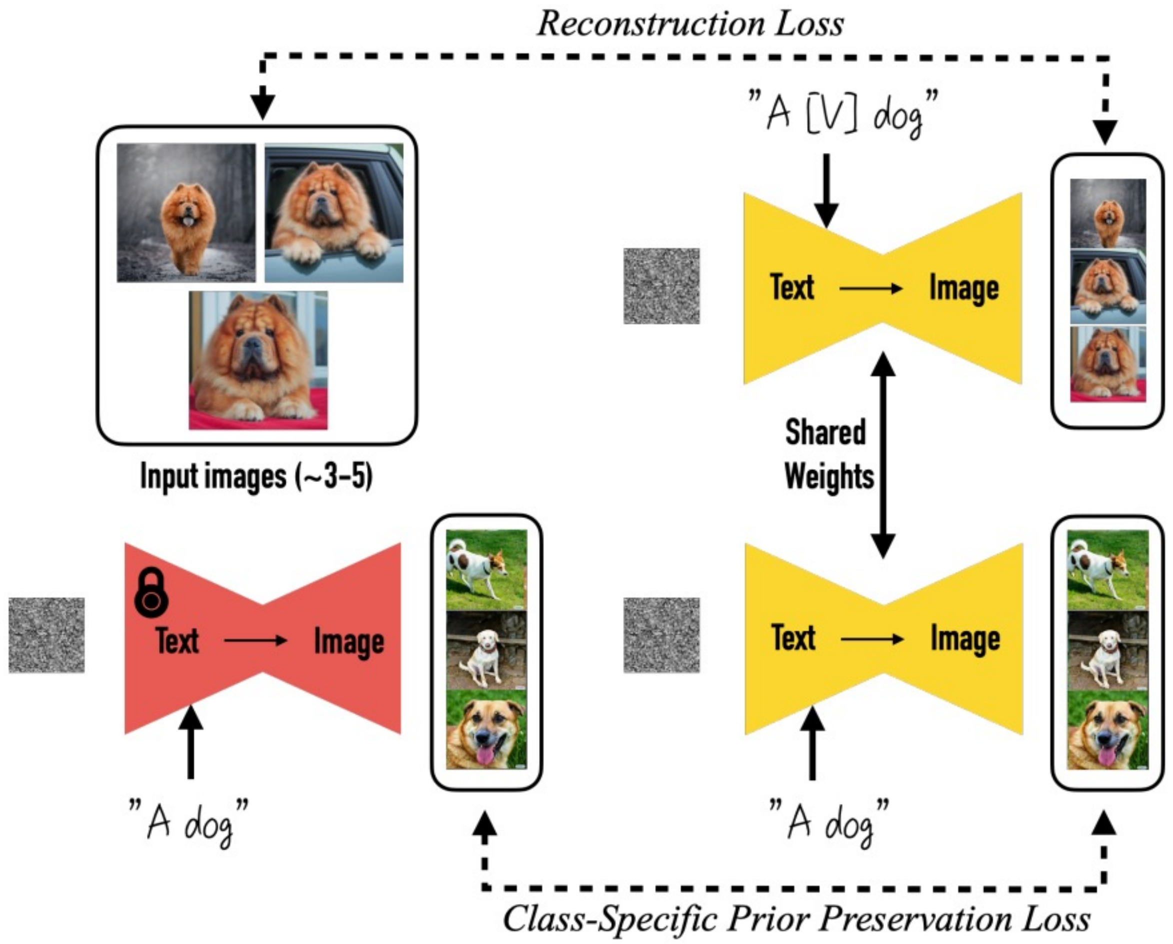


$c_p \longrightarrow$ Generator (original) $\longrightarrow x'$

A [noun]

DreamBooth (CVPR23)

$$\mathcal{L}_{DB}(\theta) = \mathbb{E}_{x_0,\epsilon,\epsilon',t,t'}\left\{\|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2 + \lambda\left\|\epsilon' - \epsilon_\theta(x'_t, t', c_p)\right\|_2^2\right\}$$



Anti-DreamBooth

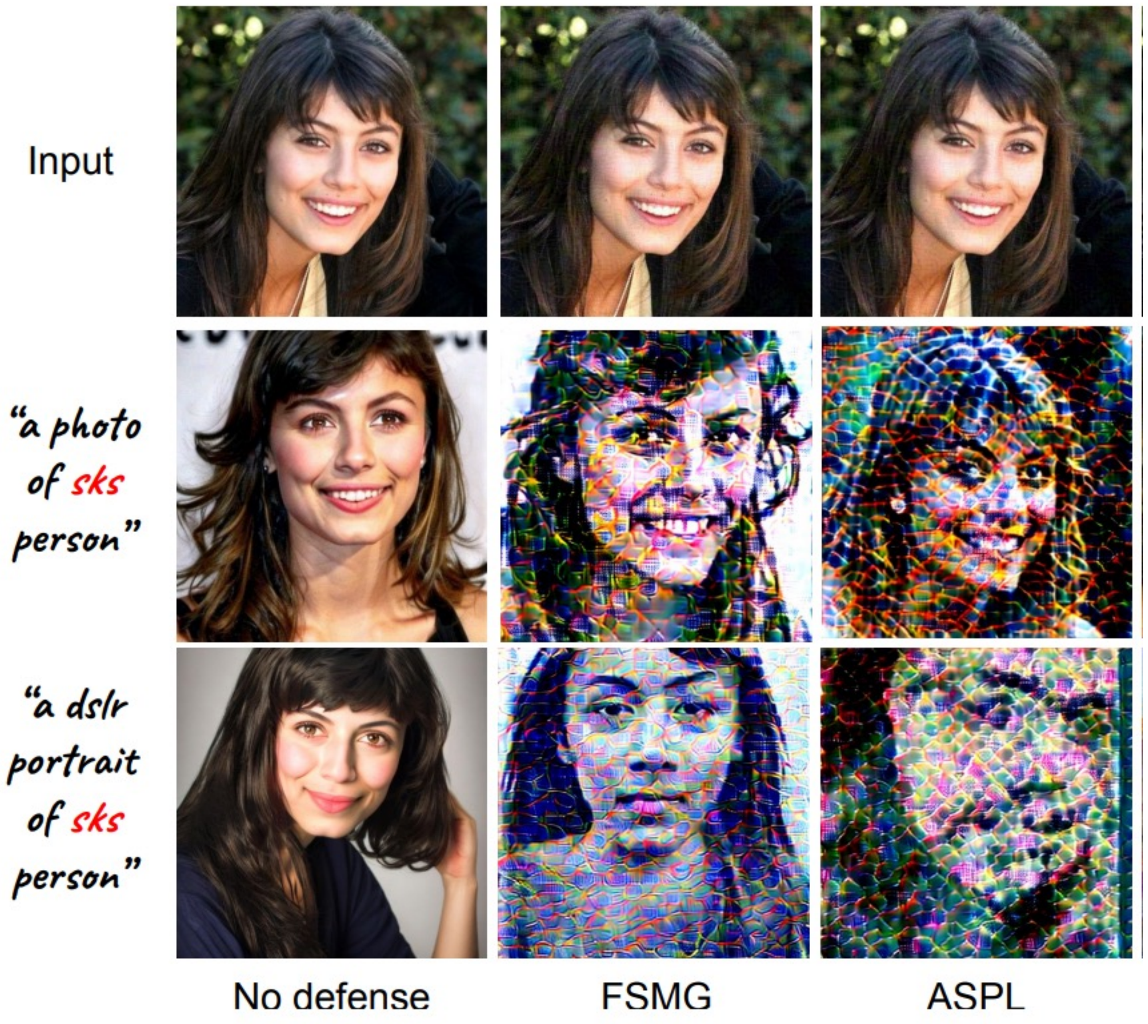
Reconstruction Loss

$$\delta^{*(i)} = \underset{\delta^{(i)}}{\operatorname{argmax}}\, \mathbb{E}_{x^{(i)},\epsilon,t} \left\{ \left\| \epsilon - \epsilon_{\theta^*}\left(x_t^{(i)} + \delta^{(i)}, t, c\right) \right\|_2^2 \right\}$$

$$\text{subj. to} \quad \theta^* = \min_{\theta} \sum_i \mathcal{L}_{DB}\left(\theta, x^{(i)} + \delta^{(i)}\right)$$

$$\left\| \delta^{(i)} \right\|_p \leq \eta$$

Results



See the paper for targeted attaks.

Discussion

- Glaze v.s. Anti-DreamBooth