2025 겨울 세미나

Protecting Against Editing by Diffusion Model

Pusan National University
Computer Vision & Signal Processing Lab, Dept. of Electronic Engineering

Presented By
정성훈

CVSP Lab
Computer Vision & Signal Processing

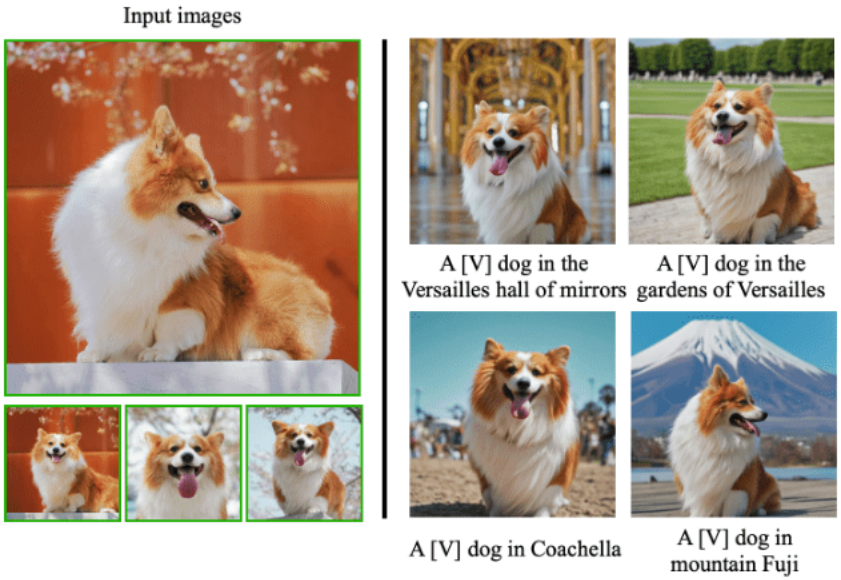
Outline

- Motivation
- Background
  - Diffusion Models
  - Latent Diffusion Models (Stable Diffusion)
  - Adversarial Attack
- Anti-dreambooth : protecting users from personalized text-to-image synthesis (ICCV 2023)
- DiffusionGuard: A Robust Defense Against Malicious Diffusion-based Image Editing (ICLR 2025)

Motivation

[1] Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023.

- 논문 및 주제 선정 이유
  - 생성형 AI의 발전
    - 최근 생성형 AI의 성능이 고도화됨에 따라 사용자들은 이미지 및 비디오를 쉽고 빠르게 생성할 수 있음
    - 기존의 데이터들을 활용하여 개인화된 콘텐츠 또한 생성 가능
      - 개인화: 사용자의 사진, 영상, 음성 등 개인 데이터를 활용하여 특정 요구에 부합하는 이미지나 비디오 제작



기존 이미지를 활용하여 개인화된 이미지를 편집할 수 있는 DreamBooth[1]

Motivation

[2] https://www.youtube.com/watch?v=yea-flCXHSg
[3] https://www.youtube.com/watch?v=PdojnAQE92U

- 논문 및 주제 선정 이유
  - 생성형 AI 기반 콘텐츠 제작에 대한 장단점
    - 생성형 AI는 창의적 콘텐츠 생성과 생산성 향상 등 다양한 장점을 제공
    - 반면 고성능의 생성형 AI에 대한 오남용으로 인해 심각한 사회적 부작용을 초래할 수 있음
      - 사기, 부적절한 콘텐츠, 가짜뉴스 등
    - 생성형 AI에 대한 오남용에 대해 탐지하는 것을 넘어 사전에 방지하기 위한 연구가 필요



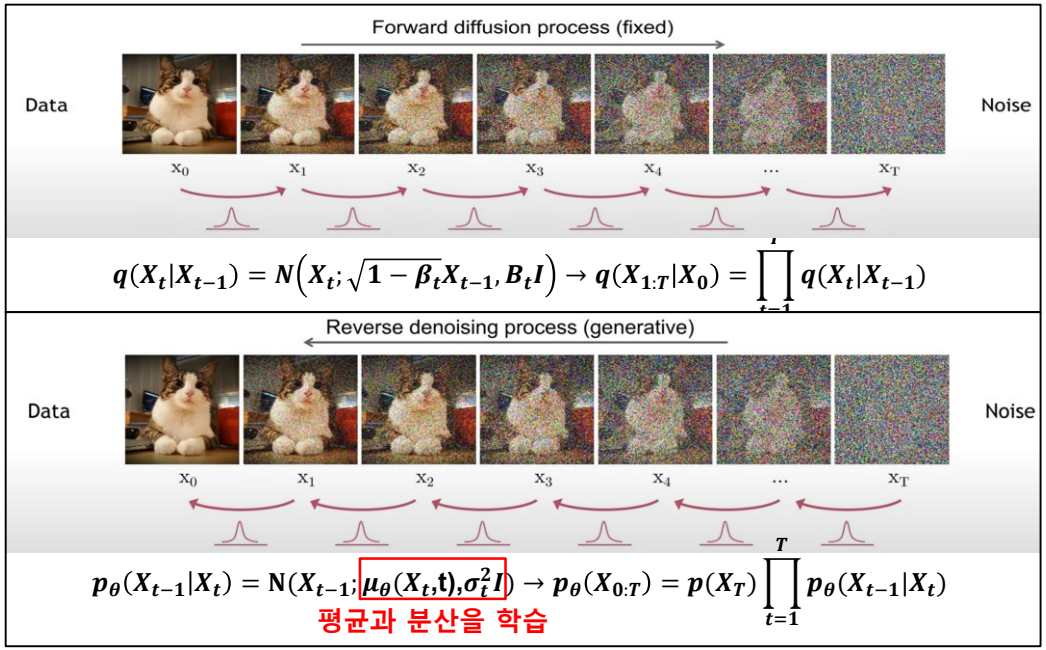


생성형 AI의 오남용으로인한 피해 사례[2,3]

Background

[4] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.

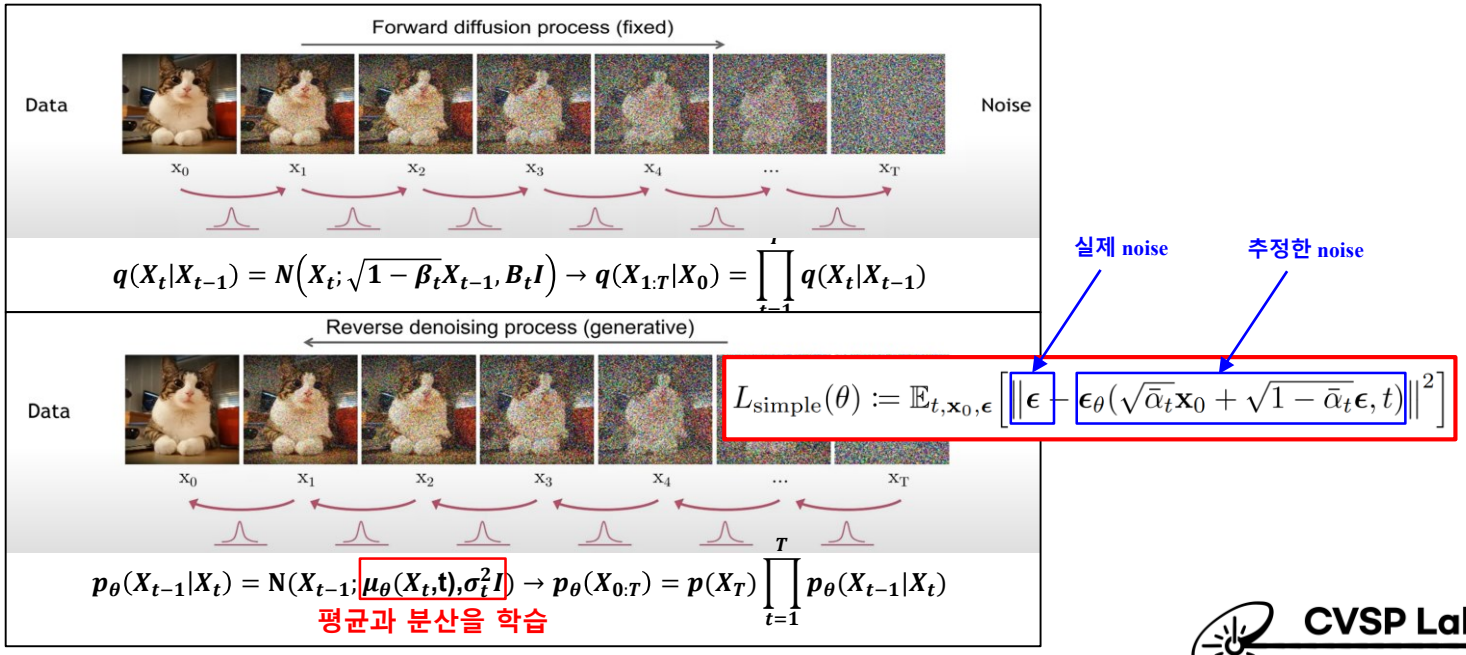
- Diffusion Models[4]
  - 고품질 이미지를 생성할 수 있는 모델
  - Gaussian noise를 더하는 과정과 역으로 추정하여 제거하는 과정을 학습함으로써 이미지 생성할 수 있음.
  - 생성되는 이미지 계속적으로 변경
    - 실제 noise 생성 시 Gaussian distribution에서 항상 다른 값이 샘플링 됨



Background

[4] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.

- Diffusion Models[4]
  - 고품질 이미지를 생성할 수 있는 모델
  - Gaussian noise를 더하는 과정과 역으로 추정하여 제거하는 과정을 학습함으로써 이미지 생성할 수 있음.
  - 생성되는 이미지 계속적으로 변경
    - 실제 noise 생성 시 Gaussian distribution에서 항상 다른 값이 샘플링 됨

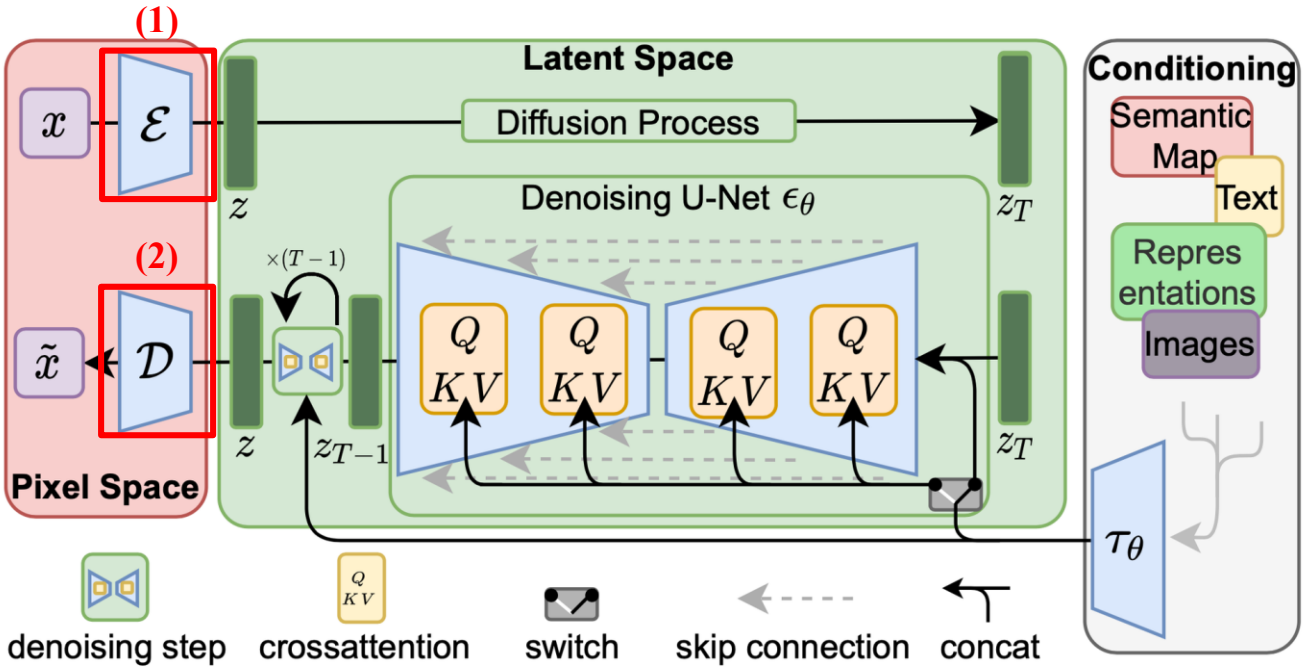


CVSP Lab
Computer Vision & Signal Processing

Background

[5] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

- Latent Diffusion Models[5] (Stable Diffusion)
  - 기존 Pixel 단위 Diffusion에서 AutoEncoder(1)와 Decoder(2)를 통해 Latent 단위로 변환하여 Diffusion을 진행
    - Latent space에서 feature level로 수행하여 pixel에서 보다 더 좋은 scaling property를 보이고 계산이 훨씬 덜 복잡해 이미지 생성에 더 효과적임.
  - Cross-Attention을 통해 추가적인 Conditioning 가능
    - Text, Images, Semantic Map 등

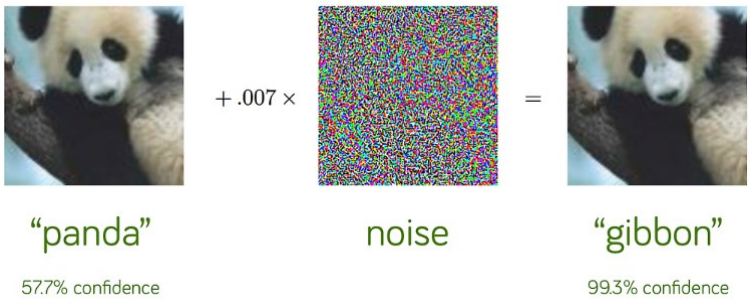


CVSP Lab Computer Vision & Signal Processing

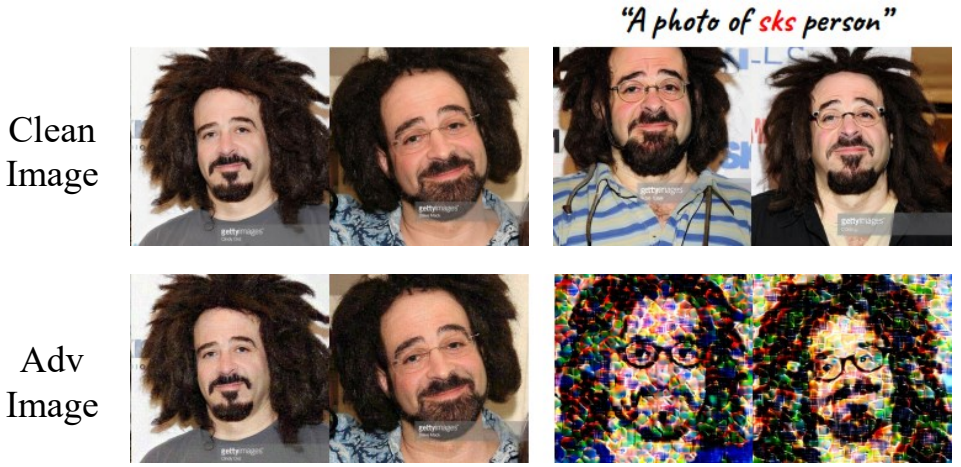
[6] Mądry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." stat 1050.9 (2017).

Background

- Adversarial Attack
  - Input image에 감지할 수 없는 perturbation을 추가하여 모델이 오작동하도록 유도
  - Adversarial Examples
    - 시각적으로 원본 Data(image)와 구분되지 않는 perturbation이 들어간 Data(image), $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\delta}$
  - Projected Gradient Descent (PGD)[6]
    - Iterative optimization process 기반에 주로 사용되는 attack으로 **adversarial example ($\boldsymbol{x}'$)**생성
      - n 번의 step 만큼 attack을 반복하여 정해진 L_∞-norm 내에서 inner maximization을 수행



Adversarial attack in Classification



Adversarial attack in Generation Model

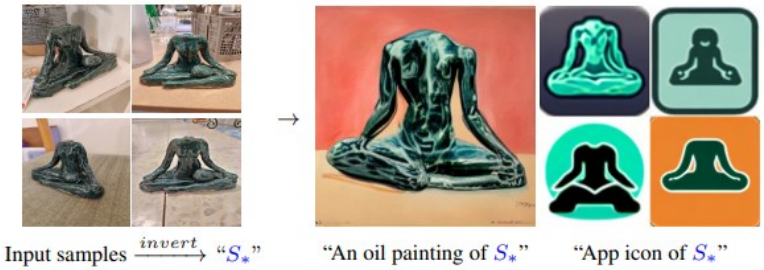
Anti-DreamBooth: Protecting users from personalized text-to-image synthesis

ICCV 2023

Introduction

[7] Gal, Rinon, et al. "An image is worth one word: Personalizing text-to-image generation using textual inversion." arXiv preprint arXiv:2208.01618 (2022).

- Motivation
  - 이미지 생성모델의 매우 유용한 기능 : 개인화
    - Textual inversion[7], Dreambooth[1]
  - Dreambooth
    - Textual inversion은 대상 subject를 나타내는 텍스트 임베딩 입력만 최적화하지만, Dreambooth는 text-image 모델 자체를 fine-tuning함.
    - 생성된 사진과 같은 이미지를 사용해 가짜 뉴스, 딥페이크 등의 악용 가능성 존재



Textual inversion examples



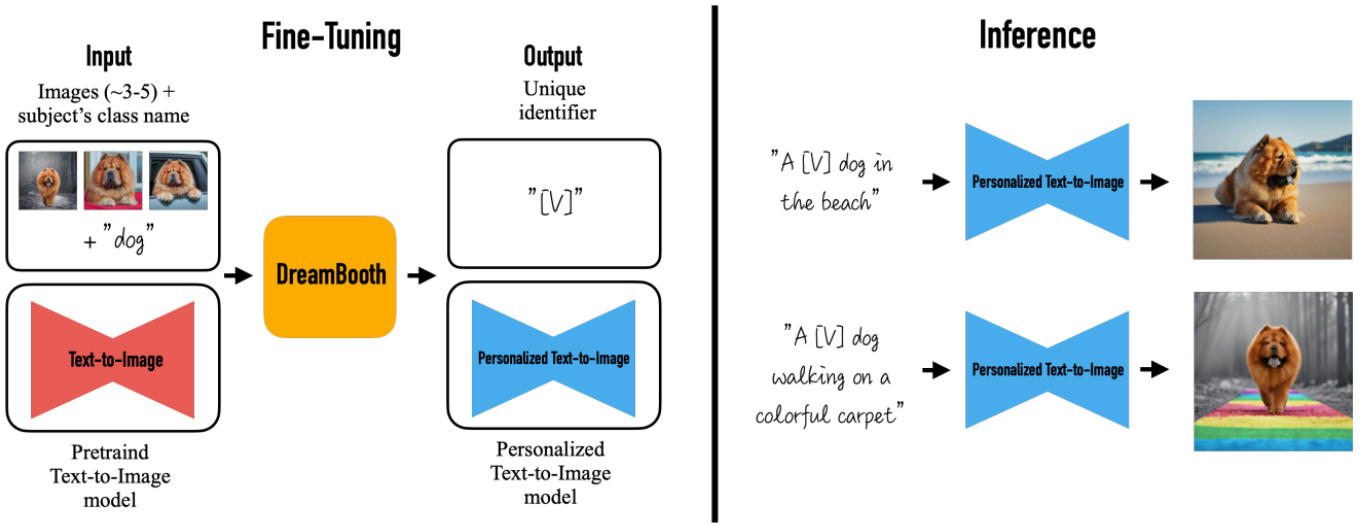
DreamBooth examples

Background

- Dreambooth
  - Personalized한 image에 text-to-image 생성모델을 Fine-tuning하는 방법
    - Fine-tuning한 이미지에 맞는 이미지가 생성됨
  - Prior preservation loss 사용
    - 소수의 input image로 발생할 수 있는 overfitting과 language drift문제 해결
      - Fidelity를 가지면서도 기존 모델의 semantic knowledge 유지

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

Input image　　Prior preservation



Method

- Overview (Anti-DreamBooth)
  - Dreambooth의 learning process를 중점으로 attack하여 adversarial image가 Dreambooth model에 overfit되게 함
    - Clean images reconstruction이 잘 수행되지 않음

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

Input image

Prior preservation

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

$$\delta^{*(i)} = \underset{\delta(i)}{\mathbf{argmax}}\, \mathcal{L}_{cond}\left(\theta^*, x^{(i)}\right), \forall i \in \{1, \dots, N_{db}\},$$

원본이미지와 loss가 커지게하는 δ를 최적화하는 term

$$\text{s.t. } \theta^* = \underset{\theta}{\mathbf{argmin}} \sum_{i=1}^{N_{db}} \mathcal{L}_{db}\left(\theta, x^{(i)} + \delta^{(i)}\right),$$

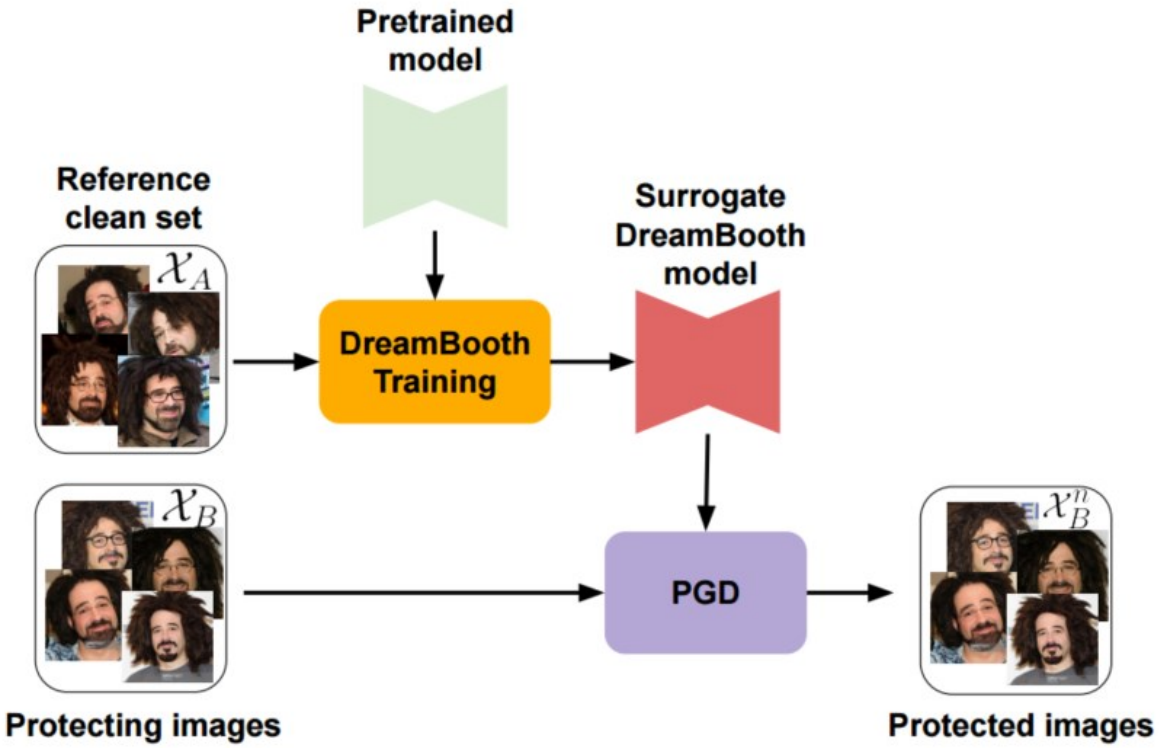
Dreambooth model이 perturbated image에 fine tuning되게 하는 term

$$and \quad \left\|\delta^{(i)}\right\|_p \leq \eta \;\; \forall i \in \{1, \dots N_{db}\}$$

CVSP Lab
Computer Vision & Signal Processing

Method

- Fully-trained Surrogate Model Guidance (**FSMG**)
  - Clean한 이미지 set으로 training한 DreamBooth model에 PGD(attack)를 통해 adversarial image를 만드는 알고리즘



Method

- Alternating Surrogate and Perturbation Learning(**ASPL**)
  - 모델 복제를 통해 DreamBooth를 시행할 모델과 PGD를 시행할 모델을 분리시켜 Purterbated data로만 학습을 함
    - 실제 attacker가 사용할 환경의 DreamBooth모델을 더욱 잘 모방할 수 있음
      - 실제 attacker는 Purterbated image로 Finetuning 할 것이기 때문
  - 1) 초기 weight를 가진 pretrained model과 보호할 이미지를 입력



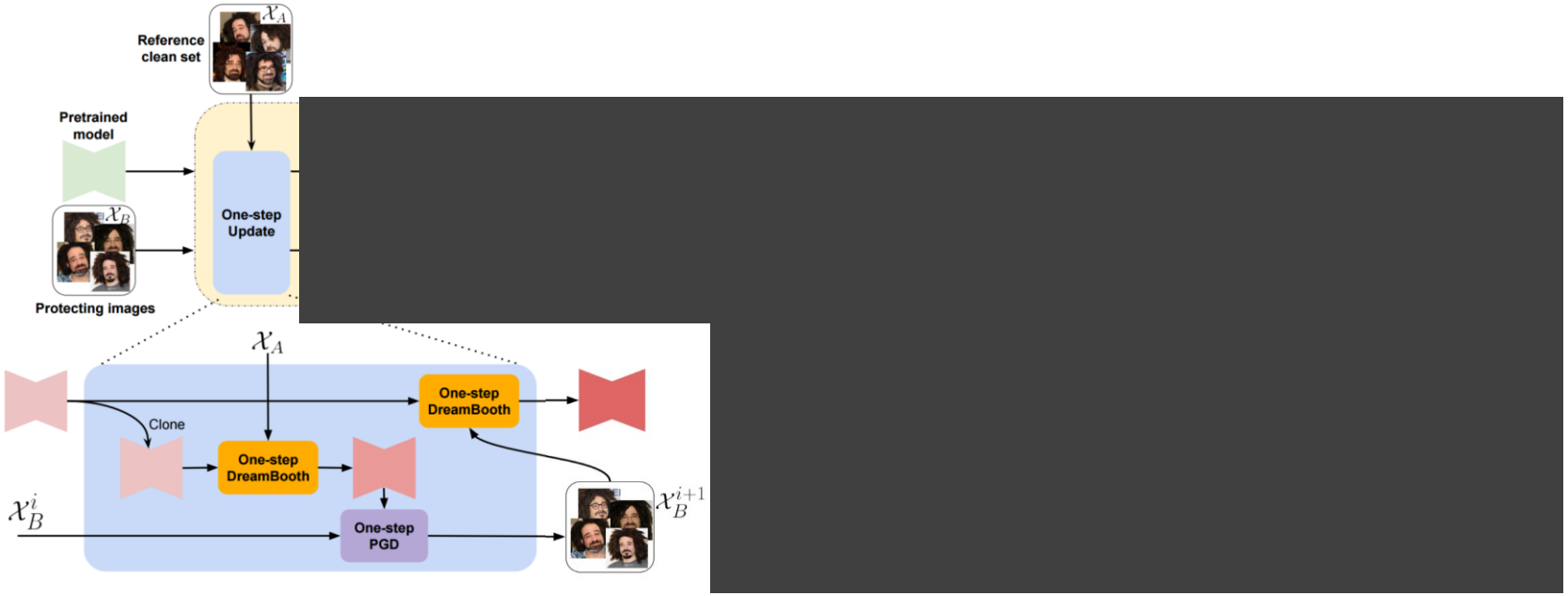
Method

- Alternating Surrogate and Perturbation Learning(**ASPL**)
  - 모델 복제를 통해 DreamBooth를 시행할 모델과 PGD를 시행할 모델을 분리시켜 Purterbated data로만 학습을 함
    - 실제 attacker가 사용할 환경의 DreamBooth모델을 더욱 잘 모방할 수 있음
  - 2) Pretrained model을 복제하여 clean image set으로 DreamBooth training을 시행한 모델로 PGD(attack실행)



Method

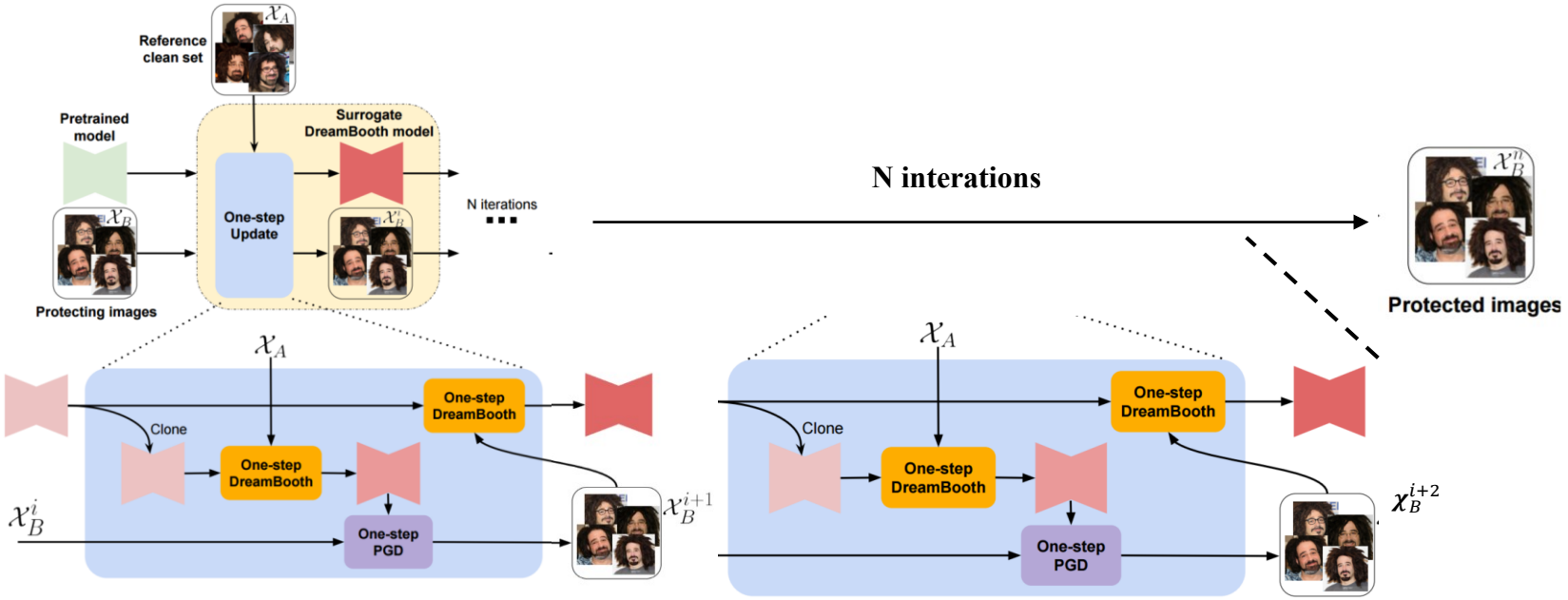
- Alternating Surrogate and Perturbation Learning(**ASPL**)
  - 모델 복제를 통해 DreamBooth를 시행할 모델과 PGD를 시행할 모델을 분리시켜 Purterbated data로만 학습을 함
    - 실제 attacker가 사용할 환경의 DreamBooth모델을 더욱 잘 모방할 수 있음
  - 3) PGD(attack)를 통해 나온 image set을 복제 이전의 기존 모델에 Dreambooth training 진행



CVSP Lab Computer Vision & Signal Processing

Method

- Alternating Surrogate and Perturbation Learning(**ASPL**)
  - 모델 복제를 통해 DreamBooth를 시행할 모델과 PGD를 시행할 모델을 분리시켜 Purterbated data로만 학습을 함
    - 실제 attacker가 사용할 환경의 DreamBooth모델을 더욱 잘 모방할 수 있음
  - 4) Purterbated image와 이 image로 DreamBooth된 모델을 다음 스텝에 넣어 N번 진행
    - 최종 Protected images 추출



Results

- Quantitative Results
  - Dataset
    - CelebA-HQ
      - 307명의 얼굴 데이터, 1024x1024, 한 명의 얼굴 데이터에 대해 최소 15장 이상
    - VGGFace2
      - 500x500 사이즈 이상, 한 명당 최소 15장 이상의 이미지를 포함
    - 각 데이터 셋에서 50명에 대한 얼굴 이미지 데이터 선택
      - 처음 12장의 이미지를 3가지 set으로 분할
        - ✓1) clean image set / 2) protect image set / 3) additional clean image set

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

CVSP Lab
Computer Vision & Signal Processing

Results

- Quantitative Results
  - Face Detection Failure Rate (FDFR) :
    - 감지할 수 없는 얼굴이 있는 비율(높을수록 방해가 잘 된 것을 의미)
  - Identity Score Matching (ISM)
    - 얼굴이 감지되면 얼굴인식 임베딩을 추출하여 clean image와 cosine distance 계산
      - (낮을수록 방해가 잘 된 것을 의미)

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

CVSP Lab
Computer Vision & Signal Processing

Results

[8] Terhorst, Philipp, et al. "SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[9] A. Mittal, A. K. Moorthy and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," in IEEE Transactions on Image Processing, vol. 21, no. 12, pp. 4695-4708, Dec. 2012, doi: 10.1109/TIP.2012.2214050.

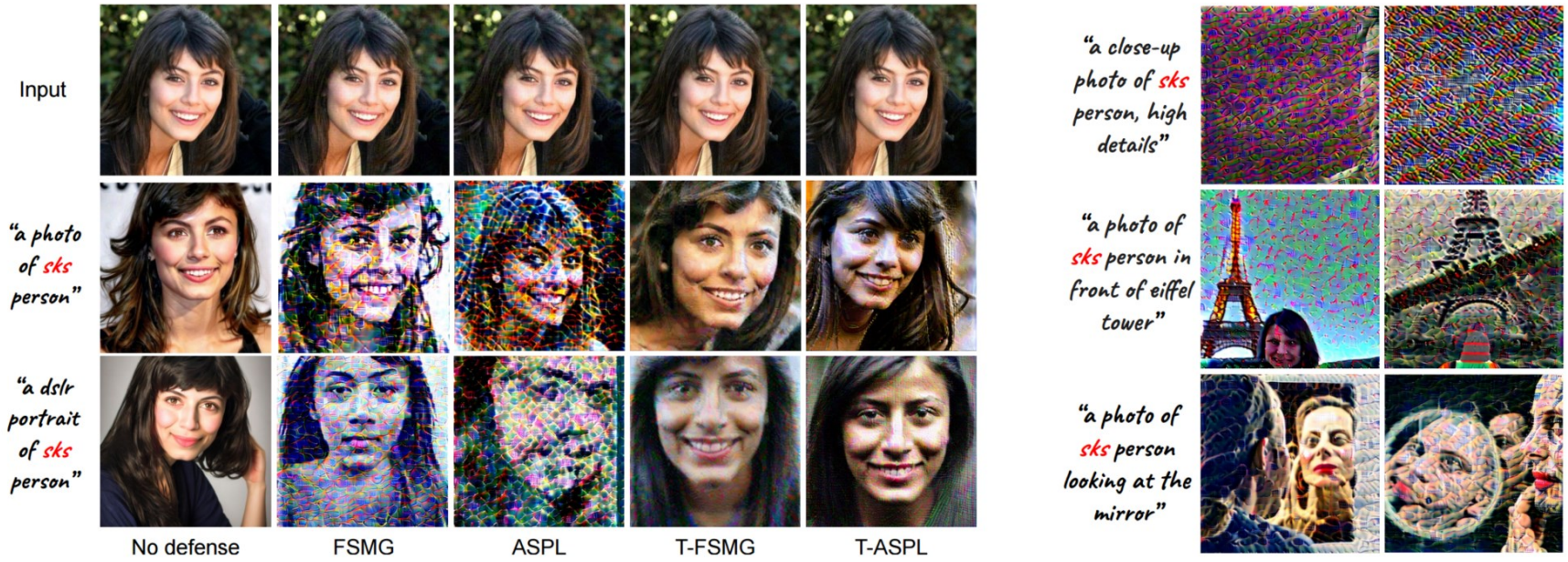
- Quantitative Results
  - SER-FQA[8]
    - 이미지를 Pre-trained Face Recognition Model에 입력 후 모델에 여러 패턴의 dropout을 적용
    - 모델의 출력에 대해 입력 이미지의 identity가 유지되는 정도를 측정
  - BRISQUE[9]
    - No-Reference Image Quality Assessment
    - 이미지의 통계적 특성 (픽셀 히스토그램 등)이 왜곡된 정도를 측정

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

CVSP Lab
Computer Vision & Signal Processing

Results

- Quantitative Results



(a) Comparison between proposed methods

(b) Inference prompts

]Figure 3: Qualitative defense results for two subjects in VGGFace2 in the convenient setting. Best viewed in zoom.

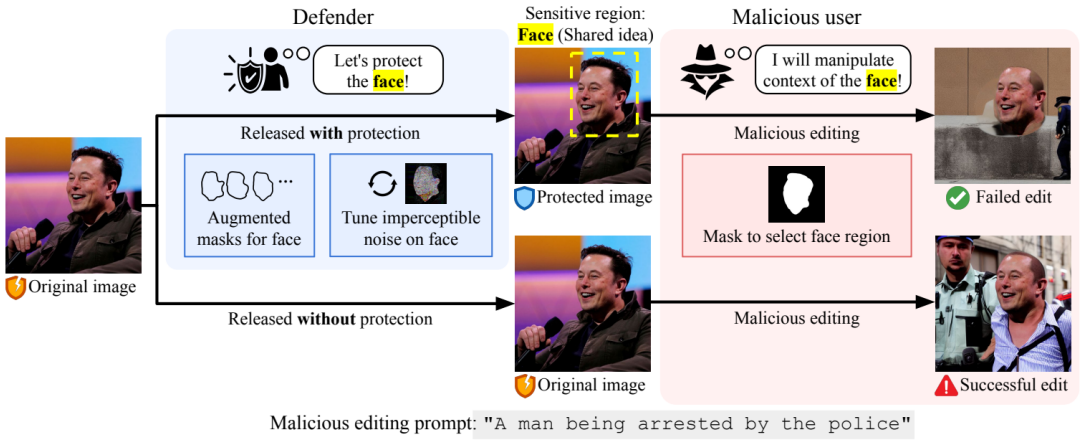
CVSP Lab
Computer Vision & Signal Processing

DiffusionGuard: A Robust Defense Against Malicious Diffusion-based Image Editing

ICLR 2025

Introduction

- Motivation
  - 기존 방법의 한계
    - 공격자가 마스크를 사용하여 편집할 경우 방어 기법을 우회할 수 있음
    - 얼굴과 같이 최소한의 영역을 선택하는 경우, 기존 보호 기법의 효과가 감소함
      - 기존의 adv noise는 이미지 전체 영역에 적용되기 때문
  - 제안하는 방법
    - DiffusionGuard
      - Diffusion model은 초기 diffusion process에서 mask 내부의 주요 영역을 유지하려는 경향이 있음
      - 때문에 초기에 adversarial attack을 가함으로써 편집을 방어함
    - Mask-augmentation
      - 다양한 형태의 마스크 입력에 대해 최소한의 영역에 대해 효과적인 adv noise를 적용하는 방법

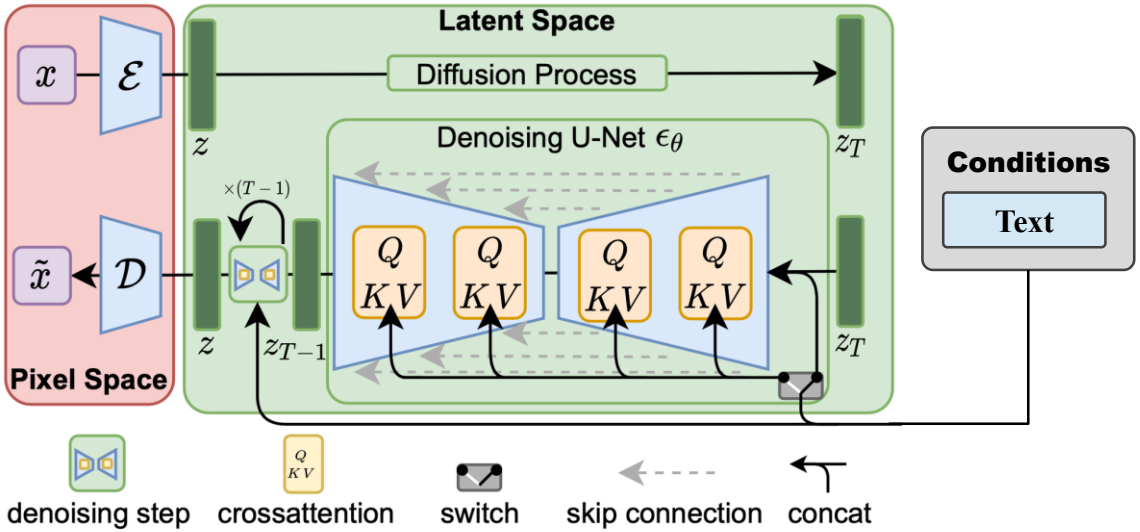


Background

- Text-to-Image Diffusion Models
  - Prompt를 condition으로 이미지를 생성할 수 있는 모델
    - Pre-trained text encoder에서 추출한 text embedding을 활용하여 이미지 생성
  - Train
    - 주어진 이미지 $\mathbf{x}$와 텍스트 y_{text} 쌍을 입력으로 할 때 denoising process에서 noise prediction model $\epsilon_\theta(\mathbf{x}_t; y_{text}, t)$를 활용하여 걷어낼 noise 추정
    - Loss:

$$L_{\text{diff}}(\theta; \mathbf{x}) = \exp_{t\sim u(1,T),\epsilon\sim N(0,I)}[\omega(\lambda_t)\|\epsilon_\theta(\mathbf{x}_t; y_{text}, t, M, \mathbf{x}_{src}) - \epsilon\|_2^2]$$

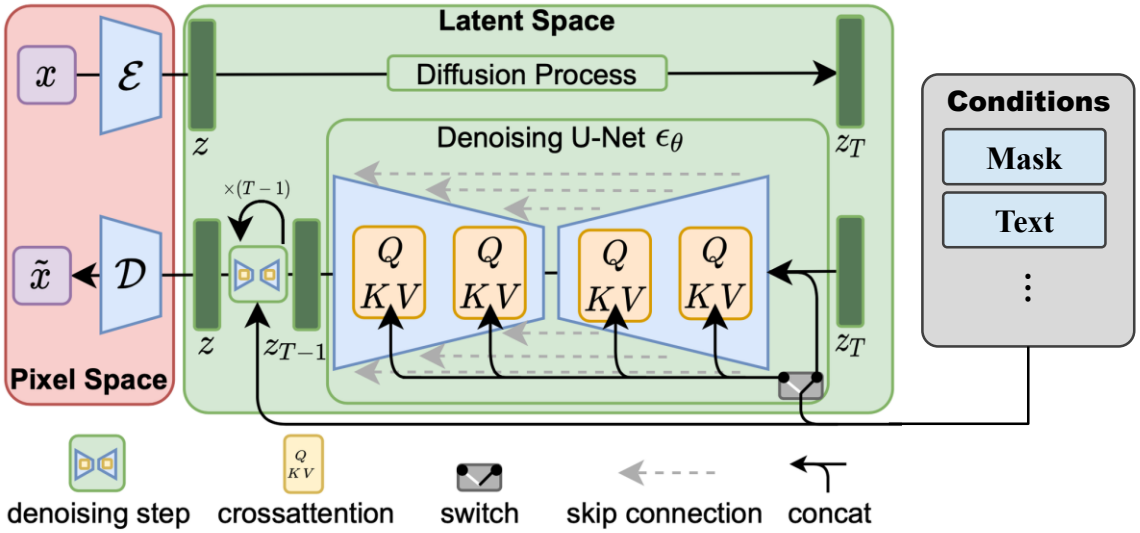
      - $\omega(\lambda_t)$: weighting function of time step t



Background

- Text-guided inpainting models
  - Mask를 활용한 이미지의 특정 영역 편집 모델
    - Pre-trained T2I Diffusion Model을 활용하여 mask와 prompt를 통해 mask가 적용된 영역을 편집
    - 원본 이미지 $\mathbf{x}_{\text{scr}}$와 binary mask M을 조건으로 추가하여 fine-tuning 수행
      - binary mask M: 이미지에서 새롭게 채워야할 영역 지정
    - Fine-tuning 중 이미지에 무작위로 영역을 지정하여 masking
    - Loss:

$$L_{inpaint}(\theta; \mathbf{x}_{\text{src}}, M) = \exp_{t\sim u(1,T),\epsilon\sim N(\mathbf{0},\boldsymbol{I})}[\omega(\lambda_t)\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t; y_{text}, t, M, \mathbf{x}_{\text{src}}) - \boldsymbol{\epsilon}\|_2^2]$$



Background

- Adversarial Examples against Diffusion Models
  - Adversarial Examples
    - 모델의 동작을 제어하기 위해 고의적으로 제작된 데이터
    - Clean image $\mathbf{x}$와 모델이 주어졌을 때, adv example은 $\mathbf{x}$에 perturbation δ를 추가
      - $\mathbf{x}+\delta$ 가 모델을 속이도록 설계됨
      - δ는 인간의 눈에는 거의 인식되지 않도록 미세함
    - Diffusion model의 각 denoising process를 교란(perturb)하여 loss를 최대화 하거나 diffusion model이 undesirable image를 생성하도록 강제
    - Perturbation δ

$$\delta = \underset{\|\delta\|_\infty \le \eta}{\operatorname{argmin}} \exp_{\epsilon \sim N(0,I)} \left[\left\| \hat{\mathbf{X}}(\mathbf{x}_{\text{src}} + \delta; \epsilon, y_{text}, M) - \mathbf{x}_{\text{target}} \right\|_2^2 \right]$$

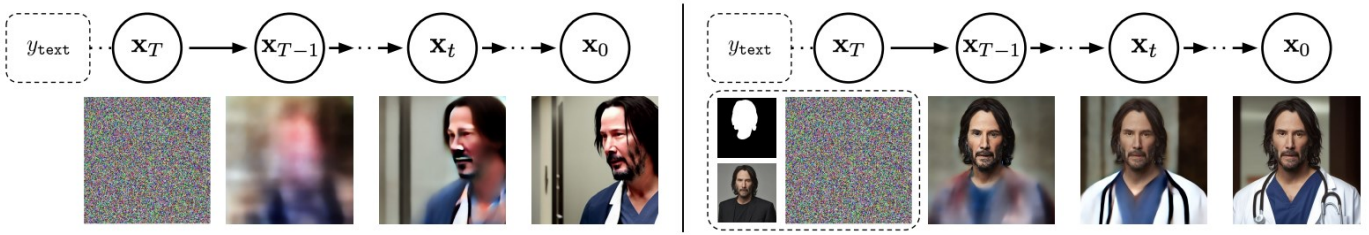
      - $\hat{\mathbf{X}}$: perturbed 이미지 $\mathbf{x}_{\text{src}} + \delta$, prompt y_{text}, mask M가 주어졌을 때 생성되는 image
      - $\|\delta\|_\infty \le \eta$: 인간이 쉽게 인식하지 못하도록 δ가 크지 않은 값을 가지도록 하는 조건

Method

- DiffusionGuard
  - Perturbating the Early Stages of the Diffusion Process
    - 일반적인 diffusion model은 denoising 단계에서 초기에는 coarse feature가 형성되고 이후 디테일을 생성하는 특징을 가짐
    - Inpainting model에서는 초기 denoising 단계에서부터 디테일이 먼저 생성됨
      - Mask가 적용되지 않는 영역은 원본 이미지의 데이터를 가지고 있기 때문
    - 초기 denoising process에서 추정된 noise의 크기 (L2 Norm)를 극대화 하는 새로운 perturbation δ 생성
      - 기존 방법은 loss를 최대화 하거나 특정 image에 더 가까워지도록 함
    - New perturbation δ:

$$\delta = \underset{\|\delta\|_\infty \leq \eta}{\operatorname{argmax}} \|\epsilon_\theta(\mathbf{x}_T; y_{text}, T, M_{tr}, \mathbf{x}_{\text{src}} + \delta)\|_2^2$$

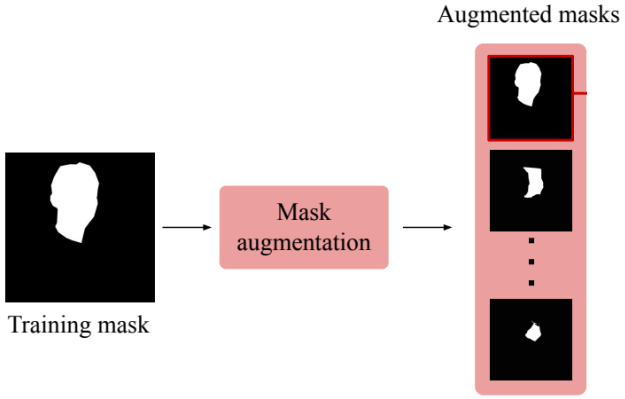
      - T: 초기 denoising step / $\mathbf{x}_T$: random noise / M_{tr}: binary mask, adv example을 생성할 때 사용되는 mask



(a) Denoising process of standard diffusion models (b) Denoising process of inpainting diffusion models

Method

- DiffusionGuard
  - Mask-Robust Adversarial Perturbation
    - 악용을 위해 사용하는 mask는 adv noise를 생성할 때 사용했던 mask와 다를 가능성이 큼
    - 따라서 다양한 형태의 mask에서도 효과적으로 편집을 방어할 수 있는 perturbation이 필요
    - Mask augmentation $A(\cdot)$를 제안하여 학습 시 사용했던 mask와 유사한 새로운 mask를 생성
      - 1. Mask의 contours (윤곽) 검출 및 contours 위 point 설정
      - 2. Contours를 따라 무작위 오프셋 적용하여 point를 내부로 이동하여 새로운 contours 설정
      - 3. 생성된 contours 내부 영역을 augmented mask로 설정



(a) Mask augmentation

Method

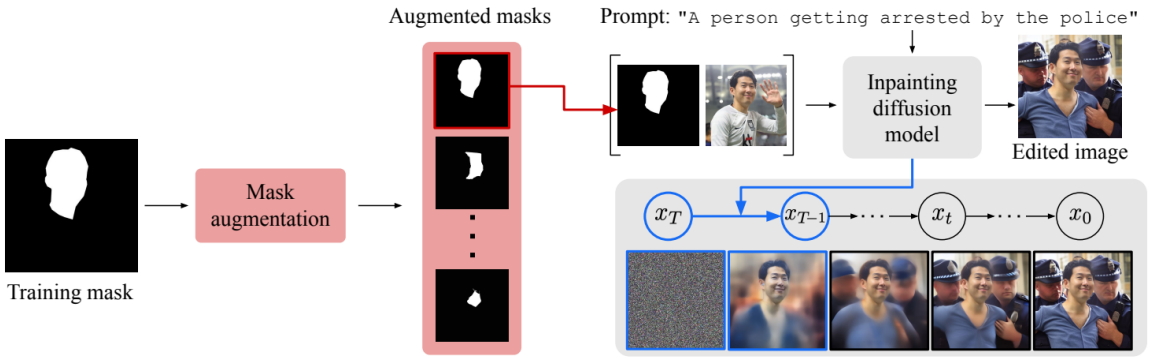
- DiffusionGuard
  - Mask-Robust Adversarial Perturbation
    - Augmented mask 집합 $A(M)$에 대해 adversarial loss를 최대화

$$\delta = \underset{\|\delta\|_\infty \le \eta}{\operatorname{argmax}} L_{adv}(\theta; \mathbf{x}+\delta, M_{tr}) = \exp_{M\sim A(M_{tr})}[\|\epsilon_\theta(x_T; y_{text}, T, M, \mathbf{x}+\delta)\|_2^2]$$

      - L_{adv}: 초기 단계에서 adv noise를 생성하기 위해 사용되는 loss
      - $M\sim A(M_{tr})$: augmented mask 중 하나를 stochastically sampling하여 최적화 수행
    - 각 iteration에서 mask M을 sampling 하여 adversarial perturbation δ 업데이트
    - Projected Gradient Descent (PGD) 적용

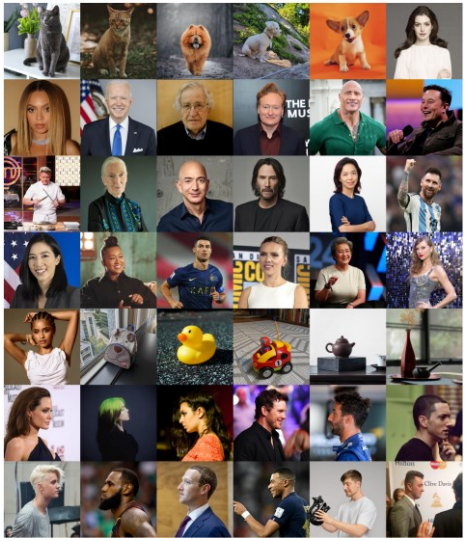
$$\delta \leftarrow \text{Proj}_{\|\delta\|_\infty \le \eta}(\delta - \gamma \cdot \text{sign}(\nabla_\delta L_{\text{adv}}))$$

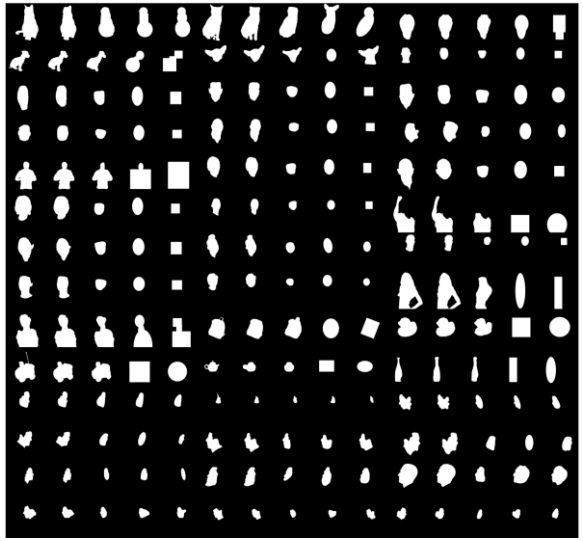
      - γ: step size



Experiments

- Setting
  - Dataset
    - 총 42개의 이미지로 구성되며, 각각 5개의 고유한 마스크를 가짐
      - 유명인 초상화(celebrity portraits) - 32개
      - 무생물 오브젝트(inanimate objects) - 5개
      - 동물(animals) - 5개
    - 각 이미지에 대해 10개의 edit prompts를 고려하여, 총 2,100개의 편집 작업을 수행
      - 42개 이미지 × 5개 마스크 × 10개 프롬프트 = 2,100개





CVSP Lab
Computer Vision & Signal Processing

Experiments

[10] Gal, Rinon, et al. "Stylegan-nada: Clip-guided domain adaptation of image generators." ACM Transactions on Graphics (TOG) 41.4 (2022): 1-13.
[11] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

- Quantitative Results
  - Evaluation metrics
    - CLIP directional similarity[10]
      - 편집된 이미지에서의 편차와 텍스트에서의 편차 간 alignment를 측정
    - CLIP similarity[11]
      - 편집된 이미지와 원본 이미지 간의 text fidelity를 측정
      - CLIP 임베딩의 cosine similarity를 기반으로 평가

Method	PSNR ↓	CLIP Dir. Sim. ↓	ImageReward ↓	CLIP Sim. ↓
		`Seen` (1 Mask, Train set)		
Unprotected	N/A	24.40	-1.365	30.15
PhotoGuard (Salman et al., 2024)	12.87	21.17 (Δ-3.23)	-1.537 (Δ-0.172)	27.89 (Δ-2.26)
DiffusionGuard	**12.60**	**18.95 (Δ-5.45)**	**-1.807 (Δ-0.442)**	**26.55 (Δ-3.60)**
		`Unseen` (4 Masks, Test set)		
Unprotected	N/A	24.29	-1.315	30.72
PhotoGuard (Salman et al., 2024)	14.53	23.30 (Δ-0.99)	-1.357 (Δ-0.042)	30.30 (Δ-0.42)
DiffusionGuard	**13.19**	**21.84 (Δ-2.45)**	**-1.557 (Δ-0.242)**	**29.05 (Δ-1.67)**
AdvDM (Liang et al., 2023)	13.37	24.27 (Δ-0.02)	-1.361 (Δ-0.046)	30.97 (Δ+0.25)
Mist (Liang & Wu, 2023)	14.51	23.93 (Δ-0.36)	-1.307 (Δ+0.008)	30.79 (Δ+0.07)
SDS(-) (Xue et al., 2024)	14.32	23.85 (Δ-0.44)	-1.237 (Δ+0.078)	30.78 (Δ+0.06)

CVSP Lab
Computer Vision & Signal Processing

Experiments

[12] Xu, Jiazheng, et al. "Imagereward: Learning and evaluating human preferences for text-to-image generation." Advances in Neural Information Processing Systems 36 (2024).

- Quantitative Results
  - Evaluation metrics
    - ImageReward
      - Human preference dataset을 기반으로 fine-tuning된 human-aligned vision-language model 을 기반으로 image quality 및 text fidelity를 평가
    - PSNR
      - Unprotected image와 protected image의 편집 결과 차이를 정량적으로 평가하기 위해 사용

Method	PSNR ↓	CLIP Dir. Sim. ↓	ImageReward ↓	CLIP Sim. ↓
		`Seen` (1 Mask, Train set)		
Unprotected	N/A	24.40	-1.365	30.15
PhotoGuard (Salman et al., 2024)	12.87	21.17 (Δ-3.23)	-1.537 (Δ-0.172)	27.89 (Δ-2.26)
DiffusionGuard	**12.60**	**18.95 (Δ-5.45)**	**-1.807 (Δ-0.442)**	**26.55 (Δ-3.60)**
		`Unseen` (4 Masks, Test set)		
Unprotected	N/A	24.29	-1.315	30.72
PhotoGuard (Salman et al., 2024)	14.53	23.30 (Δ-0.99)	-1.357 (Δ-0.042)	30.30 (Δ-0.42)
DiffusionGuard	**13.19**	**21.84 (Δ-2.45)**	**-1.557 (Δ-0.242)**	**29.05 (Δ-1.67)**
AdvDM (Liang et al., 2023)	13.37	24.27 (Δ-0.02)	-1.361 (Δ-0.046)	30.97 (Δ+0.25)
Mist (Liang & Wu, 2023)	14.51	23.93 (Δ-0.36)	-1.307 (Δ+0.008)	30.79 (Δ+0.07)
SDS(-) (Xue et al., 2024)	14.32	23.85 (Δ-0.44)	-1.237 (Δ+0.078)	30.78 (Δ+0.06)

Experiments

- Qualitative Results



Experiments

- Qualitative Results



감사합니다.

CVSP Lab
Computer Vision & Signal Processing

Reference

- [1] Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.
- [2] https://www.youtube.com/watch?v=yea-flCXHSg
- [3] https://www.youtube.com/watch?v=PdojnAQE92U
- [4] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.
- [5] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
- [6] Mądry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." stat 1050.9 (2017).
- [7] Gal, Rinon, et al. "An image is worth one word: Personalizing text-to-image generation using textual inversion." arXiv preprint arXiv:2208.01618 (2022).
- [8] Terhorst, Philipp, et al. "SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
- [9] A. Mittal, A. K. Moorthy and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," in IEEE Transactions on Image Processing, vol. 21, no. 12, pp. 4695-4708, Dec. 2012, doi: 10.1109/TIP.2012.2214050.
- [10] Gal, Rinon, et al. "Stylegan-nada: Clip-guided domain adaptation of image generators." ACM Transactions on Graphics (TOG) 41.4 (2022): 1-13.
- [11] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.
- [12] Xu, Jiazheng, et al. "Imagereward: Learning and evaluating human preferences for text-to-image generation." Advances in Neural Information Processing Systems 36 (2024).