Cognitive Computing and Internet of Things, Vol. 165, 2025, 11–20
https://doi.org/10.54941/ahfe1006086

Evaluating Glaze's Effectiveness: A Critical Analysis of AI Art Protection Through Non-Artist Perspectives and Common Image Transformations

Robert Nasyrov, Janina Bach, and Pascal Laube

Furtwangen University, Furtwangen, Germany

ABSTRACT

This paper presents a systematic evaluation of Glaze 2.1, a tool designed to protect artists' styles from AI mimicry. We examine its effectiveness against common image transformations typically applied by social media platforms and assess its protection capabilities through the perspective of non-artist users. Our methodology combines technical analysis of how transformations like JPEG compression, scaling, blurring, and sharpening affect Glaze's protective perturbations with a comprehensive user study involving participants without specific artistic expertise. Results indicate that Glaze exhibits significant vulnerabilities when protected images undergo standard social media processing, with certain transformations substantially reducing its effectiveness. These findings highlight the challenges in developing robust protection mechanisms that can withstand real-world usage scenarios while remaining practical for artists. We contribute valuable insights into the limitations of current AI art protection tools and suggest directions for developing more resilient solutions that can better safeguard artists' intellectual property in digital environments.

Keywords: Adversarial machine learning, AI art protection, Style mimicry, Image transformations, Glaze, Social media processing, User perception, Digital rights, Generative AI, Intellectual property

INTRODUCTION

The rapid advancement of text-to-image generative AI models has raised significant concerns about the protection of artists' intellectual property. While these models enable unprecedented creative possibilities, they also allow for the unauthorized replication of artists' unique styles through fine-tuning techniques like DreamBooth (Ruiz, Nataniel et al., 2023). To counter this threat, protection mechanisms like Glaze (Shan et al., 2023) have emerged, which add adversarial perturbations to artwork to prevent AI models from learning artists' styles. For many artists, these concerns extend beyond economic implications to fundamental questions about creative ownership, as AI systems can now generate entire portfolios in minutes that mimic their distinctive styles, potentially devaluing years of artistic development and threatening their position in an increasingly competitive marketplace.

Recent work by Hönig et al. (2025) challenged the effectiveness of such protection tools, demonstrating vulnerabilities to simple image processing techniques. However, this sparked controversy, with the Glaze team arguing that "security is an ongoing battle" and that protection tools remain valuable even if imperfect, as artists understand the need for continuous updates against new attacks[1].

Our work provides a systematic evaluation of Glaze 2.1's effectiveness in real-world scenarios, focusing on three critical aspects:

- Common Image Transformations: We examine how social media platforms' standard image processing operations (JPEG compression, scaling, blurring, and sharpening) affect Glaze's protective capabilities.
- Non-Expert Evaluation: We evaluate effectiveness through participants without specific selection criteria regarding artistic background or expertise, providing a more realistic assessment of how the general public perceives AI-generated art. Participants were asked about their demographic information, frequency of social media use, and prior knowledge of art and AI to contextualize results and analyze potential correlations with their evaluations.
- Practical Protection Assessment: We test Glaze 2.1 against fine-tuned models trained exclusively with DreamBooth on Stable Diffusion 1.5, simulating realistic attack scenarios.

Research Questions. Our study examines how robust Glaze 2.1 is against common image transformations applied by social media platforms and evaluates the effectiveness of the protection mechanism from the perspective of regular social media users.

Our contributions include a comprehensive evaluation of Glaze 2.1's robustness against real-world image transformations, the first larger-scale user study focusing on digitally literate participants' perception of protection effectiveness, and practical insights into the limitations and capabilities of current AI art protection mechanisms. Our results indicate that Glaze shows significant vulnerabilities when images undergo common transformations, particularly those automatically applied by social media platforms. This suggests a need for more robust protection mechanisms that can withstand real-world usage scenarios while remaining practical for artists to implement.

STYLE PROTECTION MECHANISMS

The proliferation of generative AI has raised significant concerns regarding unauthorized style replication, prompting the development of various protection tools. Glaze (Shan et al., 2023) adds imperceptible adversarial perturbations to artwork to confuse AI models during training (see Figure 1). Under ideal conditions, Glaze reportedly disrupts AI-driven style imitation with success rates exceeding 92%, maintaining effectiveness above 85% even against countermeasures. However, these claims have been challenged by subsequent research. Hönig et al. (2025) demonstrated that protection

[1] https://glaze.cs.uchicago.edu/update21.html

mechanisms like Glaze might be circumvented with simple image processing techniques. Alternative approaches include Anti-DreamBooth (Radiya-Dixit and Tramèr, 2021), IMPASTO (Guo et al., 2024), LAACA (Li et al., 2024), and Neural Style Protection (NSP) (Passananti et al., 2024). Unlike typical adversarial attacks aimed at misclassification, these protection methods disrupt feature extraction during model training, preventing accurate style imitation. Recent research has expanded beyond basic perturbation techniques to include color-based protection (Li et al., 2024), perception-aware manipulation (Guo et al., 2024), and specialized approaches for video content (Kim and Woo, 2024). A key limitation is the asymmetry between defenders and attackers. As Hönig et al. (2025) highlight, artists must apply protection preemptively, but once images are downloaded, the protection remains static. Their study demonstrates that common image transformations can significantly weaken adversarial perturbations. Additionally, Kim and Woo (2024) introduced GLEAN, a GAN-based approach that effectively removes these protections.

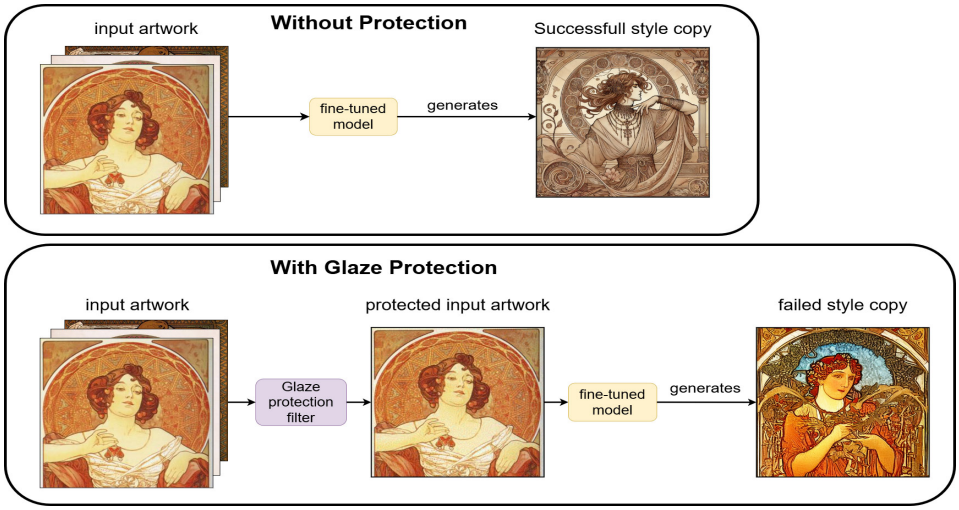

Figure 1: Process of art style replication without protection (top) and protected by glaze protection filter (bottom).

Social media platforms consistently apply image transformations, with JPEG compression being the most prevalent. Studies indicate that 9 out of 10 platforms use JPEG compression at varying intensities—Facebook employs quality factors between 71–92, while Instagram applies a more aggressive 50% compression (Moltisanti et al., 2015). Downscaling is the second most common transformation, typically triggered when images exceed platform-specific size limits; Facebook automatically reduces images larger than 2048 pixels (Castiglione, Cattaneo and De Santis, 2011; Verde et al., 2021). These transformations significantly affect image quality, with resolution scaling having a greater impact than luminance and chrominance adjustments (Laghari et al., 2018). Beyond compression and downscaling, additional processing techniques may influence protective perturbations.

Facebook applies 'enhancement filtering' to improve image appearance (Castiglione, Cattaneo and De Santis, 2011), which could interfere with adversarial defenses. Although explicit mentions of blurring and sharpening are scarce (Sun et al., 2018), these operations remain relevant as common modifications that might circumvent protection mechanisms.

METHODOLOGY

In this section, we detail our experimental design for evaluating Glaze 2.1's effectiveness against style imitation attacks, including our dataset preparation, model training approach, image processing pipeline, and user study methodology.

Dataset and Artist Selection

We selected four classic artists representing diverse styles: Van Gogh, Dürer, Rousseau, and Mucha. Images for the first three came from the "Best Artworks of All Time" Kaggle[2] dataset, while Mucha's were from WikiArt[3]. We used public domain artworks rather than contemporary artists' work, as Hönig et al. (2025) found no significant difference in Glaze's effectiveness between historical and modern artists, despite criticism from Glaze developers about studies focusing exclusively on historical works.

For each artist, we selected 20 images to simulate a style imitation attack. We standardized the dataset by selecting similar motifs and techniques per artist (e.g., only sketches for Dürer) and resizing all images to 512×512 px to match Stable Diffusion 1.5's default input resolution. Images were center-cropped when necessary to maintain focus on the main motif and preserve stylistic integrity.

Model Architecture and Training

For our experiments, we utilized Stable Diffusion 1.5 as the base model for all fine-tuning operations, using the DreamBooth implementation from Stable Diffusion Art. This model was selected due to its widespread adoption in the AI art community and its demonstrated capability for high-quality style transfer. For each artist, we created five model variants to test different scenarios. This setup simulates realistic attack scenarios while maintaining consistent training conditions across all experiments.

All images were rescaled to 512×512 px to match the standard training resolution of Stable Diffusion. For fine-tuning, mixed-precision fp16 was used to optimize computational efficiency and reduce both memory requirements and training time. Training was performed for 300 steps with a learning rate of $5e^{-6}$ using AdamW optimization. To ensure methodological consistency, fixed random values were used for all training runs, allowing a controlled assessment of image quality differences between model variants. Instance prompts and class prompts were defined for each artist to ensure adherence to style while maintaining generalization.

[2]https://www.kaggle.com/datasets/ikarus777/best-artworks-of-all-time

[3]https://www.wikiart.org/en/alphonse-mucha/

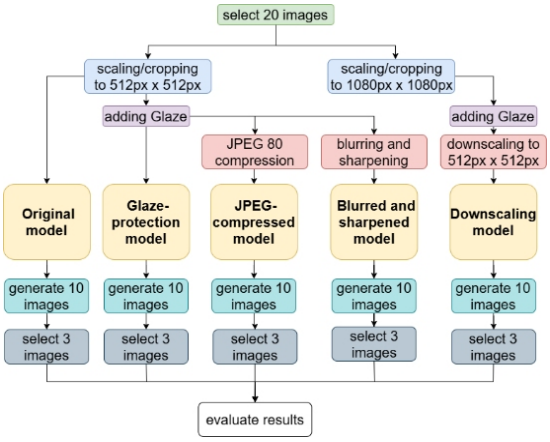

Figure 2: Systematic approach of creating five distinct AI models per artist.

This figure illustrates our systematic approach to creating five distinct AI models per artist from a consistent set of 20 images. For each artist, we developed: 1) an "original model" trained on unmodified images as our baseline; 2) a "glaze-protection model" using images processed with Glaze 2.1 at maximum perturbation intensity; 3) a "JPEG-compressed model" where Glaze-protected images underwent 80% JPEG compression to simulate social media platforms; 4) a "blurred and sharpened model" where Glaze-protected images were processed with a 5×5 Gaussian blur followed by sharpening; and 5) a "downscaling model" where images were first Glaze-protected at 1080×1080 px before being downscaled to 512×512 px.

To ensure methodological rigor, we implemented consistent seed values across all image generation processes, allowing us to attribute output variations exclusively to the different pre-processing techniques rather than random initialization differences. We observed significant quality variations between models, necessitating an adaptive generation strategy. The original model consistently produced high-quality outputs with minimal generations, while models trained on Glaze-protected images exhibited greater variance in output quality. For models producing inconsistent results, we generated up to 25 images per prompt and manually selected the 3 highest quality examples for inclusion in our user study. This selection process was critical to ensure that our comparative analysis evaluated the effectiveness of protection mechanisms rather than being confounded by general quality differences between model variants. All generated images were subsequently evaluated through our user survey to assess the effectiveness of each protection approach.

1) Original model. Exclusively trained on unmodified (only uniform scaling) images and provides a baseline for comparison to evaluate whether differences in the image quality produced are due to glaze protection or general differences in model quality.

2) Glaze-protection model. In this model, images with Glaze 2.1 are used with the highest perturbation intensity and the longest rendering time to

introduce negative perturbations before training. This model serves as a reference for evaluating Glaze's ability to prevent AI models from learning an artist's style when no additional transformations are applied.

3) **JPEG-compressed model.** Glaze-protected images were compressed to 80% JPEG quality to simulate the default compression of images on social media platforms like Instagram. This compression value was determined through empirical testing of Instagram uploads. The resolution was adjusted to 1080 × 1080 px, as Instagram supports this square format, with the platform automatically applying approximately 80% JPEG compression during upload.

4) **Blurred and sharpened mode**l. This model is used to test whether blurring and sharpening can remove Glaze perturbations from an image. A Gaussian blur B was applied to the glaze-protected images to smooth out the adversarial noise. After blurring, a sharpening filter S was applied. The transformation filters are enumerated in Equation (1).

$$S = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 4 & 7 & 4 & 1 \\ 4 & 16 & 26 & 16 & 4 \\ 7 & 26 & 41 & 26 & 7 \\ 4 & 16 & 26 & 16 & 4 \\ 1 & 4 & 7 & 4 & 1 \end{bmatrix} \tag{1}$$

5) **Downscaling model.** This model examines the effects of scaling on the robustness of Glaze's protection mechanism. The images were initially processed with Glaze at a resolution of 1080×1080 px before being scaled down to 512×512 px using bilinear interpolation prior to fine-tuning. This scaling simulates the usual downscaling of images by social media platforms such as Instagram, which automatically adjust the resolution of uploaded images. By evaluating whether resizing reduces the protective effect of Glaze, this model addresses a critical aspect of preventing style imitation, as datasets used for fine-tuning generative models often standardize image sizes to ensure consistency.

The input images were processed with Glaze 2.1, applying the highest perturbation intensity and the longest rendering time.

User Study Design

Our user study examined whether participants perceive AI-generated images as authentic representations of artists' styles. We recruited participants without specific artistic expertise requirements to approximate general public perception. Participants provided demographic information, including age, social media usage, and prior knowledge of art and AI, to contextualize results and analyze correlations with their evaluations. The study included 90 participants, with 76% reporting frequent social media use (>6 hours/week) and only 4% using it rarely (<1 hour/week or not at all).

Study Design and Materials. We evaluated 15 AI-generated images per artist (3 images × 5 model conditions) across four artists, totalling 60 images. Participants were introduced to the concept of artifacts—visual anomalies

in AI-generated images—and completed a comprehension check to ensure understanding before proceeding. To establish baseline style recognition, participants viewed nine original artworks per artist during a familiarization phase, enabling them to develop sufficient understanding of each artist's characteristic style prior to evaluation.

Evaluation Procedure. Participants were presented with 15 AI-generated images per artist, created under different model conditions (see Figure 2). They were instructed to select all images that they felt best represented the artist's style. For unselected images, participants had to justify their decision by choosing from a predefined list:

- Logical errors (e.g., anatomical errors, distorted objects)
- Unusual choice of color
- Inappropriate motif
- Visible artifacts or noise
- Other (free-text response option).

To avoid bias, participants could reject all presented images if they felt none sufficiently represented the artist's style. Original artworks were excluded from the selection pool to prevent direct comparisons, which could have led to unrealistic evaluation scenarios not reflective of real-world contexts, such as encountering AI-generated images on social media platforms.

Participant Recruitment and Filtering. Participants were recruited via Prolific, an academic research platform, with no expertise criteria in art or AI. From 143 initial participants, we included only those who completed the survey on desktop or laptop computers, resulting in 90 participants (62.94%) for final analysis. This approach ensured consistent image viewing conditions and high data quality.

RESULTS

In this section, we present the results of our study on the robustness of Glaze 2.1 against real-world image transformations and its effectiveness in preventing AI-driven style imitation. We analyze the selection rates of AI-generated images across experimental conditions and provide insights into user perceptions based on our study.

Figure 3 shows selection rates for each artist under different conditions. Unprotected original images had the highest selection rates, while Glaze-protected images had the lowest. Common transformations (JPEG compression, blurring, downscaling) significantly increased selection rates of protected images. For Van Gogh, JPEG-compressed images even exceeded original images' selection rate (68.5% vs. 54.4%). Figure 4 illustrates overall selection rates across models. Original images achieved the highest rate (57.4%), while Glaze-protected images showed the lowest (26.5%). Transformations like JPEG compression (45.5%) and downscaling (42.8%) substantially increased selection rates of protected images.

In a small percentage of cases (3.3%–7.8% across artists), participants did not select any of the generated images. This effect was most pronounced for Glaze-protected images.

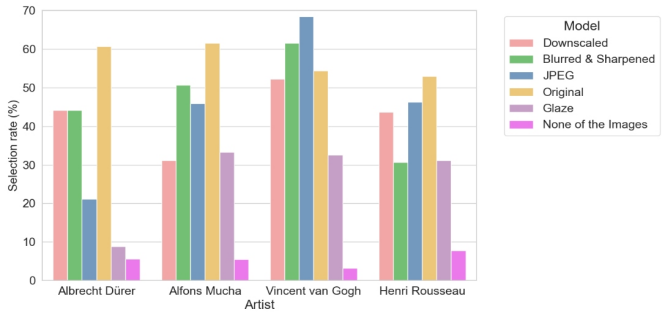

Figure 3: Selection rate per model and artist.

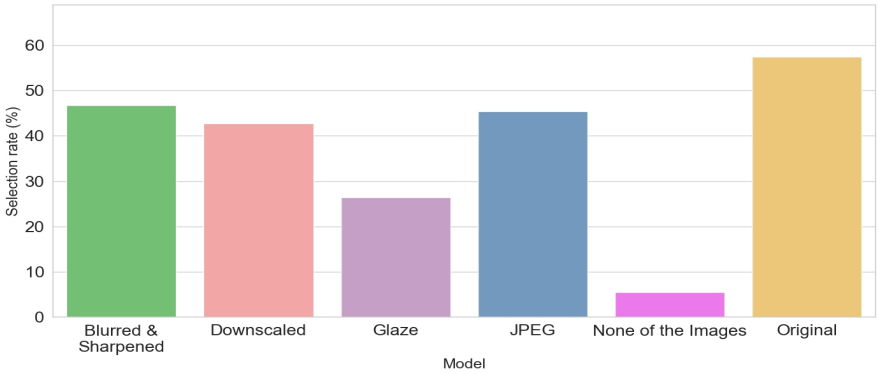

Figure 4: Average selection rate per model.

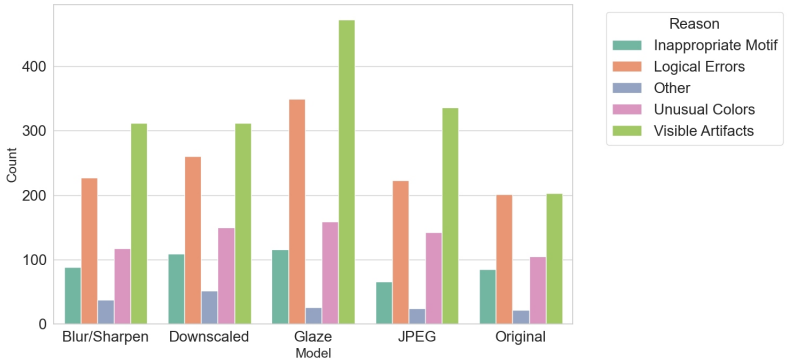

Figure 5: Rejection reasons per model.

As depicted in Figure 5, visible artifacts were the most frequently cited reason for image rejection across all models. While Glaze-protected images had the highest rejection rate due to artifacts, other models such as Downscaled and JPEG-compressed also received significant rejections for the same reason. Additionally, logical errors—such as anatomical distortions or implausible compositions—constituted an important rejection factor, particularly in JPEG-compressed images. Even images from the original

model were not exempt from rejection, as some participants found certain motifs or color variations inconsistent with the expected artistic style.

Analysis of Protection Effectiveness

Results show style-dependent Glaze effectiveness variations—Van Gogh's impasto particularly vulnerable to JPEG compression (68.5% vs. 54.4% selection rate for originals). Textural styles with high-contrast colors are more susceptible than detailed linework, suggesting the need for style-specific calibration. Selection rates increased dramatically from Glaze-protected images (26.5%) to transformed versions (JPEG: 45.5%, downscaling: 42.8%, blurring: 38.2%), revealing critical vulnerability to common transformations.

Artifacts were the primary reason for image rejection across models, with Glaze-protected images showing a higher rate of artifact-related rejections—indicating a trade-off between protection and image quality. Current protections offer limited defense in typical online contexts. JPEG compression most significantly compromised protection by preserving low-frequency components while discarding high-frequency adversarial signals, followed by downscaling, with blurring showing least impact.

Our findings reveal a fundamental asymmetry: artists using defensive measures face inherent disadvantages against circumvention attempts. The brittleness of current adversarial perturbation tools against simple transformations represents a structural challenge, not merely a technical limitation. The research community must evaluate protection mechanisms for transformation robustness and explore alternatives beyond adversarial perturbations.

CONCLUSION

Our study systematically evaluates Glaze 2.1's robustness against real-world image transformations and effectiveness in preventing AI-driven style mimicry. Testing multiple fine-tuned models with 90 participants revealed that while Glaze reduces style imitation, its protection is vulnerable to common transformations (JPEG compression, blurring, downscaling). These findings demonstrate the need for more robust, distortion-resistant protection mechanisms to safeguard artists from unauthorized style extraction. While Glaze 2.1 represents progress in protecting artistic styles, its vulnerability to routine image alterations highlights the necessity for stronger, more adaptive protection approaches.

REFERENCES

Castiglione, A., Cattaneo, G. and De Santis, A. (2011) 'A Forensic Analysis of Images on Online Social Networks', in 2011 Third International Conference on Intelligent Networking and Collaborative Systems, pp. 679–684.

Guo, Zhongliang et al. (2024) 'Artwork Protection Against Neural Style Transfer Using Locally Adaptive Adversarial Color Attack', in ECAI 2024. IOS Press, pp. 1414–1421.

Hönig, Robert, et al. (2024) "Adversarial perturbations cannot reliably protect artists from generative ai." arXiv preprint arXiv:2406.12027.
Kim, J. L. and Woo, K. (2024) 'GLEAN: Generative Learning for Eliminating Adversarial Noise'. arXiv.
Laghari, A. A. et al. (2018) 'Assessment of quality of experience (QoE) of image compression in social cloud computing', Multiagent and Grid Systems, 14(2), pp. 125–143.
Li, Y. et al. (2024) 'Neural Style Protection: Counteracting Unauthorized Neural Style Transfer', in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3966–3975.
Moltisanti, M. et al. (2015) 'Image Manipulation on Facebook for Forensics Evidence', in V. Murino and E. Puppo (eds) Image Analysis and Processing—ICIAP 2015. Cham: Springer International Publishing, pp. 506–517.
Passananti, J. et al. (2024) 'Disrupting Style Mimicry Attacks on Video Imagery'. arXiv.
Radiya-Dixit, E. and Tramèr, F. (2021) 'Data Poisoning Won't Save You From Facial Recognition', ArXiv.
Ruiz, Nataniel, et al. (2023) "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.
Shan, S. et al. (2023) 'Glaze: Protecting artists from style mimicry by {Text-to-Image} models', in 32nd USENIX Security Symposium (USENIX Security 23), pp. 2187–2204.
Sun, W. et al. (2018) 'Robust Privacy-Preserving Image Sharing over Online Social Networks (OSNs)', ACM Trans. Multimedia Comput. Commun. Appl., 14(1), pp. 14:1–14:22.
Verde, S. et al. (2021) 'Multi-clue reconstruction of sharing chains for social media images'. arXiv.