J Inf Process Syst, Vol.21, No.4, pp.449~456, August 2025
https://doi.org/10.3745/JIPS.03.0208 ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Adversarial Watermarking through the Integration of GAN and FGSM for Preventing Unauthorized AI Training

Ji-Hun Kim[1,*] and YongTae Shin[2]

Abstract

This paper proposes a novel adversarial watermarking method combining generative adversarial networks (GANs) and fast gradient sign method (FGSM) to prevent unauthorized artificial intelligence (AI) training while maintaining high visual quality of the watermarked content. GANs are used to generate imperceptible adversarial watermarks that are embedded into the original content, minimizing visual distortions. FGSM enhances these watermarks by introducing targeted perturbations to confuse AI models, significantly degrading their learning performance. Experiments conducted using ResNet-18 demonstrate the effectiveness of the proposed method across key metrics, including peak signal-to-noise ratio, probability shift, and MAX probability shift. The results show that the combined GAN and FGSM approach strikes a balance between maintaining the visual quality of watermarked content and achieving superior adversarial robustness compared to standalone GAN or FGSM methods. This study provides a practical reference for advancing adversarial watermarking techniques, contributing to the protection of intellectual property in the era of AI-driven content creation.

Keywords

Adversarial Watermarking, Digital Content Protection, FGSM, GAN, Probability Shift, PSNR

1. Introduction

Digital content has become a crucial asset in today's economy and society, making its protection an increasingly important challenge. Copyright infringement and unauthorized duplication threaten the value of digital media and the rights of creators. Particularly, advancements in artificial intelligence (AI) have increased the risk of unauthorized use of content during the training of AI models or in copyright infringement activities that exploit these models. The practice of AI models learning from large-scale datasets, which may include copyrighted images, videos, and texts, is no longer a theoretical issue but has emerged as a tangible problem in recent years [1].

To address this, adversarial watermarking techniques have recently gained attention as a critical tool for protecting digital content. Adversarial watermarking introduces noise into the original content to impair the learning capabilities of AI models, thereby reducing their performance [2]. Among such techniques, the fast gradient sign method (FGSM) stands out as a simple yet effective approach that can

Manuscript received Manuscript received December 23, 2024; accepted March 20, 2025.
* **Corresponding Author:** Ji-Hun Kim (bizkjh9827@gmail.com)
[1] Department of Computer Science and Engineering, Soongsil University, Seoul, Korea (bizkjh9827@gmail.com)
[2] School of Computer Science and Engineering, Soongsil University, Seoul, Korea (shin@ssu.ac.kr)

introduce confusion during the AI model's learning process [3].

On the other hand, generative adversarial networks (GANs), which generate data through the competitive learning of a generator and a discriminator, have been widely employed in tasks involving image generation and transformation. By applying the generative capabilities of GANs to adversarial watermarking, watermarks can be embedded into digital content with minimal distortion. Such watermarks are imperceptible to humans but can effectively prevent AI models from infringing on copyrights [1].

This paper proposes a novel adversarial watermarking technique that combines GANs and FGSM. The proposed approach aims to prevent AI models from unauthorized usage of copyrighted digital content. Ultimately, this study introduces a new method to counter copyright infringement facilitated by AI, contributing to the advancement of digital content protection technologies.

2. Related Work

Adversarial watermarking techniques for protecting digital content have recently gained significant attention in various studies. Particularly, techniques such as GAN and FGSM have been proposed as effective methods to prevent the unauthorized learning of digital content, either individually or in combination. This section reviews the existing research on adversarial watermarking techniques that leverage GAN and FGSM.

2.1 Adversarial Watermarking using GANs

GANs consist of two neural networks, a generator and a discriminator, that learn in competition with each other. GANs are highly effective for generating adversarial examples, where the generator embeds adversarial watermarks into original images to mislead AI models into making incorrect predictions.

GAN-based adversarial watermarking enables the insertion of highly sophisticated patterns that appear visually similar to the original image but disrupt the model's predictions. The generator produces subtle changes that preserve the visual features of the original image while introducing distortions that can mislead AI predictions. By maintaining the visual characteristics of the original content while deceiving the discriminator, GANs prove to be an effective tool for adversarial watermark generation to mislead AI models [4].

2.2 Fast Gradient Sign Method

The FGSM is a simple and fast technique used to generate adversarial examples (Fig. 1). This algorithm operates by calculating the gradient of the loss function with respect to the input image and then adding small perturbations to each pixel of the original image [5].

Fig. 1. Example of the fast gradient sign method (FGSM).

FGSM is defined mathematically as shown in Eq. (1):

$$x' = x + \epsilon \cdot sign\big(\nabla_x J(\theta, x, y)\big), \tag{1}$$

where x represents the original image, ϵ is a parameter controlling the magnitude of the noise, and $\nabla_x J$ denotes the gradient of the loss function J with respect to the input image x.

The key advantage of FGSM lies in its computational efficiency. By generating adversarial watermarks in a single step, it is well-suited for real-time attacks. Despite its simplicity, FGSM can significantly impact the predictions of AI models, making it a powerful tool for adversarial watermarking [5].

2.3 Carlini & Wagner Attack

The Carlini & Wagner (C&W) attack is known as one of the most powerful and effective methods for generating adversarial examples. The C&W attack employs an optimization-based approach to generate adversarial examples that are nearly indistinguishable from the original image while causing the model to make incorrect predictions [6].

The objective of the C&W attack is to solve the optimization problem defined in Eq. (2):

$$min\|x' - x\|_2 + c \cdot f(x'), \tag{2}$$

where x'represents the adversarial image, x is the original image, c is a balancing parameter, and $f(x')$ denotes the loss function that encourages the adversarial image to induce a targeted misclassification by the model.

The C&W attack supports various regularization methods, such as L_2 or L_∞, which allow for controlling the strength and subtlety of the attack. While this method is highly effective in generating sophisticated adversarial examples, it comes with the trade-off of high computational cost, making it more resource-intensive compared to simpler methods [6].

2.4 DeepFool

DeepFool is an algorithm designed to generate adversarial examples by introducing minimal perturbations that lead to incorrect predictions for a given model. DeepFool identifies the decision boundary of the model and computes the smallest perturbation δ required to push the input image beyond this boundary, thereby generating an adversarial image [7].

The core idea of DeepFool is to approximate the decision boundary of the model as a linear hyperplane and compute the minimal perturbation needed to flip the model's prediction. For a linearized classifier $f(x)$, the required perturbation can be expressed in Eq. (3):

$$\delta = -\frac{f(x)}{\|\nabla_x f(x)\|_2^2} \cdot \nabla_x f(x), \tag{3}$$

where $f(x)$ denotes the output of the classifier for input x, $\nabla_x f(x)$ represents the gradient of the classifier with respect to the input x, and δ signifies the minimal perturbation required to cross the decision boundary and mislead the model [7].

DeepFool iteratively refines this perturbation by recalculating the linear approximation of the decision boundary until the adversarial image successfully alters the model's prediction. This iterative process allows DeepFool to generate highly optimized adversarial examples, enabling precise and effective attacks with minimal noise [7].

3. Proposed Method

This paper proposes a novel adversarial watermarking technique that integrates GANs and FGSM to disrupt AI model training while minimizing distortions to the original digital content. The proposed approach combines two key ideas: GAN is employed to generate visually imperceptible watermarks, and FGSM is applied to enhance these watermarks with adversarial noise that prevents AI models from effectively learning the content.

3.1 Adversarial Watermark Generation using GANs

GANs are used to generate watermarks by taking random noise vectors as input and transforming them into visually imperceptible patterns that can be embedded into the original image. Specifically, the generator takes randomly sampled noise as input and iteratively refines it to match the dimensions of the image while creating a noise pattern that serves as the watermark.

The generated watermark is then added to the original image, resulting in a modified image that appears nearly identical to the original to the human eye but introduces confusion for AI models. The intensity of the watermark is carefully adjusted during this process to maintain the visual quality of the original image while ensuring that AI models face difficulties in learning from the altered image.

3.2 Adversarial Enhancement using FGSM

To strengthen the adversarial effects of the watermark generated by GAN, the FGSM is applied. FGSM enhances the watermark by introducing perturbations that exploit the AI model's gradient information. This ensures that the watermark includes patterns that are difficult for AI models to interpret, thereby preventing unauthorized use of the content in training.

The adversarial enhancement step optimizes the strength of the watermark to balance two objectives: minimizing visual distortions to the original image and maximizing the impact on the AI model's recognition performance. As a result, the final watermark not only preserves the quality of the original content but also effectively hinders AI models from learning or utilizing it without authorization.

4. Experiments and Results

This section presents the experimental results comparing the performance of the proposed adversarial watermarking technique (GAN+FGSM) against standalone GAN and FGSM methods (Fig. 2). The evaluation was conducted using two primary metrics: peak signal-to-noise ratio (PSNR) and probability shift & MAX probability shift, leveraging the ResNet-18 model available in PyTorch.

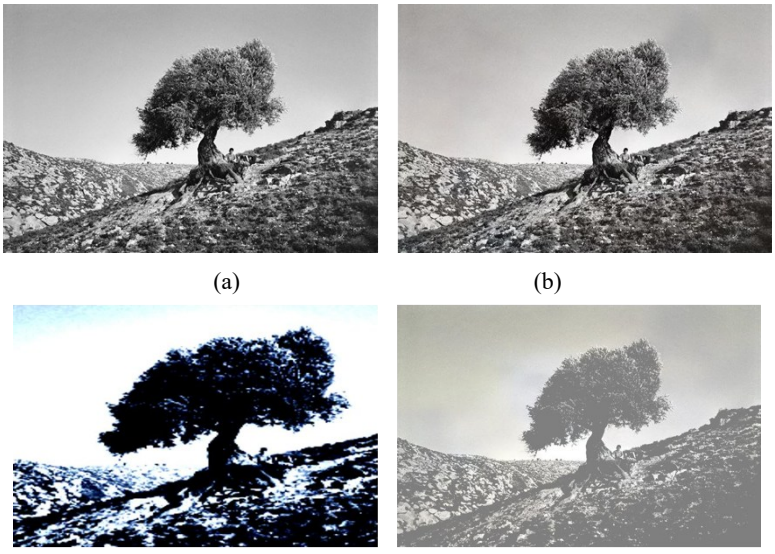
4.1 Experimental Setup

The experiments evaluated each watermarking technique based on the similarity to the original image (PSNR) and the extent of AI model prediction changes (probability shift and MAX probability shift). The evaluation was conducted using the ResNet-18 model implemented in PyTorch.

The proposed GAN+FGSM-based adversarial watermarking model was implemented using PyTorch. The generator and discriminator were trained for 100 epochs using the Adam optimizer with a learning

rate of 0.0002 ($\beta_1 = 0.5$, $\beta_2 = 0.999$) and a batch size of 64. Input images were resized to 128×128 pixels. Binary cross-entropy (BCE) was used as the loss function, and model weights were initialized using a normal distribution with mean 0 and standard deviation 0.02.

The FGSM component was applied with an epsilon (ε) value of 0.05 to generate imperceptible perturbations, while the generator was trained to deceive the discriminator and preserve visual similarity.



(a) (b)

(c) (d)

Fig. 2. Content with embedded adversarial watermark: (a) original image, (b) GAN method, (c) FGSM method, and (d) GAN+FGSM method.

4.1.1 PSNR

This metric assesses the visual quality of the watermarked image by calculating the ratio between the maximum possible pixel intensity and the distortion introduced by watermarking. A higher PSNR value indicates that the watermarked image is more similar to the original image with less perceptible degradation.

4.1.2 Probability shift

This metric measures the change in AI model prediction probabilities when watermarked images are used instead of the original image. A higher value indicates a greater shift in prediction probabilities.

4.1.3 MAX probability shift

This metric evaluates the change in the AI model's confidence for the most likely predicted class. A higher value indicates greater confusion in the AI model.

4.2 PSNR Analysis

Table 1 shows the results of measuring the similarity between the watermarked and original images using the PSNR metric. When using GAN alone, the PSNR value was the highest at 33.55 dB, indicating

minimal visual distortion and high similarity to the original image. In contrast, FGSM alone resulted in the lowest PSNR value of 12.12 dB, suggesting significant degradation in visual quality. The combined GAN and FGSM method yielded a PSNR value of 18.79 dB, which is lower than GAN alone but slightly higher than FGSM alone, indicating moderate visual distortion.

Table 1. Comparison of PSNR values for different watermarking methods

	GAN	FGSM	GAN+FGSM
PSNR (dB)	33.55	12.12	18.79

4.3 Probability Shift Analysis

Table 2 presents the results of measuring changes in the AI model's prediction probabilities using the Probability Shift metric. When using GAN alone, the probability shift value was the lowest at 0.3804, indicating that the AI model's prediction probabilities experienced minimal changes. In contrast, using FGSM alone resulted in a probability shift value of 1.4681, demonstrating a significant change in the AI model's predictions. The proposed combination of GAN and FGSM achieved the highest probability shift value of 1.5537, indicating that it caused the greatest confusion for the AI model.

Table 2. Comparison of probability shift values for different watermarking methods

	GAN	FGSM	GAN+FGSM
Probability shift value	0.3804	1.4681	1.5537

4.4 MAX Probability Shift Analysis

Table 3 presents the results of measuring changes in the probability of the AI model's most confident predicted class using the MAX probability shift metric. When using GAN alone, the MAX probability shift value was the lowest at 0.0888, indicating that the AI model still predicts a class similar to the original image. When using FGSM alone, the MAX probability shift value increased to 0.1260. In the combined approach using GAN and FGSM, the MAX probability shift value reached the highest at 0.4270, demonstrating that the proposed method causes the most significant change in the AI model's prediction probabilities.

Table 3. Comparison of MAX probability shift values for different watermarking methods

	GAN	FGSM	GAN+FGSM
MAX probability shift value	0.0888	0.126	0.427

5. Conclusion

In this study, we proposed a novel adversarial watermarking technique combining GANs and FGSM, and evaluated its performance using PSNR, probability shift, and MAX probability shift metrics. The proposed method proves to be an effective approach for preventing unauthorized AI model training on digital content, successfully disrupting the AI model's learning process while maintaining the visual quality of the original image.

The PSNR analysis revealed that using GAN alone resulted in the highest similarity and minimal distortion of the original image. On the other hand, the combined approach of GAN and FGSM showed

a reduction in PSNR due to additional noise, but the drop was not significant, and it still demonstrated excellent visual quality. Moreover, the probability shift and MAX probability shift metrics showed that the GAN+FGSM approach caused the largest change in the AI model's prediction probabilities, effectively disrupting its learning process.

These results suggest that the integrated GAN and FGSM approach is an effective solution for adversarial watermarking in digital content protection. Notably, this method can prevent unauthorized AI learning while maintaining an acceptable level of image quality, making it a promising tool for real-world digital content protection scenarios.

Future research will explore the applicability of this technique to various AI models and work towards enhancing the robustness of watermarking methods. Such efforts will contribute not only to copyright protection of digital content but also to promoting the ethical use of AI technologies.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

This research was supported by the Ministry of Culture, Sport and Tourism R&D Program through the Korea Creative Content Agency grant, funded by the Ministry of Culture, Sport and Tourism in 2024 (Project No. RS-1375027563, Development of Copyright Technology for OTT Contents Copyright Protection Technology Development and Application; 100%).

Acknowledgments

This paper is the extended version of “Adversarial Watermarking Combining GAN and FGSM: Preventing Unauthorized Learning of AI Models,” in the 2024 Annual Conference of KIPS (ACK 2024) held in Gwangju, Republic of Korea, dated October 31-November 2, 2024.

References

[1] S. Park and Y. G. Shin, “Generative convolution layer for image generation,” 2021 [Online]. Available: https://arxiv.org/abs/2111.15171.

[2] X. Jia, X. Wei, X. Cao, and X. Han, “Adv-watermark: a novel watermark perturbation for adversarial examples,” in *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle, WA, USA, 2020, pp. 1579-1587. https://doi.org/10.1145/3394171.3413976

[3] S. Lupart and S. Clinchant, “A study on FGSM adversarial training for neural retrieval,” 2023 [Online]. Available: https://arxiv.org/abs/2301.10576.

[4] X. Zhong, A. Das, F. Alrasheedi, and A. Tanvir, “A brief, in-depth survey of deep learning-based image watermarking,” *Applied Sciences*, vol. 13, no. 21, article no. 11852, 2023. https://doi.org/10.3390/app132111852

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2015 [Online]. Available: https://arxiv.org/abs/1412.6572.

[6] T. W. Kim, S. M. Hyun, and E. J. Hong, "Comparison of adversarial example restoration performance of VQ-VAE models depending on image segmentation," *Journal of the Institute of Signal Processing and Systems*, vol. 23, no. 4, pp. 194-199, 2022. https://doi.org/10.23087/jkicsp.2022.23.4.002

[7] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: a simple and accurate method to fool deep neural networks," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2574-2582. https://doi.org/10.1109/CVPR.2016.282

Ji-Hun Kim **https://orcid.org/0009-0004-4914-9102**

He received his B.S. degree in Computer Science and Engineering from Dongguk University in 2024. Since March 2024, he has been with the Department of Computer Science & Engineering at Soongsil University as an M.S. candidate. His current research interests include artificial intelligence, big data, and copyright protection technologies.

YongTae Shin **https://orcid.org/0000-0002-1199-1845**

He received his B.S. degree in Industrial Engineering from Hanyang University in February 1985, M.S. degree in Computer Science from the University of Iowa in December 1990, and Ph.D. degree in Computer Science from the University of Iowa in February 1994. Since March 1995, he has been a professor in the School of Computer Science & Engineering at Soongsil University. His current research interests include computer networks, distributed computing, internet protocols, and e-commerce technologies.