High-Frequency Anti-DreamBooth: Robust Defense against Personalized Image Synthesis

Takuto Onikubo and Yusuke Matsui

The University of Tokyo

Abstract. Recently, text-to-image generative models have been misused to create unauthorized malicious images of individuals, posing a growing social problem. Previous solutions, such as Anti-DreamBooth, add adversarial noise to images to protect them from being used as training data for malicious generation. However, we found that the adversarial noise can be removed by adversarial purification methods such as DiffPure. Therefore, we propose a new adversarial attack method that adds strong perturbation on the high-frequency areas of images to make it more robust to adversarial purification. Our experiment showed that the adversarial images retained noise even after adversarial purification, hindering malicious image generation.

Keywords: Fake Image · Image Generative Model · Adversarial Attack

1 Introduction

Recently, the progress of text-to-image generative models has been remarkable. One of the practical applications of these models is personalization, a fine-tuning method to generate images of a specific subject with a few examples. However, some people personalize models to generate unauthorized malicious images called fake images, posing a significant social problem. For example, some people download photos of celebrities from the internet and fine-tune generative models with these photos to generate scandalous images [2]. Furthermore, some painters are losing their jobs due to the generation of artworks that mimic their techniques [1].

To address the issue of fake images, Thanh et al. [20] proposed an adversarial attack named Anti-DreamBooth, a method that prevents personalized generation. We can use Anti-DreamBooth to add slight noise perturbation to our images before publishing them, hindering any unauthorized personalized generations. Additionally, Shan et al. [17] proposed Glaze to prevent the generation of images that mimic paintings created by artists. Although these methods are very powerful, we have found that the added perturbations can be removed easily by adversarial purification techniques such as DiffPure [12]. We also found they are vulnerable to simple noise-removal filters like Gaussian or bilateral filters.

To address the problem, we propose an adversarial attack that adds strong adversarial noise on high-frequency areas of images. Our method utilizes a high-pass filter to create a mask that represents the edges of the image and adds

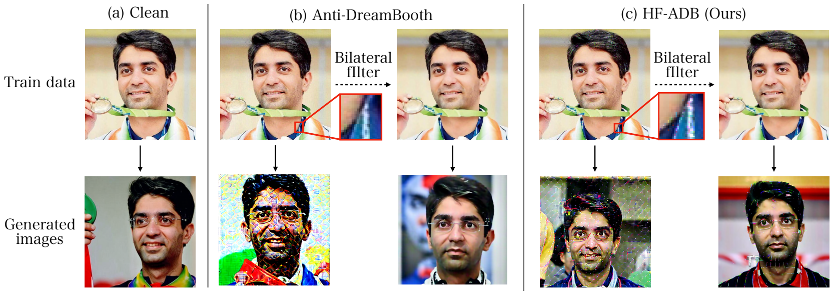

Fig. 1: Summary: (a) We can generate realistic images by personalizing a model. (b) Applying Anti-DreamBooth [20] to the train images hinders personalized generation. However, we can break the defense by noise removal methods such as bilateral filters. (c) Our method maintains its defense even after applying noise removal methods.

especially strong adversarial noise to the masked area. The intense adversarial noise in the complicated region is difficult to remove despite being inconspicuous.

The overview of our work is shown in Fig. 1 with the following contributions.

- We found that the adversarial noise previous works add can be removed easily by adversarial purification technique.
- By adding strong perturbation to the high-frequency areas on images, we propose a robust adversarial attack against adversarial purification.
- As our idea to vary the strengths of noise does not depend on a specific adversarial attack, we can apply the concept to various adversarial attacks.

2 Related Work

2.1 Adversarial Attack and Adversarial Purification

Adversarial attack [5, 8, 18] is a method that causes a model to fail by taking advantage of its learning and operation processes. For example, there are attacks against speech recognition [4, 23] or image recognition [9, 21]. Also, data poisoning [14, 22] is an adversarial attack that prevents the model from being trained adequately by injecting malicious data into the dataset.

As a countermeasure against adversarial attacks, a method called adversarial purification [16] uses generative models to purify the adversarial noise. DiffPure [12] is an adversarial purification method that uses diffusion models. Although computationally expensive, DiffPure is one of the most potent adversarial purification methods.

2.2 Abuse of Text-to-Image Generative Models

Generating fake images or imitating illustrations by personalizing generative models has become a significant problem. In order to handle the issue, adversarial attack-based methods [17, 20] and watermarking-based methods [10, 24] have been mainly studied.

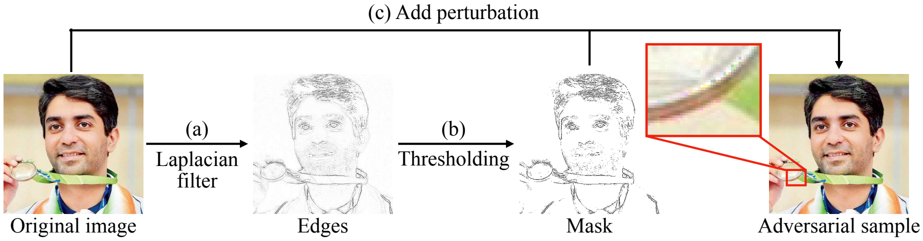

Fig. 2: Procedure steps of our method.

3 Preliminaries

In this section, we describe the problem setting. Suppose a user uploads their images onto the internet. Malicious people might collect the images, personalize pre-trained models, and generate manipulated images (e.g., depicting the user committing a crime). To prevent personalization, we suggest a method for users to protect their images, i.e., adding adversarial perturbations to their images before uploading them to the internet.

We can formulate the problem as follows. First, we denote $\left\{\mathbf{x}^{(i)}\right\}_{i=1}^{N_{\text{db}}} \subset [0,\ 1]^{H\times W\times 3}$ as the images for the user to protect. Here, $N_{\text{db}} \in \mathbb{N}$ is the number of images to upload to the internet. Next, we optimize perturbations $\left\{\boldsymbol{\delta}^{(i)}\right\}_{i=1}^{N_{\text{db}}} \subset [0,\ 1]^{H\times W\times 3}$. We then construct adversarial examples by adding the perturbations to the images: $\left\{\mathbf{x}^{(i)} + \boldsymbol{\delta}^{(i)}\right\}_{i=1}^{N_{\text{db}}}$. Our task is to maximize the output of the loss function $\mathcal{L}_{\text{db}}$ during fine-tuning using these adversarial examples so that the fine-tuned model fails to generate the user's images.

4 Method

In this section, we propose a method to hinder personalized image generation using DreamBooth [15] by adding strong adversarial perturbation to the high-frequency areas of the images. Our method adds perturbation that is difficult to remove using traditional filters or adversarial purification techniques.

Fig. 2 illustrates the workflow of our method. First, we apply a 3×3 Laplacian filter to the input image and extract the edges of the image (Fig. 2(a)). Then, by thresholding the pixel values of the edges, we obtain a mask of the edges (Fig. 2(b)). This mask shows the area with high frequency. The threshold is determined for each image so that the ratio of the masked area is constant (in practice, 3-5% of the entire image). Next, we add adversarial perturbation using the created mask (Fig. 2(c)). As shown in the red box area, strong perturbation is added only to the edges, while weak perturbation is added to the entire image.

Here, we used ASPL (Alternating Surrogate and Perturbation Learning) from Anti-DreamBooth [20] to calculate the noise. Our goal is to generate adversarial

Algorithm 1 Algorithm of our method

```
Require: $\left\{\mathbf{x}^{(i)}\right\}_{i=1}^{N_{\mathrm{db}}} \subset [0,\ 1]^{H\times W\times 3}$ : original images,
         $\left\{\mathbf{m}^{(i)}\right\}_{i=1}^{N_{\mathrm{db}}} \subset [0,\ 1]^{H\times W\times 3}$ : masks,
         $\mathcal{X} \subset [0,\ 1]^{H\times W\times 3}$ : reference images for the surrogate model fine-tuning,
         $\eta \in [0,\ 1]$ : maximum magnitude of the weak perturbation,
         $\eta_{\mathrm{mask}} \in [\eta,\ 1]$ : maximum magnitude of the strong perturbation,
         $\eta_{\mathrm{unit}} \in [0,\ \eta]$ : the amount of perturbation added in one step,
         $\boldsymbol{\theta}_{\mathrm{pre}}$ : pre-trained parameter of Stable Diffusion,
         $T$ : the number of steps to add perturbation,
Ensure: $\left\{\boldsymbol{\delta}_T^{(i)}\right\}_{i=1}^{N_{\mathrm{db}}} \subset [0,\ 1]^{H\times W\times 3}$: adversarial perturbation.
1: $\boldsymbol{\delta}_0^{(i)} \leftarrow \mathbf{0}$ (for all $i$) : initialization of noise $\boldsymbol{\delta}$
2: $\boldsymbol{\theta}_{\mathrm{sur}} \leftarrow \boldsymbol{\theta}_{\mathrm{pre}}$ : initialization of the parameters of the surrogate model
3: for $t \in \{1, \ldots, T\}$ do
4:     $\boldsymbol{\theta}_{\mathrm{tmp}} \leftarrow \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{\mathbf{x}_{\mathrm{ref}} \in \mathcal{X}} \mathcal{L}_{\mathrm{db}}(\boldsymbol{\theta},\ \mathbf{x}_{\mathrm{ref}})$ (with $\boldsymbol{\theta}_{\mathrm{sur}}$ as the initial weight)
5:     $\boldsymbol{\delta}^{(i)} \leftarrow \underset{\boldsymbol{\delta} \in [0,\ \eta_{\mathrm{unit}}]^{H\times W\times 3}}{\operatorname{argmax}} \mathcal{L}_{\mathrm{cond}}(\boldsymbol{\theta}_{\mathrm{tmp}},\ \mathbf{x}^{(i)} + \boldsymbol{\delta}_{t-1}^{(i)} + \boldsymbol{\delta})$ (for all $i$)
6:     $\boldsymbol{\delta}_t^{(i)} \leftarrow \operatorname{clamp}\left(\boldsymbol{\delta}_{t-1}^{(i)} + \boldsymbol{\delta}^{(i)},\ \mathbf{m}^{(i)},\ \eta,\ \eta_{\mathrm{mask}}\right)$ (for all $i$)
7:     $\boldsymbol{\theta}_{\mathrm{sur}} \leftarrow \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{i=1}^{N_{\mathrm{db}}} \mathcal{L}_{\mathrm{db}}(\boldsymbol{\theta},\ \mathbf{x}^{(i)} + \boldsymbol{\delta}_t^{(i)})$
8: end for
```

examples that degrade the performance of a model fine-tuned on these images. This is too complicated to solve. Therefore, ASPL exploits a model called surrogate model $\boldsymbol{\theta}_{\mathrm{sur}}$, an approximation of a model personalized with adversarial examples. ASPL trains a surrogate model using the user's reference images $\mathcal{X}$ and simultaneously calculates adversarial perturbations that degrade the performance of the surrogate model.

We can formulate the process of adding perturbation as Alg. 1. Firstly, we initialize the adversarial noise and the surrogate model parameters in L1 and L2. Then, we add adversarial noise over T steps in L3–8. Specifically, in L4, the surrogate model $\boldsymbol{\theta}_{\mathrm{sur}}$ is fine-tuned for one step, resulting in the model $\boldsymbol{\theta}_{\mathrm{tmp}}$. Here, $\mathcal{L}_{\mathrm{db}}$ is the loss function of DreamBooth [15]. Next, in L5, perturbation is added to each pixel of the image by η_{unit} so that the inference loss of the model $\boldsymbol{\theta}_{\mathrm{tmp}}$ is maximized. Here, $\mathcal{L}_{\mathrm{cond}}$ is the loss function of a typical diffusion model. Then, in L6, adversarial perturbation is clamped so that the perturbation on the masked area is less than η_{mask}, and otherwise less than η. This part is the core of our method. Through this process, we can create adversarial examples with strong adversarial perturbations at the edges of the image. Finally, in L7, the surrogate model is fine-tuned for one step using the updated adversarial examples. We repeat these steps over T steps.

5 Experiment

5.1 Dataset

In this study, we use VGGFace2 [3] as the dataset. VGGFace2 is a large-scale dataset for face recognition, containing over 3.3 million face images of more than 9,000 individuals. We sampled 400 face images from 50 individuals from this dataset, with eight images per individual. To eliminate the influence of image quality on the experimental results, we selected images with a resolution of no less than 512 × 512 pixels and resized them to 512 × 512 pixels.

We divided eight samples of each individual into two groups of four images. The first four images were training images $\left\{\mathbf{x}^{(i)}\right\}_{i=1}^{N_{\mathrm{db}}}$ for personalized generation, and the remains were training images $\mathcal{X}$ for the surrogate model. We define the former as the original images and the latter as the reference images.

5.2 Experiment Settings

Firstly, we generated adversarial examples for original images. Here, we changed the strength of the adversarial perturbation $\eta, \eta_{\mathrm{mask}}$ and the ratio of the masked area as the hyper-parameters. Then, we purify the added perturbations. We used bilateral filter [13] and DiffPure [12] for purification. Next, we personalized Stable Diffusion with these images. Finally, we generated images with a prompt of `"a photo of sks person"` with the personalized model. Here, `"sks person"` is a pseudo-class of target individuals used for personalization.

In this experiment, we used Stable Diffusion v2.1 as the Text-to-Image generative model and DreamBooth [15] as the personalization method. Adding adversarial noise to each image took about 2 minutes on an NVIDIA A100 GPU.

To evaluate the experimental results, we used the following metrics. Firstly, FDSR (Face Detection Success Ratio) is the probability of being successfully detected by RetinaFace [7]. Then, ISM (Identity Score Matching) measures the similarity of the two images by comparing the ArcFace [6] features from the images. Lastly, both SER-FIQ (Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness) [19] and BRISQUE(Blind/Referenceless Image Spatial Quality Evaluator) [11] measure the quality of images.

5.3 Discussion

Fig. 3 shows the generated images in this experiment. Firstly, the first and second rows show that clean images are generated by the model personalized with Anti-DreamBooth's adversarial examples after purifying. In other words, the perturbations of adversarial examples are easily removed and lose their defensive capacity. On the other hand, in our method, the third and fourth rows show that generated images exhibit unnatural patterns on the faces, indicating successful defense. These results suggest that our method is effective in preventing personalized generation. Furthermore, our method adds adversarial perturbations that cannot be removed by adversarial purification techniques.

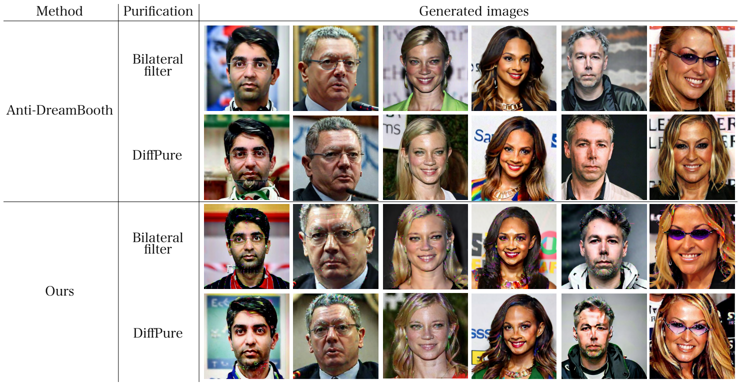

Fig. 3: Generated images: Noise budgets were $\eta = 0.02$ for Anti-DreamBooth, and $\eta = 0.01, \eta_{\text{mask}} = 0.5$ for ours, and the masked area was 3% of images. In this setting, the L_1 norms of the adversarial examples for the clean images are almost the same.

Table 1: Comparison between HF-ADB (Ours) and Anti-DreamBooth (ADB). The arrows indicate the direction in which generative quality deteriorates, meaning the defense is successful. The experimental setting is identical to that in Fig. 3.

Method	Purification	FDSR↓	ISM↓	SER-FIQ↓	BRISQUE↑
ADB	Bilateral filter	0.97 ± 0.10	0.35 ± 0.09	0.76 ± 0.07	42.3 ± 7.7
ADB	DiffPure	0.96 ± 0.10	0.36 ± 0.09	0.76 ± 0.07	10.7 ± 4.3
Ours	Bilateral filter	0.96 ± 0.10	0.35 ± 0.09	0.76 ± 0.08	51.5 ± 8.5
Ours	DiffPure	0.96 ± 0.10	0.36 ± 0.09	0.76 ± 0.08	8.3 ± 3.7

Tab. 1 is the numerical evaluation of the experiment. We couldn't observe a significant difference between those methods. These results suggest that existing metrics cannot assess the authenticity of images, while they can detect significant degradation. Thus, considering more sensitive metrics is part of our future work.

Our method has some limitations. Firstly, the noise we add is so strong that the noise could be noticeable when viewed up close. Secondly, while our method can reduce the effect of adversarial purification, it cannot completely maintain the original defense capacity.

6 Conclusion

In this paper, we propose a method to hinder the generation of face images, specifically addressing the problem that image generative models are misused to create fake images. We found that previous works had an issue that added adversarial perturbation can be removed easily. In contrast, our proposed method achieves a robust adversarial attack against adversarial purification techniques by adding strong adversarial perturbation to the images. We hope that this method will be widely used and help prevent malicious image generation.

References

1. BBC: New ai systems collide with copyright law. https://www.bbc.com/news/business-66231268 (2023), viewed: June 2024 1
2. BBC: Ai can be easily used to make fake election photos - report. https://www.bbc.com/news/world-us-canada-68471253 (2024), viewed: June 2024 1
3. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: IEEE International Conference on Automatic Face & Gesture Recognition(FG). pp. 67–74. IEEE (2018) 5
4. Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., Wagner, D., Zhou, W.: Hidden voice commands. In: 25th USENIX security symposium (USENIX security 16). pp. 513–530 (2016) 2
5. Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., Mukhopadhyay, D.: Adversarial attacks and defences: A survey. arXiv preprint arXiv:1810.00069 (2018) 2
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 4690–4699 (2019) 5
7. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: Retinaface: Single-stage dense face localisation in the wild. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR) (2019) 5
8. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (ICLR) (2014) 2
9. Kong, Z., Guo, J., Li, A., Liu, C.: Physgan: Generating physical-world-resilient adversarial examples for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR). pp. 14254–14263 (2020) 2
10. Liu, H., Sun, Z., Mu, Y.: Countering personalized text-to-image generation with influence watermarks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 12257–12267 (2024) 2
11. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing **21**(12), 4695–4708 (2012) 5
12. Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., Anandkumar, A.: Diffusion models for adversarial purification. In: International Conference on Machine Learning (ICML). pp. 16805–16827. PMLR (2022) 1, 2, 5
13. Paris, S., Kornprobst, P., Tumblin, J., Durand, F., et al.: Bilateral filtering: Theory and applications. Foundations and Trends® in Computer Graphics and Vision **4**(1), 1–73 (2009) 5
14. Rubinstein, B.I., Nelson, B., Huang, L., Joseph, A.D., Lau, S.h., Rao, S., Taft, N., Tygar, J.D.: Antidote: understanding and defending against poisoning of anomaly detectors. In: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement. pp. 1–14 (2009) 2
15. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22500–22510 (2023) 3, 4, 5
16. Samangouei, P., Kabkab, M., Chellappa, R.: Defense-gan: Protecting classifiers against adversarial attacks using generative models. In: International Conference on Learning Representations (ICLR) (2018) 2

17. Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: Glaze: Protecting artists from style mimicry by text-to-image models (2023) 1, 2, 9
18. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations (ICLR) (2014), http://arxiv.org/abs/1312.6199 2
19. Terhorst, P., Kolf, J.N., Damer, N., Kirchbuchner, F., Kuijper, A.: Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5651–5660 (2020) 5
20. Van Le, T., Phung, H., Nguyen, T.H., Dao, Q., Tran, N.N., Tran, A.: Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV). pp. 2116–2127 (2023) 1, 2, 3
21. Wu, H., Yunas, S., Rowlands, S., Ruan, W., Wahlström, J.: Adversarial driving: Attacking end-to-end autonomous driving. In: 2023 IEEE Intelligent Vehicles Symposium (IV). pp. 1–7. IEEE (2023) 2
22. Yang, C., Wu, Q., Li, H., Chen, Y.: Generative poisoning attack method against neural networks. arXiv preprint arXiv:1703.01340 (2017) 2
23. Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., Xu, W.: Dolphinattack: Inaudible voice commands. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. pp. 103–117 (2017) 2
24. Zhu, P., Takahashi, T., Kataoka, H.: Watermark-embedded adversarial examples for copyright protection against diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). pp. 24420–24430 (2024) 2

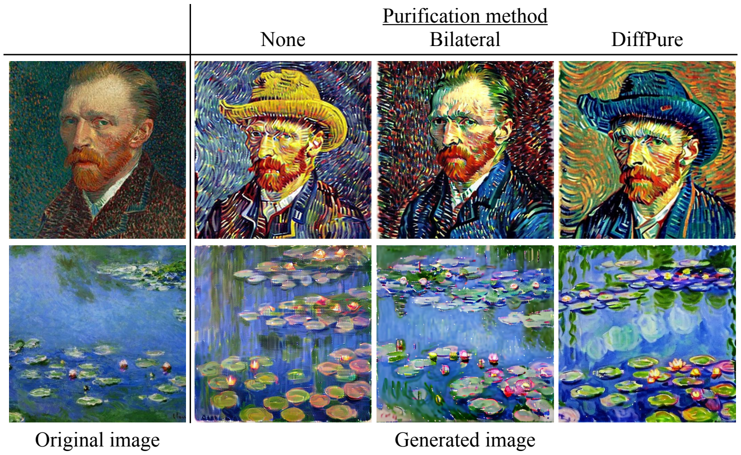

Fig. 4: Artworks generated by a model personalized with our method's adversarial examples. Noise budgets were $\eta = 0.01$ and $\eta_{\text{mask}} = 0.5$, and the masked area was 3%.

A Supplemental Material for Defense against Artwork Generation

Personalized generation of artworks without permission is also a significant problem as well as fake image generation. Here, we applied our method to prevent personalized generation of artworks and analyzed the result.

Fig. 4 shows the result. The first column contains the training images used for personalization, and the second to fourth columns show the results generated after applying different purification methods. We can say that the generated images were relatively clean and the defense failed. Moreover, the generated images look natural even when we do not apply any purification methods. Although the generated images have some unnatural patterns, they are not so noisy. We estimate that the colorful trait of artworks may cause this compared to face images. Therefore, it may be said that style transformation method like Glaze [17], which aims to make personalized models generate different art-style images, is more effective for artwork protection.