VISUAL FIDELITY VS. ROBUSTNESS: TRADE-OFF ANALYSIS OF IMAGE ADVERSARIAL WATERMARK MITIGATED BY SSIM LOSS

Jiwoo Choi **Jinwoo Kim** **Sejong Yang** **Seon Joo Kim**
Yonsei University

ABSTRACT

Adversarial watermark is an important technique for protecting digital images from unauthorized use and illegal AI training. However, conventional methods often introduce visually unpleasant artifacts, making the watermark easily perceptible. This results in an inherent trade-off between robustness and visual fidelity, where stronger protection comes at the cost of degraded image quality. In this work, we address this challenge by integrating SSIM loss into the perturbation embedding process using the Fully-trained Surrogate Model Guidance (FSMG) from baseline. By employing a tunable SSIM weight, our approach balances the adversarial loss—designed to hinder unauthorized model training—with a perceptual loss that preserves image fidelity. Experimental results on CelebA-HQ and VGGFace2, including user studies, show that our method effectively enhances image quality while preserving robustness, as validated by quantitative metrics.

1 INTRODUCTION

The rapid development of personalized generative models (Ruiz et al., 2023; Kumari et al., 2023; Gal et al., 2022) has raised concerns about content theft and unauthorized AI training. For instance, audio generative models (Zhang et al., 2023a; Liu et al., 2023) can be misused in deep-voice crimes by impersonating individuals, while text-to-image and image-to-image models (Rombach et al., 2022; Ramesh et al., 2021; Zhang et al., 2023b) have raised alarms regarding deepfakes and style theft. Furthermore, instances of social media platforms training on user content without consent have exacerbated public apprehension, underscoring the need for mechanisms that protect data integrity.

In this context, adversarial watermark has emerged as a promising approach to empower content creators against unauthorized AI training and content tampering. By embedding subtle perturbations into content, adversarial watermark aims to render such attacks ineffective. However, existing image-cloaking methods that insert perturbations, most notably Anti-DreamBooth (Van Le et al., 2023), are limited by their simplistic perturbation. Specifically, Anti-DreamBooth employs a PGD algorithm (Madry et al., 2018) to maximize a loss function, thereby embedding perturbations into the image. Although effective in deterring misuse, this method often degrades visual quality of the image because the perturbations are directly embedded into key features, significantly reducing imperceptibility. Enhancing robustness typically involves embedding stronger watermarks, which can compromise visual fidelity, while prioritizing imperceptibility may weaken resistance to attacks.

To address this trade-off between image visual fidelity and robustness, our study proposes a approach that integrates the Structural Similarity Index Measure (SSIM) loss (Wang et al., 2004) into the image cloaking framework. By incorporating a perceptual loss term into the overall loss function—controlled by an adjustable SSIM weight λ—we aim to refine the perturbation process. This integration is designed to maintain high visual quality while still ensuring that the watermark remains effective against adversarial attacks.

The key contributions of our study are as follows:

- We introduce a loss formulation that combines the baseline Anti-DreamBooth objective with SSIM loss to provide better visual fidelity, thus enhancing the practical applicability of adversarial watermark in real-world scenarios.

- We perform various experiments across multiple SSIM weights to quantitatively and qualitatively analyze the trade-offs between imperceptibility and resistance to adversarial attacks.

2 Related Work

2.1 AI Watermark for image

AI watermark for image has emerged as a critical technology for protecting digital content from unauthorized use and malicious tampering. Several recent studies have focused on optimizing these perturbations to achieve more effective watermark embedding. They can broadly be divided into two categories: message watermark and non-message watermark.

Message watermark embeds explicit information within the content, enabling the extraction of an embedded message. When the content is subjected to unlawful training, the message can serve as a proof of ownership. Watermark Anything (Sander et al., 2024) and EditGuard (Zhang et al., 2024) employ a watermark localization strategy that adjusts the messages to mask, ensuring that the messages can be extracted even after image editing.

In contrast, non-message watermark does not embed explicit messages. The representative watermark in this category is the adversarial watermark, which disrupts unauthorized model training by embedding subtle perturbations into images, hindering their ability to learn from content. advDM (Liang et al., 2023) focuses on generating adversarial examples through optimized perturbations, while Anti-DreamBooth embeds perturbations by combining a anti-personalization generation framework with a projection gradient descent algorithm (PGD).

2.2 Perceptual Losses for Image Quality

Perceptual loss functions are widely used in image generation and restoration to enhance visual fidelity beyond pixel-wise metrics. The SSIM loss evaluates structural similarity between images, capturing luminance, contrast, and texture differences. It is robust to brightness shifts and widely used in image enhancement. The VGG-based perceptual loss (Johnson et al., 2016) extracts deep feature representations to improve perceptual realism, while adversarial losses (Goodfellow et al., 2020; Ledig et al., 2017) encourage photorealistic synthesis through distribution matching. The LPIPS loss (Zhang et al., 2018) further refines perceptual evaluation by learning feature embeddings aligned with human judgments. In this work, we adopt SSIM loss as a perceptual constraint to ensure high visual quality while embedding watermarks into images, preserving structural integrity and minimizing perceptual degradation.

3 Method

3.1 Method Overview

In adversarial watermark, existing models generally add perturbation to the image in a simple manner. As shown in Figure 1, such straightforward methods discourage users from using model in practice, as they add noticeable noise to the image. Their models mentioned above have simple perturbation embedding, and do not consider imperceptibility. Thus, achieving an optimal balance between these two competing objectives remains a significant challenge in AI watermark for image. Mitigating this dilemma is crucial to alleviate users' concerns; this study investigates whether the perceptual loss can help compromise a portion of this trade-off.

3.2 Proposed methods

We adopt the Fully-trained Surrogate Model Guidance (FSMG) of Anti-DreamBooth as a baseline tool to enhance image's imperceptibility. Anti-Dreambooth have an architecture that adds the gradient in the opposite direction of the Dreambooth loss, thereby corrupting the parameters of the generative model. The steps of FSMG are as follows.

Let $\mathbb{X}' = \{x'_1, x'_2, \ldots, n_i\}$ be a reference set of images and let $\mathbb{X} = \{x_1, x_2, \ldots, n_i\}$ be a set of images to be protected. The first step of FSMG involves training the parameters of a personaliza-

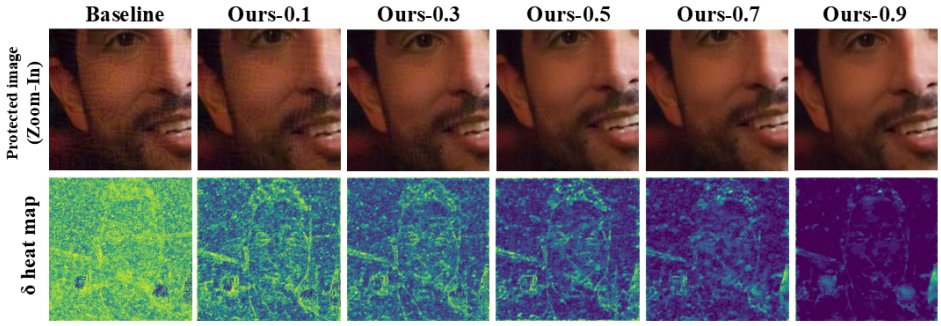

Figure 1: Each row displays the protected images (Zoom-In) and the corresponding perturbation heatmaps for both the baseline and our approach. In the perturbation heatmaps, higher values are represented in yellow, while lower values are shown in blue. Ours-0.1 to 0.9 denotes our method with λ set ranging from 0.1 to 0.9.

tion model, such as DreamBooth, on a reference set $\mathbb{X}'$ to obtain clean parameters θ. Next, the PGD algorithm is used to perturb the protecting images $\mathbb{X}$, updating them iteratively for k steps as $x_i^* = x_i + \delta^{(i)}$. Finally, the algorithm finds perturbations δ^* that maximize the loss between the perturbed images x^* and the clean parameters θ, while minimizing the loss between them to obtain contaminated parameters θ^*.

The main problem is that their perturbations embedded in the original images appear as artifacts. This issue manifests as a critical trade-off between image fidelity (imperceptibility) and robustness against AI training.

In our proposed method, we incorporate the Structural Similarity Index Measure (SSIM) loss into our overall objective function to find the optimal point that enhances imperceptibility while maintaining robustness. The SSIM loss quantifies the extent to which a perturbed image preserves the original by capturing fine structural details and contrast. In our approach, we define an adversarial loss, L_{adv}, that intentionally perturbs an input image, and we modulate the contribution of the structural similarity (SSIM) loss through a SSIM weight, hyperparameter λ. Specifically, λ is varied from 0.1 to 0.9 in increments of 0.2, allowing us to systematically investigate the trade-offs between image fidelity and robustness.

$$\mathcal{L}_{\text{adv}} = -\text{MSE}(x_i, x_{i,k}^*),$$
$$\mathcal{L}_{\text{SSIM}} = 1 - \text{SSIM}(x_i, x_{i,k}^*),$$
$$\mathcal{L}_{\text{total}} = (1-\lambda)\mathcal{L}_{\text{adv}} + \lambda\mathcal{L}_{\text{SSIM}}$$

The reason of SSIM loss as the perceptual loss is to preserve the overall structure of the original image rather than merely matching pixel values. For instance, in personalization generation tasks involving human face datasets, it is more critical to protect key facial features—such as the eyes, nose, mouth, and hair—from being learned by the AI, since prompts may change pixel values and alter certain attributes.

4 Experiment

4.1 Setup

Datasets & Training. Our datasets and the training setup adhere to our baseline's setting, enabling a fair comparison with perturbation. we select both reference image sets and clean image sets from CelebA-HQ (Karras, 2017) and VGGFace2 [1] (Cao et al., 2018). For VGGFace2, we use 280

[1]We used CelebA-HQ and VGGFace2 datasets for non-commercial research purposes. For license details, please see CelebA-HQ webpage, VGGFace2 webpage.

images from 35 subjects, whereas for CelebA, we leverage 288 images from 36 subjects. The images are preprocessed by resizing them to 512×512 pixels and applying a centercrop. while training DreamBooth (Ruiz et al., 2023) on NVIDIA RTX A6000 48GB for 1,000 steps, FSMG with SSIM Loss is trained for 100 steps. Noise budget η is set to 0.05. Training prompt is "a photo of sks person" and unseen prompt is "a $dslr$ portrait of sks person". By controlling the SSIM weight λ ranging from 0.1 to 0.9 in increments of 0.2, we identify the optimal compromise point of the trade-off between imperceptibility and the robustness.

4.2 Evaluation Metrics

For each subject and each SSIM weight λ, we generate 30 disturbed samples and computed their average score for evaluation. We utilize the **Perception metrics** to evaluate the imperceptibility of images, we use both the Structural Similarity Index (SSIM) and the Learned Perceptual Image Patch Similarity (LPIPS) metrics. Similar to the baseline metrics we utilize the **Defense metrics.** Using RetinaFace, we compute the face detection failure rate (FDFR) and, if a face is detected, calculate the ISM from cosine distances between embeddings. In additions, we include SER-FQA (Terhorst et al., 2020) and BRISQUE (Mittal et al., 2012) for a more diverse evaluation. The **User Study** is also conducted to incorporate user preferences into the trade-off analysis and identify an optimal point. In the first stage, 20 participants are asked to indicate the minimum level of imperceptibility they considered acceptable for using the model. In the second stage, participants subsequently vote the better image fidelity and rate the defense (score 1-5) against AI training for the two highest-scoring λ settings (0.1 and 0.3) relative to the baseline.

4.3 Trade-off analysis

Table 1 presents a comparative analysis of the defense performance of the Baseline, Ours-0.1, and Ours-0.3 methods on the VGGFace2 and CelebA-HQ datasets. By evaluating various metrics, our proposed approach distorts the AI images with a level of performance comparable to that of the baseline. The inference results can be found on the Figure 2. In Table 2 both Ours-0.1 and Ours-0.3 achieve higher SSIM values and lower LPIPS scores than the baseline, preserving better visual quality. In particular, Ours-0.3 achieves the best performance in both SSIM and LPIPS, and while users gave it slightly lower defense scores, it received a much higher preference for image fidelity.

As shown in Table 3, the user voting results on image fidelity for different λ values (0.1, 0.3, 0.5, 0.7, and 0.9). Overall, the results show that λ values of 0.1 and 0.3 each received 32% of the votes, indicating that users prefer these settings as they best balance image quality and defense strength. Based on these results, we adopt λ values of 0.1 and 0.3 (the highest scoring) and compare our method with both the baseline and a version without defense. Ours-0.1 and Ours-0.3 denote our method with λ set to 0.1 and 0.3, respectively. We conducted the ablation study in Table 4 to evaluate the defense performance of the baseline and our proposed methods under Gaussian blur filtering conditions with kernel sizes $K = 3, 5, 7$ on the VGGFace2 dataset. Our methods maintain consistent defense performance under various blur conditions, further demonstrating their robustness.

Table 1: Comparative analysis of defense performance between the baseline and ours on the VGGFace2 and CelebA Datasets.

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	0.06	0.60	0.70	16.12	0.13	0.41	0.69	6.07
	Baseline	0.35	0.37	**0.35**	**38.94**	0.25	0.48	**0.32**	**39.33**
	Ours-0.1	0.40	0.31	0.37	36.47	0.29	0.41	0.39	36.48
	Ours-0.3	**0.49**	**0.17**	0.53	38.04	**0.33**	**0.30**	0.50	31.38
CelebA-HQ	No Defense	0.06	0.62	0.74	14.01	0.28	0.43	0.73	6.48
	Baseline	0.51	0.30	**0.50**	36.71	0.38	0.30	**0.63**	**31.94**
	Ours-0.1	0.56	0.23	0.57	**36.88**	0.40	**0.28**	0.68	22.75
	Ours-0.3	**0.62**	**0.14**	0.65	34.90	**0.45**	0.29	0.68	14.94

Table 2: Comparative analysis of imperceptibility performance on the VGGFace2 and CelebA. User study is also included, showing the average user score (1-5) for defense against AI training and the proportion of user vote(%) for fidelity.

Method	VGGFace2		CelebA-HQ		User Study	
	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓	Defense↑	Fidelity↑
Baseline	0.86	0.21	0.90	0.17	4.30	0
Ours-0.1	0.94	0.14	0.94	0.11	**4.40**	33
Ours-0.3	**0.97**	**0.09**	**0.97**	**0.07**	3.83	**67**

Table 3: User voting results for imperceptibility of λ ranging from 0.1 to 0.9.

λ	0.1	0.3	0.5	0.7	0.9
User votes	**32%**	**32%**	24%	7%	5%

Table 4: Comparative analysis of defense performance against Gaussian Blur filtering on VGGFace2.

Kernel Size	Method	"a photo of *sks* person"			
		FDFR ↑	ISM ↓	SER-FIQ ↓	BRISQUES ↑
Gaussian K=3	Baseline	**0.64**	0.08	**0.75**	58.66
	Ours-0.1	0.62	0.09	0.79	**59.88**
	Ours-0.3	0.65	**0.06**	0.78	59.07
Gaussian K=5	Baseline	**0.54**	0.16	0.77	68.20
	Ours-0.1	0.52	0.16	**0.70**	69.98
	Ours-0.3	**0.54**	**0.10**	0.78	**71.99**
Gaussian K=7	Baseline	**0.45**	**0.00**	0.42	47.31
	Ours-0.1	0.43	0.06	0.43	**48.36**
	Ours-0.3	**0.45**	**0.00**	**0.29**	29.09

Figure 2: Selected adversarial watermark CelebA-HQ results by the baseline and our methods. Thee models are finetuned in a prompt "a photo of sks person" and a unseen prompt "a $dslr$ portrait of sks person".

5 Conclusion

In this work, we addressed the fundamental trade-off between visual fidelity and robustness in adversarial watermarking by integrating SSIM loss into the perturbation embedding process. By introducing a tunable SSIM weight, our method effectively balances imperceptibility while maintaining strong resistance against unauthorized AI training. Experimental results on CelebA-HQ and VGGFace2 demonstrate that this approach enhances image quality without compromising protection, as validated by both quantitative metrics and user studies.

These results highlight the importance of perceptual constraints in watermarking techniques, making adversarial watermarking more practical for real-world applications. By refining the trade-off between fidelity and robustness, our approach contributes to ensuring data integrity in the era of generative AI. Future work could explore adaptive SSIM weighting strategies to further optimize this balance in dynamic environments.

REFERENCES

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 694–711. Springer, 2016.

Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1931–1941, 2023.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. *arXiv preprint arXiv:2302.04578*, 2023.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Tom Sander, Pierre Fernandez, Alain Durmus, Teddy Furon, and Matthijs Douze. Watermark anything with localized messages. *arXiv preprint arXiv:2411.07231*, 2024.

Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5651–5660, 2020.

Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2116–2127, 2023.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Chenshuang Zhang, Chaoning Zhang, Sheng Zheng, Mengchun Zhang, Maryam Qamar, Sung-Ho Bae, and In So Kweon. A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai. *arXiv preprint arXiv:2303.13336*, 2023a.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11964–11974, 2024.