# 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

**MI2 DataLab Winter Research Seminar 2023**

**Tymoteusz Kwieciński**

# Authors
## SAND lab

Shawn Shan

Jenna Cryan

Emily Wenger
Meta AI

Haitao Zheng
Fellow of the IEEE
Neubauer Professor of CS

Rana Hanocka

Ben Y. Zhao

# Awards

Paper was presented at USENIX Security 2023

Winner: Distinguished Paper Award
Winner: USENIX Internet Defense Prize 2023

usenix
THE ADVANCED COMPUTING SYSTEMS ASSOCIATION

# Paper contributions

Glaze – tool for disrupting the images to remove artists' personal style from their artwork

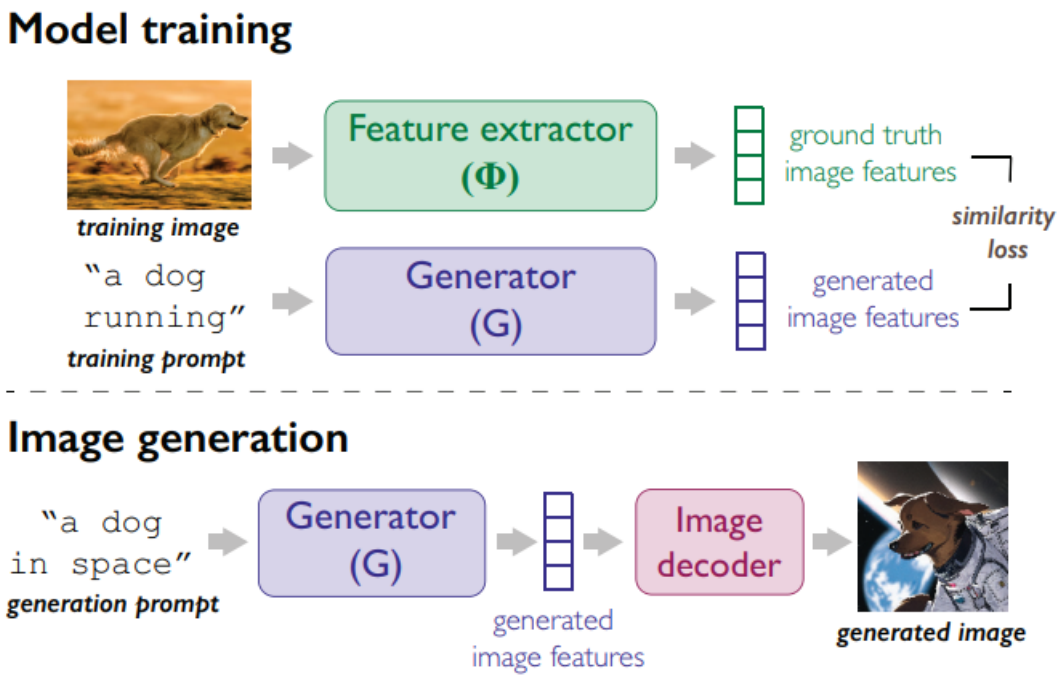
Raising awarness about illegal usage of artworks online


Glaze on instagram

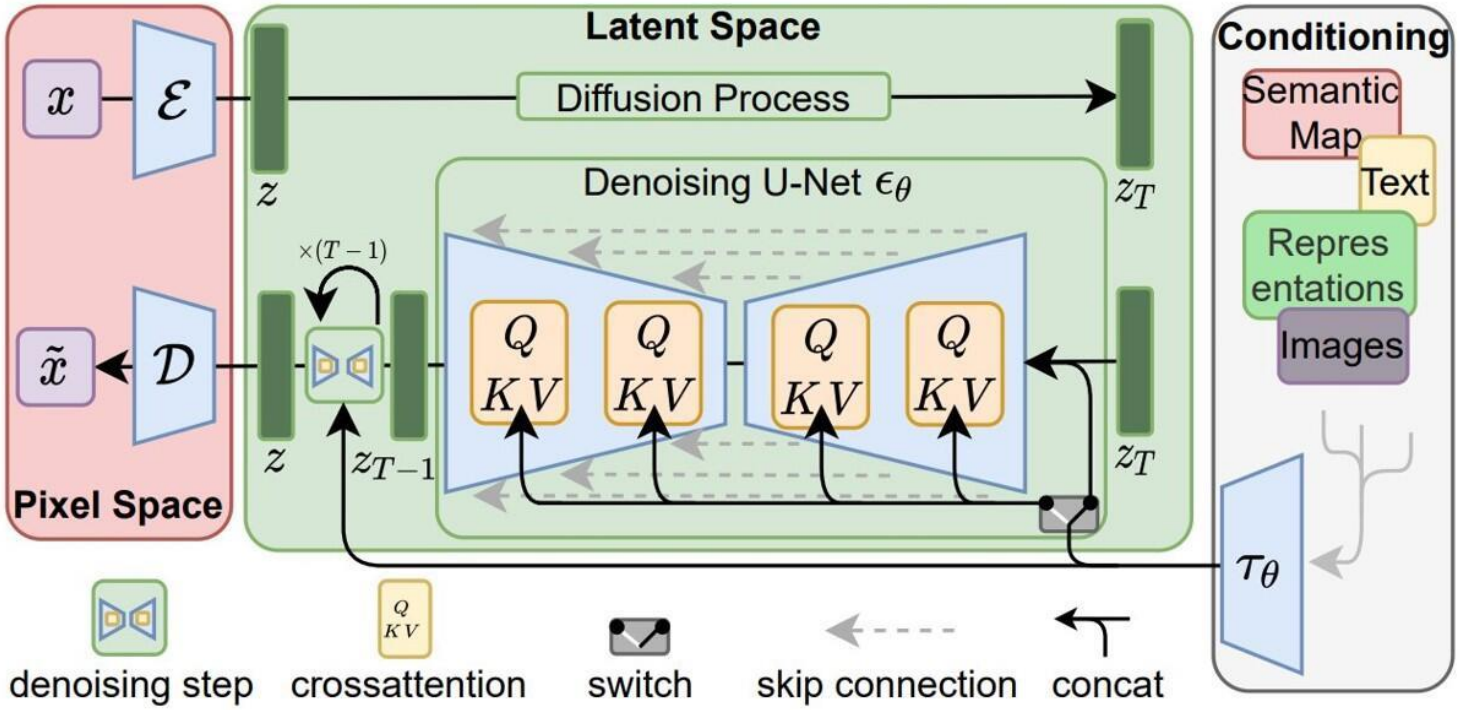
# Stable diffusion recap

Text-to-image models, such as Stable Diffusion consists of 3 elements:

- **Autoencoder (VAE)**, which casts images into smaller latent space and reverses the proces
- **Text-encoder**, e.g. CLIP, creating latent representation of text prompt
- **Diffusion model**, which input is random noise and is conditioned by text-encoder; its output is representation of image in latent space



# Stable diffusion recap



# Ethics

## Copyrighted images

Large datasets are not filtered for privacy

LAION include copyrighted artworks

## Style Mimicry

Artists develop their personal style for years

Models may be fine-tuned on samples of artists' artworks

Some more popular artists may be already recognized without fine-tuning


An example of style mimicry

# How artists understand the topic of GenAI and do they view is as threat?

**1207** artists were surveyed

91% of surveyed artists declared that they read extensively about AI art.

97% view AI mimicry as real danger

95% of artists post their artwork online

The most significant concern of artists is **scraping of existing artworks without permission or compensation**

Comments on AI art impact on artists:

*Junior positions will become extinct*

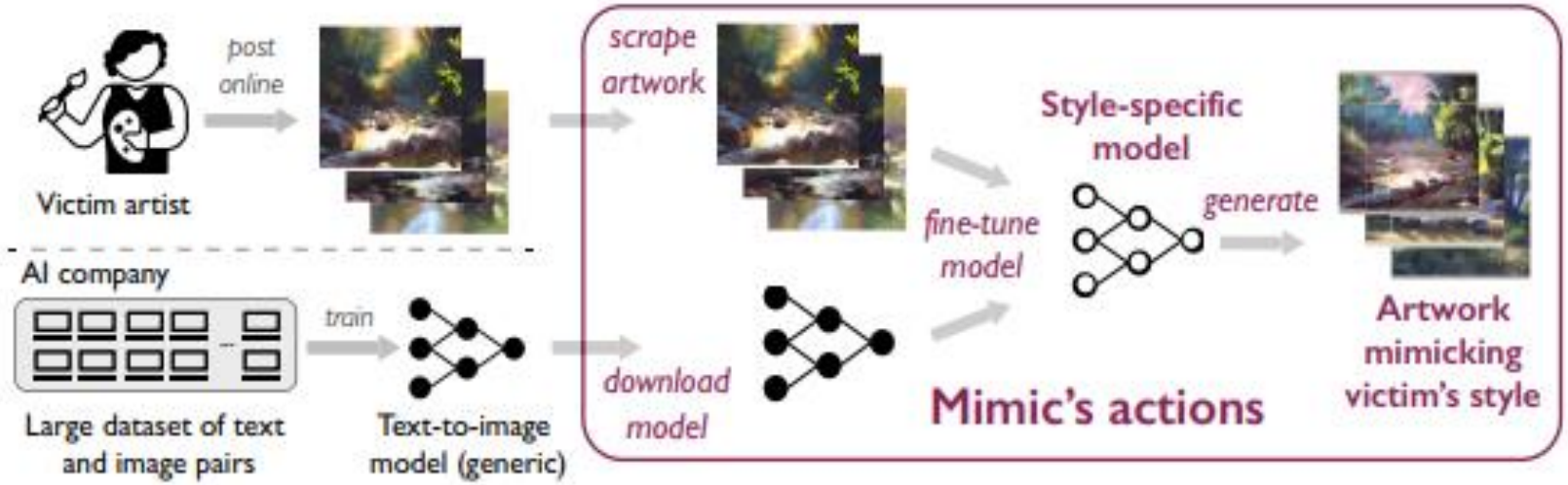
*AI art has unmotivated myself from uploading more art and made me think about all the years I spent learning art*

# Style mimicry attack

Mimic wants to train a model reasembling the victim style in *high quality*

we assume that the mimic has access to **victim's artwork** and **significant computational power**

Threat model is fine-tuned on the victims artworks



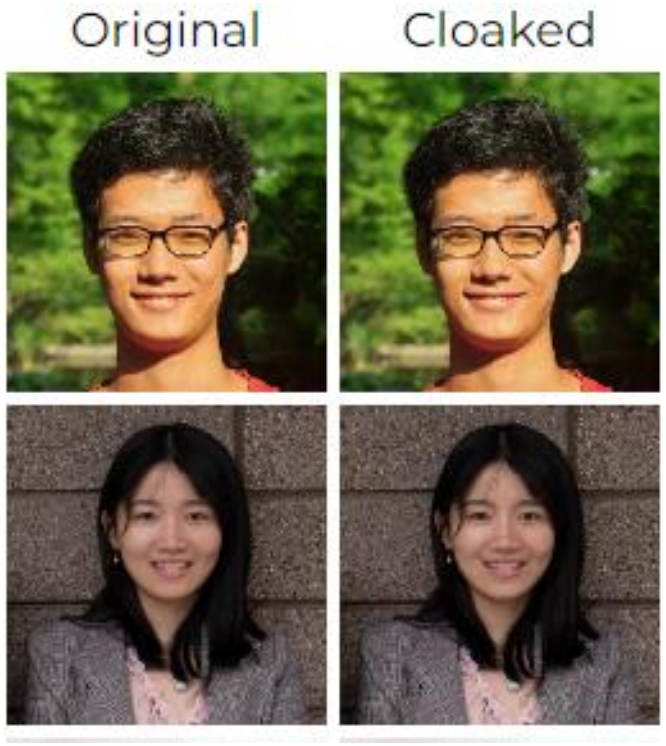

Overview of mimicry attack scenario

# Facial cloaks - Fawkes

The idea for *Glaze* comes from the similar problem in face recognition models

The individual face features used to recognize faces are being disrupted, yet rest of the image stays the same

In this way model cannot be trained well and **won't recognize** one's face



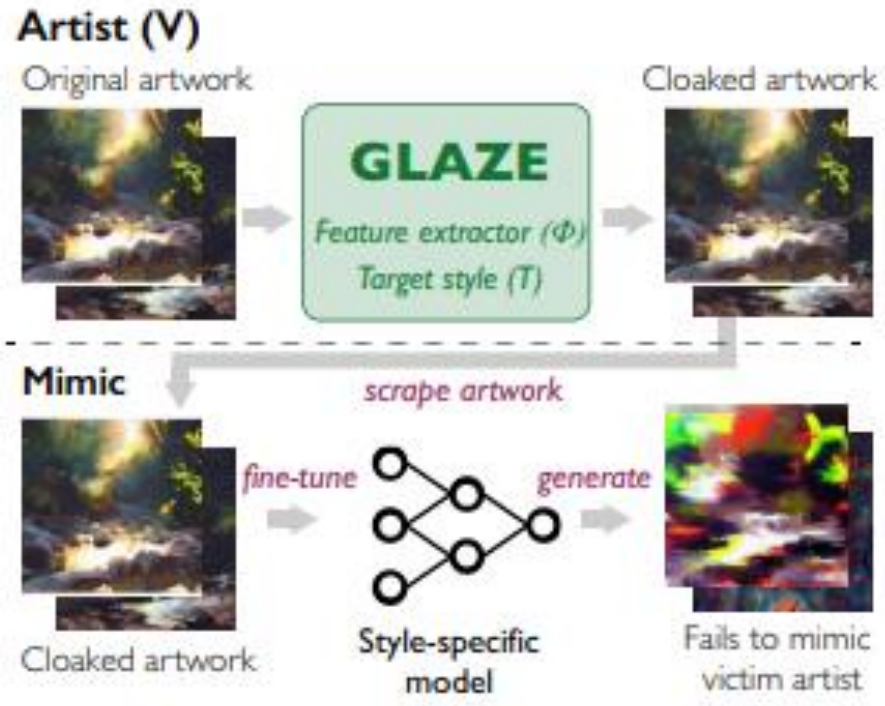
# Glaze idea

Key intuition is to isolate **style-specyfic features** of an artist's original artwork

We define V as victim artist, T as Target style
Φ as feature extractor

## Algorithm

1. Choose Target Style
2. Style transfer
3. Compute cloak perturbation



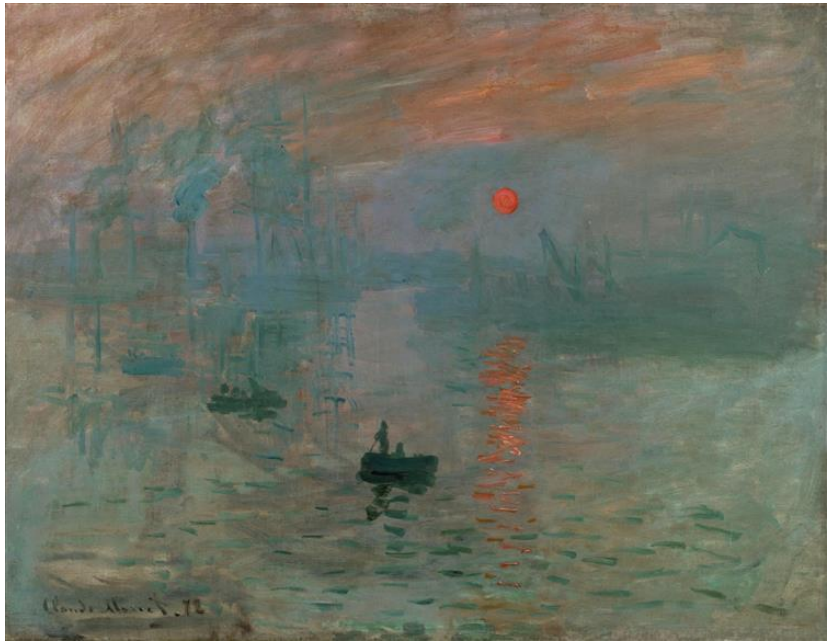
# Choice of target style

For new user *Glaze* randomly selects T from styles reasonably different from V's style

Styles are choosen from public available styles (e.g. Monet, Van Gogh, Picasso)

Database of styles is collection artworks from WikiArt of 1119 prominent artists

Candidate set are all styles between image of V's style centroid and few images in T centroid in feature space extracted by Φ

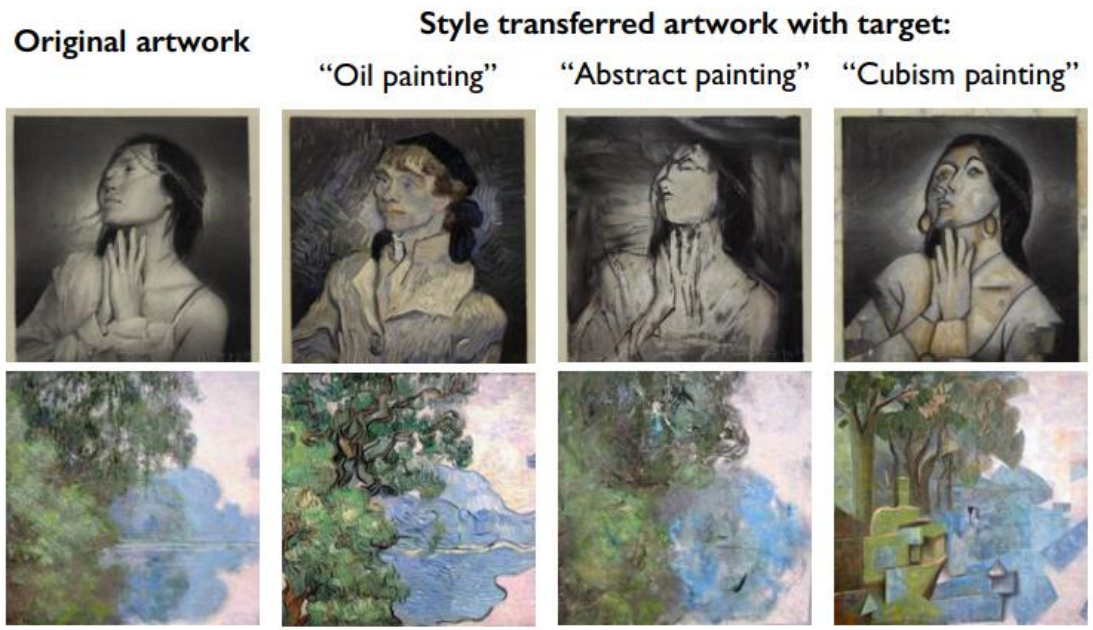


*Impression, sunrise;* by Claude Monet
One of the images used in choice of target style

# Style transfer

Using a pretrainied model Ω it shifts style image to target style T Ω(x,T)

Authors used Stable Diffusion to transfer styles using apropiate prompt



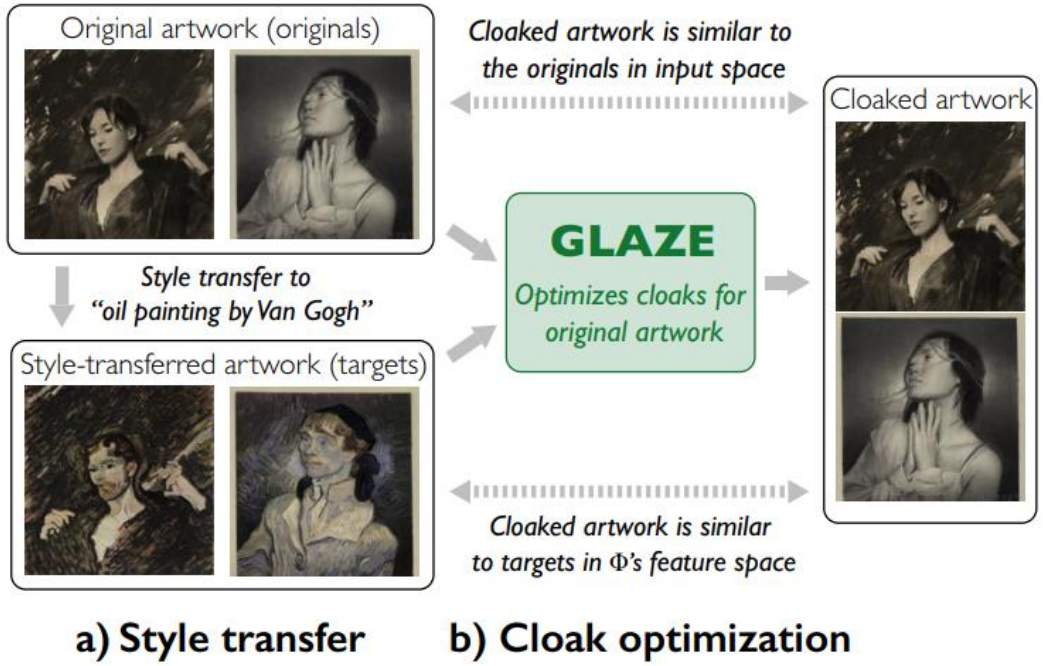
# Cloak perturbation

For choosen target style it applies a *cloak* on the image

The cloak disrupts the artist style-specific features, shifting it into the target style vectors

$$\min_{\delta_x} Dist\left(\Phi(x+\delta_x), \Phi(\Omega(x,T))\right),$$

$$\text{subject to } |\delta_x| < p,$$



# Cloak perturbation

To calculate visual image perturbation authors used LPIPS – Learned Perceptual Patch Similarity

LPIPS simply utilizes a pretrained neural network to rate if two images are visually similar

Authors optimized the following function:

$$\min_{\delta_x} ||\Phi(\Omega(x,T)),\Phi(x+\delta_x)||_2^2 + \alpha \cdot max(LPIPS(\delta_x) - p, 0)$$

# Measurement of cloak quality

## Artist based score PSR

Artists rated successfulnes of protection on 5-level Likert scale

Artist-rated PSR is percentage of artworks that artists classified as unsuccessful in case of style mimicry

## CLIP-based genre shift

*Glaze* succeeds if the mimicked artwork is classified into different art genre from the original artwork

CLIP-score is defined as percentage of mimicked artworks that are not classified with victim's style

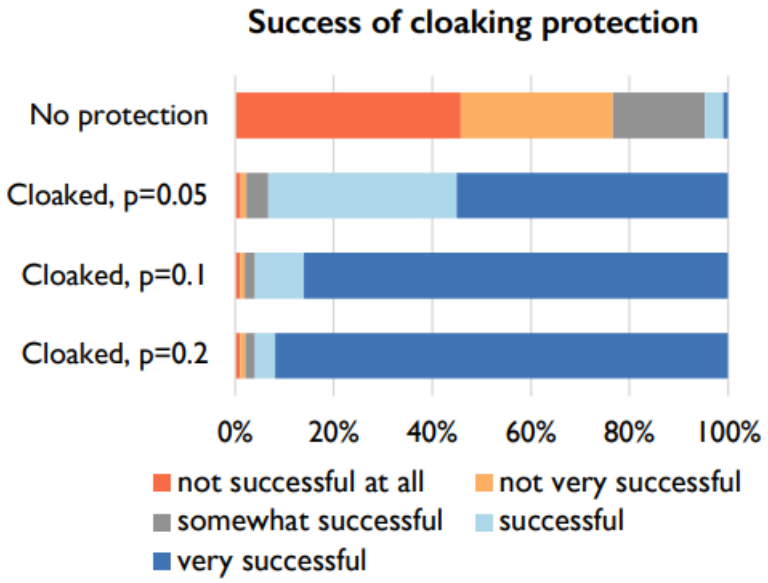
# Glaze robustness

*Glaze* makes mimicry attacks less successful

Performance is slightly better for current than for historic artists

Artists are willing to tolerate perturbations generated by *glaze*, since it gives them better protection against style mimicry

| Perturbation budget | Artist-rated PSR | CLIP-based genre shift |
|---|---|---|
| 0 (no cloak) | $4.6 \pm 1.4\%$ | $2.4 \pm 0.8\%$ |
| 0.05 | $93.3 \pm 0.6\%$ | $96.0 \pm 0.3\%$ |
| 0.1 | $95.9 \pm 0.4\%$ | $98.2 \pm 0.1\%$ |
| 0.2 | $96.1 \pm 0.3\%$ | $98.5 \pm 0.1\%$ |



# Glaze robustness

# Challenges

Mimic/artist uses different feature extractor

Mimic has access to some of uncloaked artist artwork

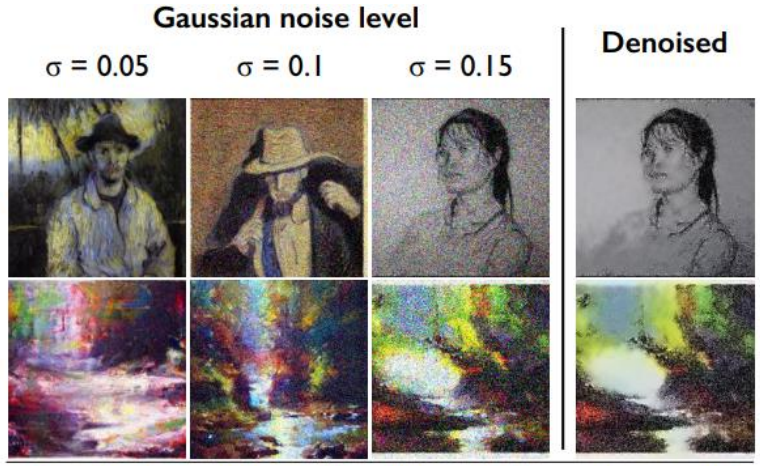
| | Feature extractors used by artist and mimic | | | | Percentage of artwork cloaked | | | |
|---|---|---|---|---|---|---|---|---|
| | Artist: **no cloaking** Mimic: **Φ-A** | Artist: **Φ-A** Mimic: **Φ-A** | Artist: **Φ-B** Mimic: **Φ-A** | Artist: **Φ-C** Mimic: **Φ-A** | 0% cloaked | 25% cloaked | 50% cloaked | 75% cloaked |
| **Attempts to mimic artist A** | | | | | | | | |
| **Attempts to mimic artist B** | | | | | | | | |
| **Artist-rated PSR** | 4.3 ± 0.2% | 93.5 + 0.6% | 91.3 + 0.5% | 90.2 + 0.8% | 4.3 ± 0.2% | 87.2 ± 1.1% | 90.1 ± 0.8% | 91.5 ± 0.9% |
| **CLIP-based genre shift** | 1.4 ± 0.2% | 96.0 ± 0.3% | 94.8 ± 0.4% | 94.0 ± 0.4% | 1.4 ± 0.2% | 90.3 ± 0.8% | 93.8 ± 0.4% | 94.7 ± 0.3% |

# Countermeasures

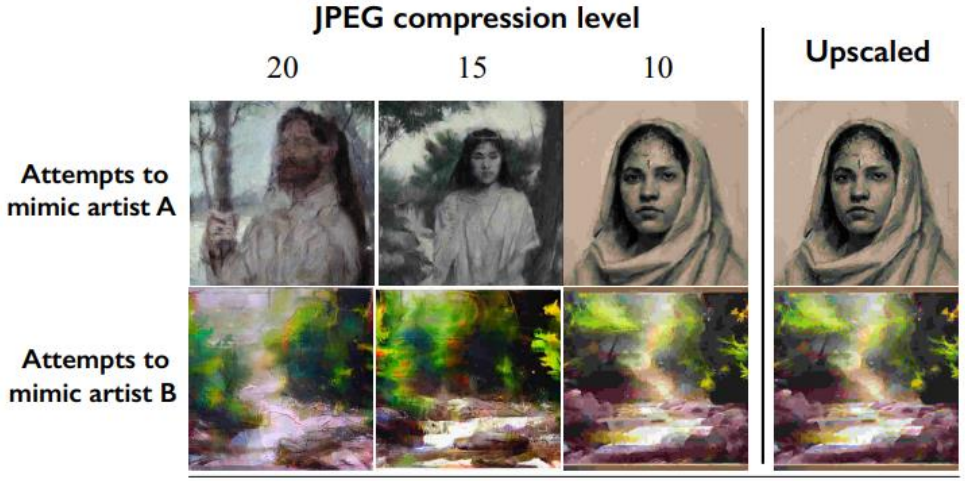
## Image preprocessing

Finetuning models on images with artificial noise/compression and denoising/decompressing them afterwards gave better mimcry results

Glaze achieved 85% score of artist based PSR




| | σ = 0.05 | σ = 0.1 | σ = 0.15 | Denoised |
|---|---|---|---|---|
| Artist-rated PSR | 92.9 ± 0.5% | 91.2 ± 0.7% | 91.6 ± 0.5% | 89.3 ± 1.2% |




| | 20 | 15 | 10 | Upscaled |
|---|---|---|---|---|
| Artist-rated PSR | 93.4 ± 0.8% | 92.3 ± 0.6% | 87.4 ± 0.9% | 85.3 ± 1.3% |

# Countermeasures

## Robust pretraining

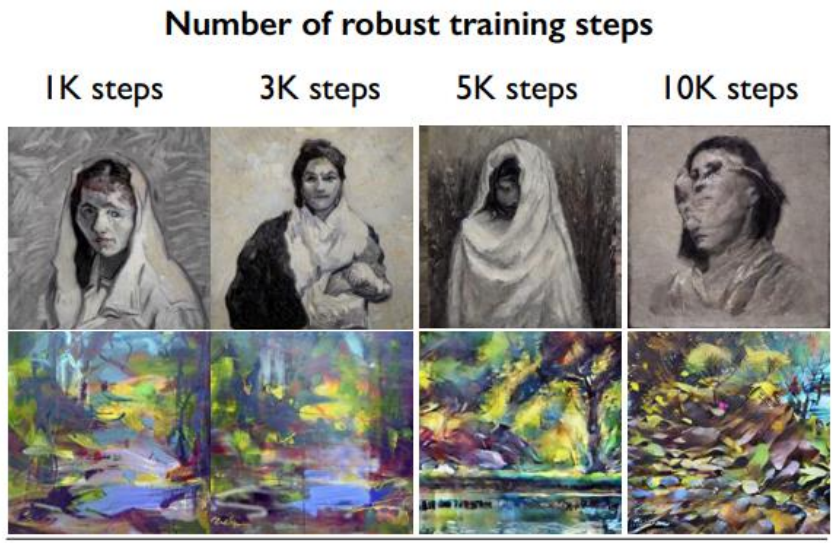
Model is fine-tuned on cloaked images with correct caption

Artists' PSR score remains >88%

## Outlier detection

Cloaked images are not used in fine-tuning

Outlier detectors have limited effectivenes against *Glaze*




| | 1K steps | 3K steps | 5K steps | 10K steps |
|---|---|---|---|---|
| Artist-rated PSR | 92.2 ± 0.8% | 89.3 ± 1.3% | 91.3 ± 0.9% | 95.3 ± 0.3% |

# Real world limitations

Cloaking artwork may take time and require better hardware

Artworks uploaded to the Internet cannot be cloaked

Will glaze be future proof?

# Other mimicing techniques

After release of Glaze there were already few attempts to bypass the style cloaks

**PEZ** – style mimicry using single image
reverse engeneering image into text prompt and then regenerating it

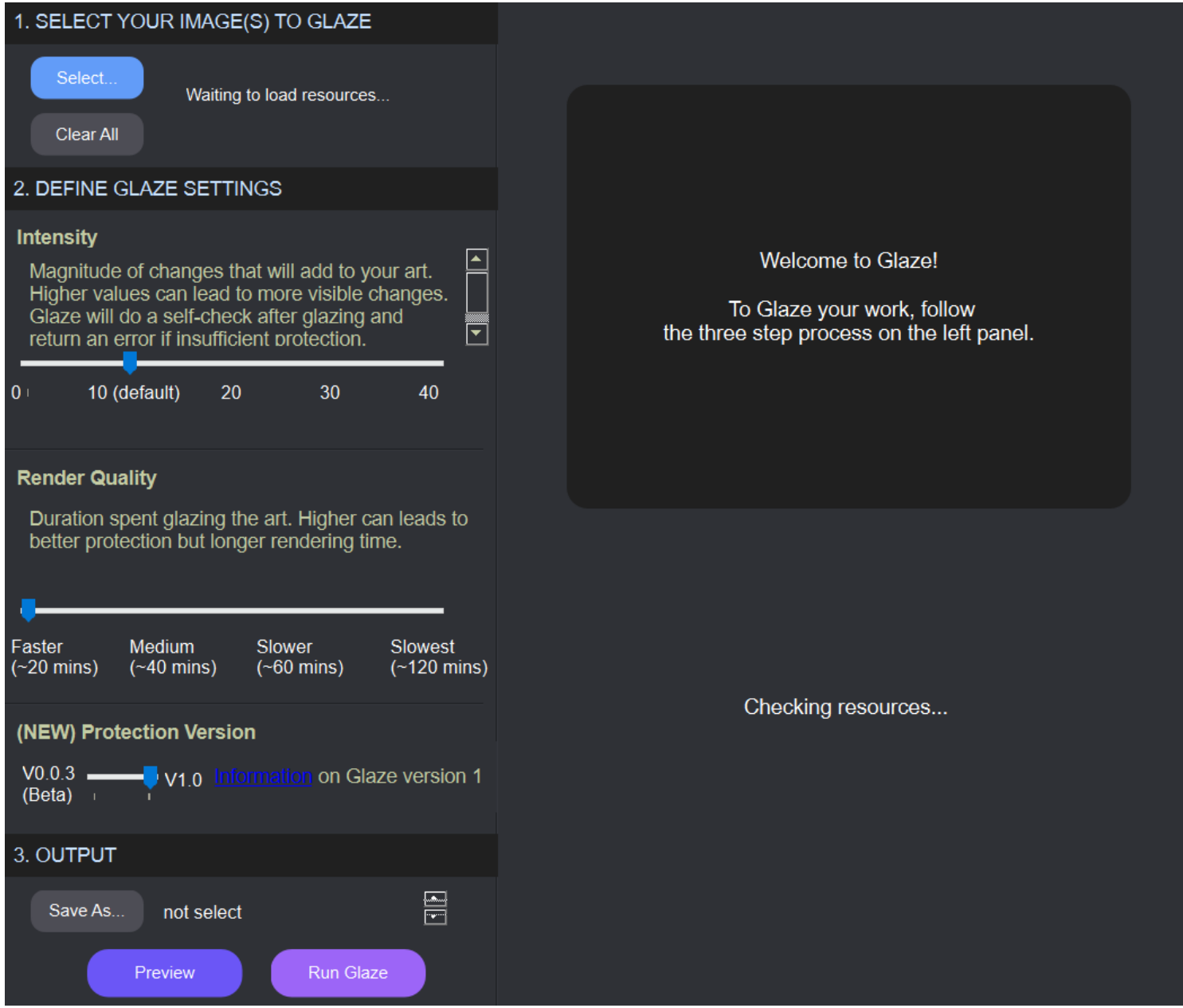
**Pixel smoothing** – repeatedly applies bilateral filters on the image to remove cloaks

# Usage

https://glaze.cs.uchicago.edu/

Authors released an application *glaze* – it was downloaded >740 times since Feb 2023

*Webglaze* – online glaze, invitation needed



# Thanks for your attention!