UNIVERSITÀ DEGLI STUDI DI
NAPOLI FEDERICO II
Scuola Politecnica e delle Scienze di Base
Corso di Laurea Magistrale in Ingegneria Informatica

tesi di laurea magistrale

Evaluating Artist Protection Methods Against Style Mimicry by Diffusion Models

2022/2023

relatore
Ch.ma prof. Luisa Verdoliva

candidato
Raffaele Russo
Matr. M63001325

Introduction

- Diffusion Models (DMs) generate high-quality pictures from a prompt
- 3 billion pictures produced with Adobe Firefly in 6 months
- Several applications are possible (e.g. copying the style of an image)

"Cute little robot artist painting on a 3D easel."

"A serene sunset over a calm lake with a solitary cabin in the distance."

"A close-up macro photo of a small, adorable, cute, and colorful spider."

Problem

- Diffusion Models (DMs) can represent a threat to artists' work
- They can be used to generate:
  - ✓ Illegal and fake content
  - ✓ Images with the style of a target artist

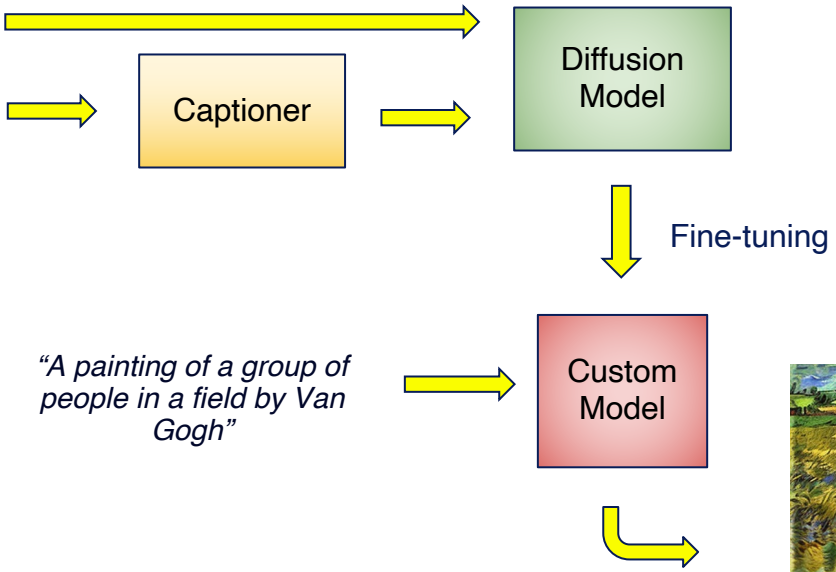
State Fair won by AI generated image

It is easy to reproduce images with the style of an artist

Threat model

- A malicious attacker can download images from a target artist from the web
- Use a pre-trained diffusion model and fine-tune it
- Then the model will be able to generate images with the style of the target artist

Van Gogh pictures



Protection methods (1)

- **GLAZE**: it enables artists to apply “style cloaks” to their art
- These cloaks apply barely perceptible perturbations to images

Van Gogh picture

Style transfer to Cubism by Picasso Style

Feature Extractor

Feature Extractor

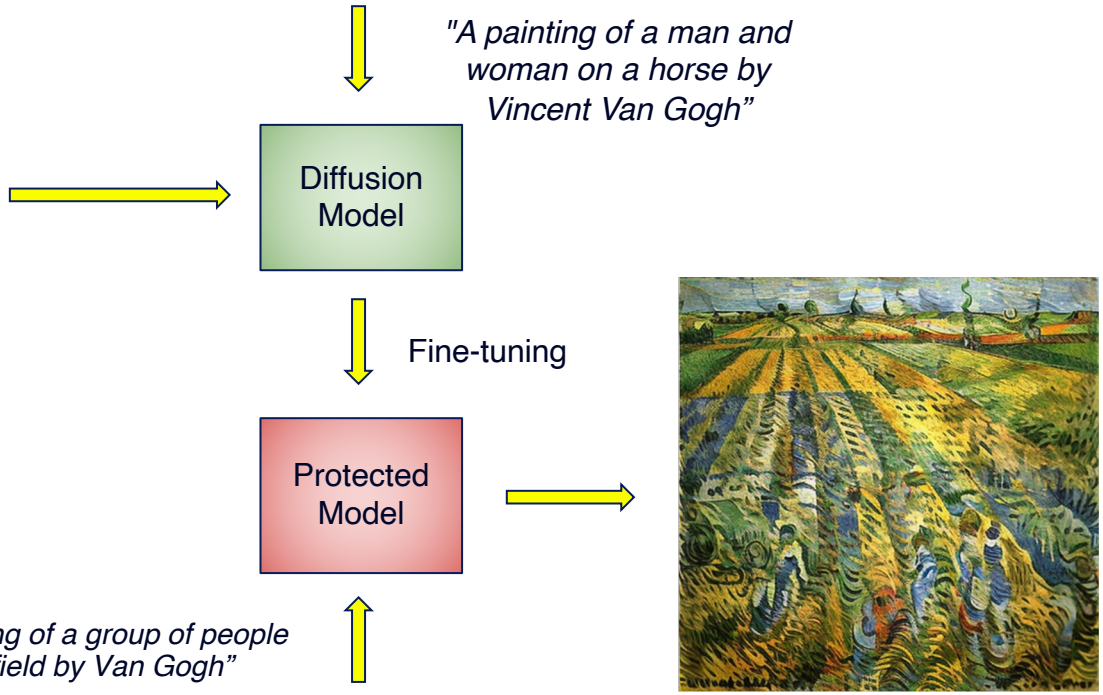
$$\min_{\boldsymbol{\delta}} \|\mathcal{E}(\Omega(\mathbf{x}, T)), \mathcal{E}(\mathbf{x}+\boldsymbol{\delta})\|_2^2 + \lambda \cdot \max\left(\mathrm{LPIPS}\left(\mathbf{x}, \mathbf{x}+\boldsymbol{\delta}\right) - \Delta_L, 0\right)$$

Van Gogh Glaze picture

Protection methods (2)

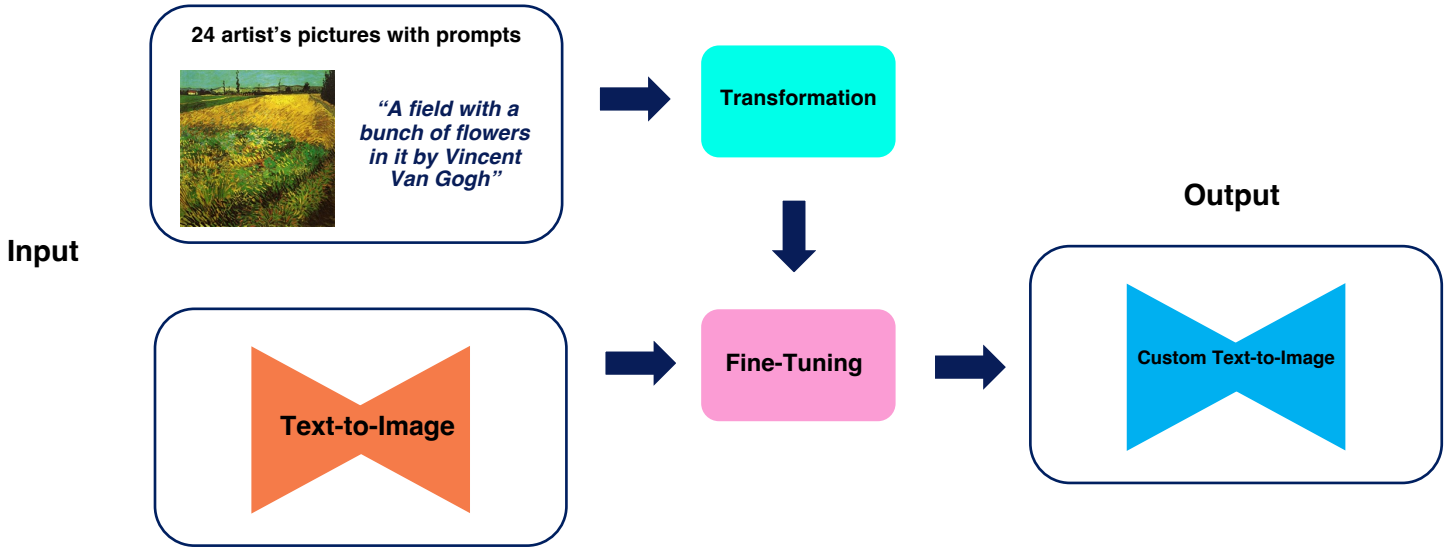
- **GLAZE**: it enables artists to apply “style cloaks” to their art
- When used as training data, they mislead generative models that try to mimic a specific artist

Van Gogh Glaze picture



Goal of this thesis

- **Assess the robustness of state-of-the-art protection methods** against style mimicry by diffusion models
- **Adversarial attacks**: protected images are analyzed against target attacks, e.g. image purification
- **Laundering**: protected images are analyzed against not malicious perturbations, i.e. JPEG compression, resizing, blurring



Experiments: Data Preparation

- 124 random paintings (here we show Van Gogh and artxman)
- 24 images will be used for fine-tuning, the latter 100 for testing
- Captions created with vit-gpt2
- Victim artist name is appended to the caption, forming the final prompt

"A painting of a man and woman on a horse by Vincent Van Gogh"

"A woman in a hat sitting on a tree stump by artxman"

No fine-tuning

"A man with a cartoon face on a colorful head by artxman"

"A painting of a man riding a horse by Vincent Van Gogh"

Clean images

"A man with a cartoon face on a colorful head by artxman"

"A painting of a man riding a horse by Vincent Van Gogh"

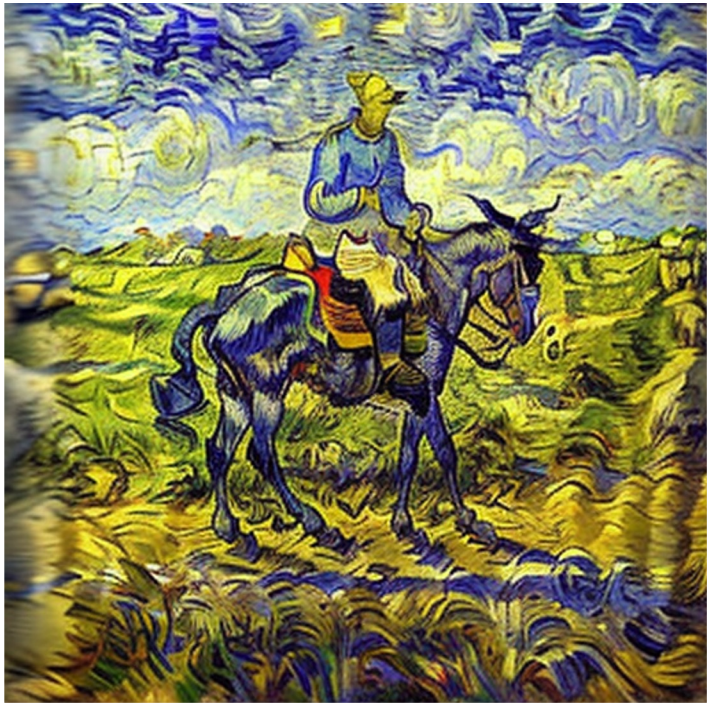
Glaze images

"A man with a cartoon face on a colorful head by artxman"

"A painting of a man riding a horse by Vincent Van Gogh"

Glaze Effectiveness

- Post fine-tuning with protected images, the model's output lacks Van Gogh's signature brushwork
- Previously distinct and textured strokes are now uniformly flattened



Clean

Glaze

Evaluation Metrics

- No-reference metric to assess the quality of diffusion model's output:
  - Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) ↓
- Reference metric to measure distortion between diffusion model's output after fine-tuning with clean pictures and transformed pictures
  - Fréchet Inception Distance (FID) ↓

↓ indicates higher quality or similarity for lower values of the metric

Adversarial attacks

- Image purification by removing imperceptible perturbations
- **Similarity Condition**: purified image visually close to the perturbed and original image
- **Consistency Condition**: DM reconstructed image for the purified image visually close to itself

Original

Purified

Attack Effectiveness

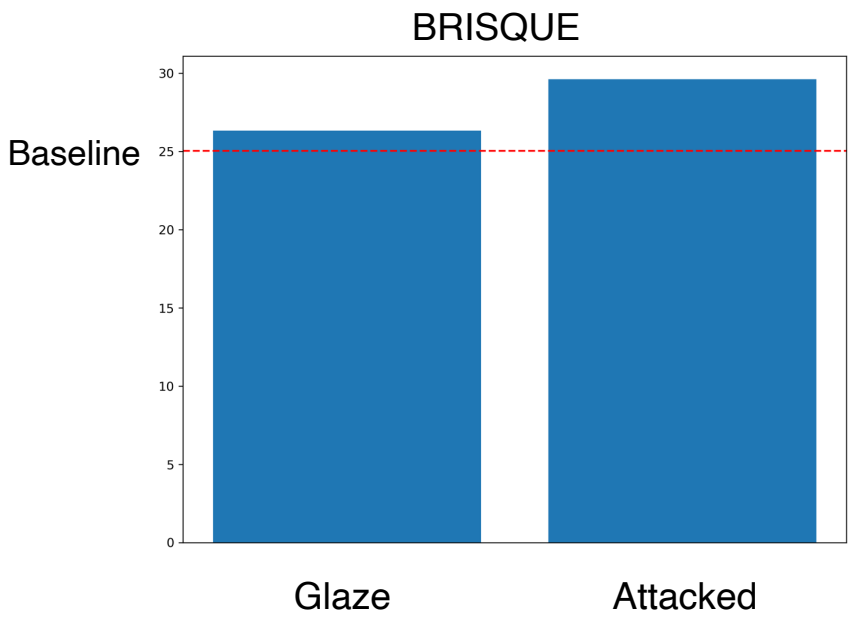
- Post fine-tuning with the purified images, the model's output partially reacquires Van Gogh's features

Glaze

Attacked

Evaluation Metrics

- Glaze: No impact on the output quality
- Adversarial attack: Slightly degrades output quality in terms of BRISQUE
- Both maintain visual similarity to clean-picture fine-tuned output in terms of FID
- BRISQUE detects minor differences, but they do not affect overall similarity captured by FID
- Consistent results across both artists



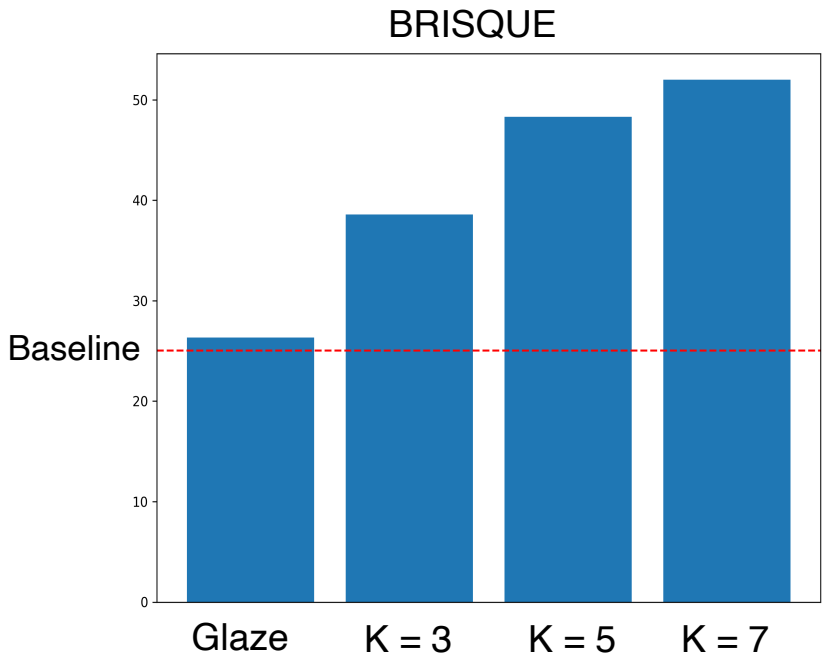
Laundering Experiments

- Glaze robustness evaluated against trivial transformations
  - **Gaussian Blur (GB)** with kernel size k = 3, 5, 7 and **σ** = 2
  - **Resizing** to 256 x 256
- Resizing and GB with k = 3 led to style mimicry, bypassing Glaze protection

Clean | Glaze | GB k = 3 | Resize

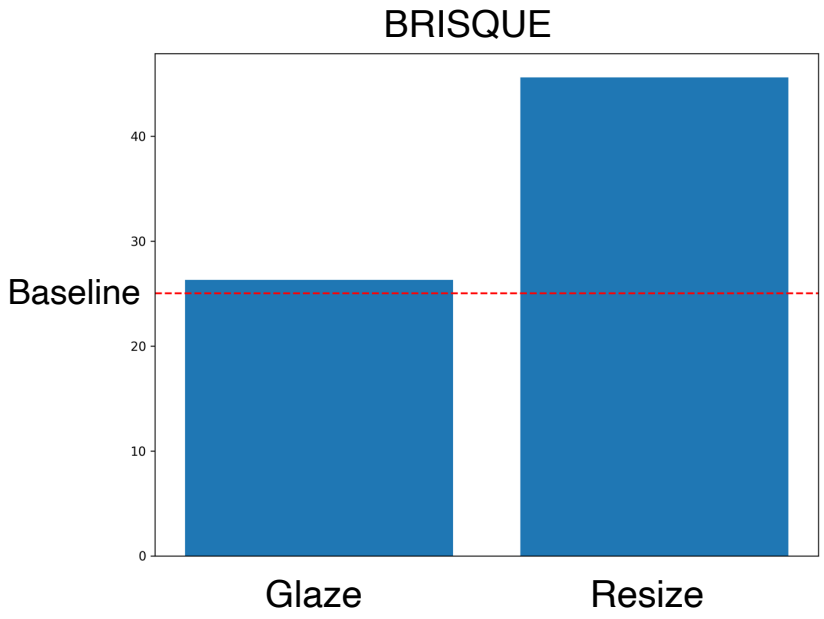
Laundering Evaluation: Gaussian Blur (GB)

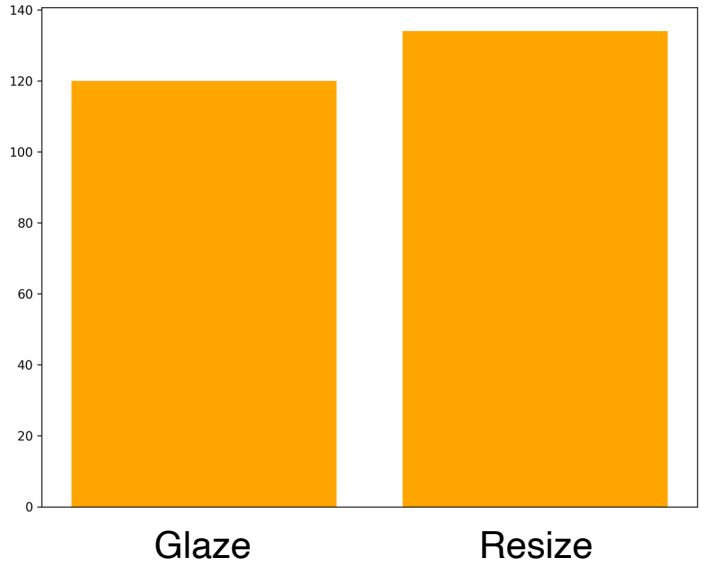
- Quality degrades for high k values in terms of BRISQUE
- GB (k = 3) maintains visual similarity to clean-picture fine-tuned output in terms of FID
- High k values result in simultaneous increases in both FID and BRISQUE
- Consistent results across both artists



Laundering Evaluation: Resizing

- Degrades quality in terms of BRISQUE
- Maintains visual similarity to clean-picture fine-tuned output in terms of FID
- Consistent results across both artists





Conclusions

- Glaze protection is vulnerable to adversarial attacks
- In addition, we found that it can easily be removed with trivial transformations
- Further exploration is required to safeguard artists from style mimicry
- Currently, there is no objective metric to assess the artist's features in a picture

Thank You