논문 2024-4-8 http://dx.doi.org/10.29056/jsav.2024.12.08

Self-Attention 기반 GAN 워터마킹의 비가시성과 강인성 개선 연구

이종호*, 이소영*, 신용태**†

Research on Enhancing the Robustness and Imperceptibility of GAN Watermarking Based on Self-Attention

Jong-Ho Lee*, So-Yeong Lee*, Yong-Tae Shin**†

요 약

디지털 이미지의 저작권 보호를 위한 워터마킹 기술은 이미지의 시각적 품질을 유지하면서도 다양한 공격 환경에서도 워터마크를 안정적으로 복원할 수 있어야 한다. 본 연구에서는 Self-Attention 메커니즘을 도입한 GAN 기반 워터마킹 모델을 제안한다. 제안된 모델은 Self-Attention을 활용하여 이미지 내 전역적(Global) 관계와 중요한 특징을 학습함으로써 워터마크 삽입 이미지의 비가시성과 강인성을 동시에 개선하였다. 실험 결과, JPEG 압축, 가우시안 노이즈, 블러, 크롭, 밝기 및 대비 조정 등 다양한 공격 시나리오에서 Bit Accuracy는 평균적으로 약 7.94% 향상되었으며, PSNR(Peak Signal-to-Noise Ratio)은 약 1.42dB, SSIM(Structural Similarity Index Measure)은 약 0.06 증가하여 비가시성과 강인성 모두에서 기존 모델 대비 우수한 성능을 입증하였다.

Abstract

Digital image watermarking technology for copyright protection must maintain the visual quality of images while ensuring robust recovery of watermarks under various attack scenarios. This study proposes a GAN-based watermarking model incorporating a Self-Attention mechanism. The proposed model leverages Self-Attention to learn global relationships and critical features within images, simultaneously enhancing the imperceptibility and robustness of watermarked images. Experimental results demonstrate that, under various attack scenarios including JPEG compression, Gaussian noise, blur, cropping, brightness, and contrast adjustments, the Bit Accuracy improved by an average of approximately 7.94%, while the Peak Signal-to-Noise Ratio (PSNR) increased by approximately 1.42dB and the Structural Similarity Index Measure (SSIM) improved by approximately 0.06, showcasing superior performance in both imperceptibility and robustness compared to existing models.

한글키워드 : 저작권 보호, Self-Attention, 생성적 적대 신경망, 워터마킹

keywords : Copyright Protection, Self-Attention, Generative Adversarial Network(GAN), Watermarking

* 숭실대학교 컴퓨터학과
** 숭실대학교 컴퓨터학부
† 교신저자: 신용태(email: shin@ssu.ac.kr)

접수일자: 2024.10.11. 심사완료: 2024.12.03.
게재확정: 2024.12.20.

1. 서 론

생성형 AI 기술의 발전으로 이미지나 예술 작품 등 매우 사실적인 콘텐츠를 제작할 수 있는 능력이 크게 향상되었다. 이에 따라 생성형 AI 기술이 연구나 산업 분야에서 널리 활용되는 한편, 저작권 보호 및 지식재산권(Intellectual Property Right)에 대한 우려도 커지고 있다[1]. 이러한 문제를 해결하기 위해 디지털 콘텐츠의 출처, 제공자, 그리고 콘텐츠 인증의 중요성이 대두되었으며, 이미지 분야에서는 워터마킹 기술이 지속적으로 발전해왔다[2]. 워터마킹은 디지털 콘텐츠의 소유권을 명시하거나 불법 복제를 방지하기 위해 디지털 콘텐츠에 고유한 패턴(워터마크)을 삽입하는 기술을 의미한다.

최근에는 CNN(Convolutional Neural Network), GAN(Generative Adversarial Network), 그리고 DNN(Deep Neural Network)과 같은 딥러닝 기반 워터마킹 기술이 주목받고 있다. 특히 GAN 기반 워터마킹 모델은 생성자(Generator)와 판별자(Discriminator)가 상호작용하며 학습하는 구조를 통해 워터마크가 삽입된 이미지를 자연스럽고 강인하게 생성할 수 있다. 이러한 딥러닝 기반 워터마킹의 주요 장점은 다양한 왜곡 및 공격에도 워터마크를 유지할 수 있는 강인성(Robustness)과 원본 콘텐츠의 시각적 품질을 유지하는 비가시성(Invisibility)이다[3]. 이러한 기술은 이미지 콘텐츠 보호를 위한 효과적인 솔루션으로 자리 잡고 있다.

Self-Attention은 전역적 관계를 효과적으로 모델링하고, 이미지 내 공간적 및 채널 간 관계를 강화하여, 이미지의 전반적인 품질을 유지하면서도 복잡한 패턴을 학습하는 데 기여한다[4]. 이러한 접근 방식은 이미지의 모든 위치 간 관계를 학습함으로써 전반적인 구조와 특징을 효과적으로 반영한다[7]. 따라서 특정 영역만을 처리하는 지역적(Local) 접근 방법과 차이를 보인다. 이에 본 연구는 GAN 기반 워터마킹 모델에 Self-Attention 메커니즘을 적용하여, Self-Attention 적용에 따른 시각적 품질을 비교한다. 또한, 시각적 품질에 대한 비교 결과를 바탕으로 Self-Attention 기반 모델을 최적화하여 워터마킹의 비가시성과 강인성을 극대화하는 방안을 제안하고자 한다.

본 논문은 다음과 같이 구성되어 있다. 2장 관련 연구에서는 GAN 기반 워터마킹과 Self-Attention에 관한 기존 연구를 검토하며, 본 연구의 차별성과 기여를 서술한다. 3장 모델 설계에서는 제안된 GAN 기반 워터마킹 모델과 Self-Attention 메커니즘의 통합 과정을 설명하고, 모델 구조와 학습 목표를 제시한다. 4장 구현 및 실험 설정에서는 연구에 사용된 데이터셋, 학습 환경, 평가 지표 및 공격 시나리오를 서술한다. 마지막으로, 5장 실험 결과 및 분석에서는 제안된 모델의 성능을 기존 연구와 비교 분석하며, 비가시성과 강인성 측면에서의 성과를 제시한다.

2. 관련 연구

2.1 GAN 기반 워터마킹 기술

전통적인 워터마킹 기술은 주로 주파수 도메인(예: DCT, DWT)이나 공간 도메인 기반 접근법을 사용해왔다. 최근 딥러닝 기반 기법은 이러한 기존 접근법에 비해 비가시성과 강인성 측면에서 뛰어난 성능을 보여주며, 워터마킹 연구의 새로운 패러다임으로 자리 잡고 있다[3]. 특히, GAN을 활용한 워터마킹 기법은 생성자와 판별자가 경쟁적으로 학습하는 구조를 통해 자연스럽고 강인한 워터마크 삽입 이미지를 생성할 수 있다[5].

Zhu et al.[6]은 데이터 은닉과 워터마킹을 위한 HiDDeN 모델을 제안하였다. 이 모델은 생성자가 워터마크가 삽입된 이미지를 생성하고, 판별자는 해당 이미지가 워터마크를 포함하는지를 판단하는 구조로 설계되었다. HiDDeN은 데이터 은닉과 비가시성 측면에서 성공적인 성과를 보였으나, 특정 노이즈에 대해 훈련되지 않은 경우 공격에 대한 강인성은 제한적이라는 한계가 있었다. 그러나 HiDDeN은 GAN 기반 적대적 학습 구조를 적용하여 데이터 은닉 및 비가시성에서 성공적인 사례를 제시한 워터마킹 모델이다.

Hao et al.[5]은 HiDDeN 모델을 기반으로 한 GAN 기반 이미지 워터마킹 모델을 제안하였다. 이 모델은 생성자와 판별자로 구성된 GAN 구조를 사용하며 생성자는 입력 이미지에 워터마크를 삽입하고, 디코더를 통해 노이즈가 추가된 이미지의 워터마크를 복원한다. 판별자는 고역통과 필터(High-Pass Filter, HPF)를 통해 이미지의 고주파 성분을 강조하며, 이를 기반으로 워터마크가 삽입된 이미지를 판별한다. 이 과정은 생성자가 워터마크를 중간 주파수 영역에 삽입하도록 유도하며, JPEG 압축과 같은 주파수 기반 공격에 대한 강인성을 높이는 동시에, 인간 시각 시스템(Human Visual System, HVS)이 저주파 변화에 민감한 특성을 활용해 비가시성을 유지한다. 이를 통해 워터마크가 이미지 중심 영역에서는 시각적으로 덜 눈에 띄면서도, JPEG 압축, 가우시안 블러, 크롭 등 다양한 공격에 대한 강인성을 유지할 수 있었다. 실험 결과, 해당 모델은 평균 비트 오류율(Bit Error Rate, BER)을 낮추고 구조적 유사성 지수(SSIM)를 향상시키며, 기존의 HiDDeN 모델보다 우수한 성능을 보였다.

이러한 GAN 기반 워터마킹 기술은 워터마킹의 효율성과 성능을 개선하는 데 핵심적인 역할을 하며, 강인성과 비가시성의 균형을 유지하는 방향으로 꾸준히 연구되고 있다.

2.2 Self-Attention 메커니즘

Self-Attention 메커니즘은 GAN에 적용되어 이미지 전체의 관계를 효과적으로 모델링하며, 전반적인 특징을 학습하는 데 중요한 역할을 한다. Zhang et al.[4]이 제안한 Self-Attention Generative Adversarial Network(SAGAN)는 Self-Attention 메커니즘을 GAN 구조에 통합하여, 이미지 생성 과정에서 지역적 특징과 전역적 특징 간의 균형을 학습하여 이미지 품질을 크게 향상시켰다. 특히, Self-Attention은 공간적 관계와 채널 간 상호작용을 강화하여 생성 모델의 효율성을 높이는 데 기여했다.

이를 위해 입력 데이터의 각 위치를 Query(Q), Key(K), Value(V)라는 세 가지 요소로 변환하고, 이들의 내적을 활용하여 가중치를 계산한다. 입력 데이터를 $X \in R^{n \times d}$ 라고 하면, 학습 가능한 가중치 행렬 W^Q, W^K, $W^V \in R^{d \times d_k}$를 통해 Query, Key, Value를 생성하며 이를 각각 $Q = XW^Q, K = XW^K, V = XW^V$로 나타낸다. 이후 Query와 Key의 내적을 계산하여 유사도를 구하고, 이를 Softmax 함수로 정규화하여 Attention Weight를 도출한다. 최종적으로 이러한 가중치를 Value에 곱하여 새로운 표현을 생성한다. 이 과정에 대한 수식(1)은 다음과 같다. 여기서 $\sqrt{d_k}$는 Query와 Key의 차원을 정규화하여 값의 크기를 안정화하고, 학습 과정을 용이하게 만든다.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

그림 1은 Self-Attention 메커니즘의 작동 방식을 시각적으로 나타낸다. 입력된 피처 맵은 1×1 합성곱(Convolution) 연산을 통해 Query, Key, Value로 변환되며, Query와 Key의 내적을 통해 각 위치 간의 상호작용 강도를 계산한다.

이후, Softmax 연산으로 정규화된 Attention Map은 Value와 결합되어 최종적으로 처리된 피처 맵(Self-Attention Feature Maps)을 생성한다. 이 과정을 통해 이미지 내 전역적 관계 정보를 반영할 수 있다.

본 연구는 Self-Attention Generative Adversarial Network 모델에서 도입된 Self-Attention 메커니즘을 기반으로, GAN 기반 모델에서 Self-Attention이 워터마크의 강인성과 비가시성에 미치는 영향을 분석하였다. 이를 위해 Self-Attention 메커니즘의 핵심 개념을 GAN 기반 워터마킹 모델에 맞게 변형하여 적용하였다. 이는 SAGAN의 완전한 Self-Attention 구현과는 차이가 있다. 예를 들어, Query, Key, Value 메커니즘을 활용하되, 워터마크 삽입이라는 목적에 맞춰 계산 구조를 간소화하였다. 이러한 간소화는 모델의 훈련 효율성을 높이고 워터마킹에 특화된 전역적 특징에 학습에 중점을 두기 위함이다.

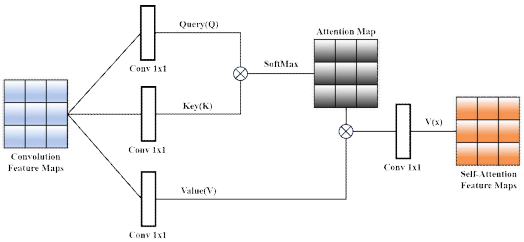

그림 1. Self-Attention 메커니즘 [4]
Fig. 1. Self-Attention Mechanism [4]

3. 제안 모델

3.1 제안 모델의 구조

본 연구는 Self-Attention이 비가시성과 강인성에 미치는 영향을 비교하기 위해 GAN 기반 워터마킹 모델에 Self-Attention 메커니즘을 적용하는 모델을 제안한다. 그림 2는 Self-Attention이 적용된 워터마킹 모델의 전체 구조를 나타낸다. 모델의 생성자는 인코더-디코더 구조를 기반으로 하고, 중간 레이어에 Self-Attention 모듈을 통합하였다.

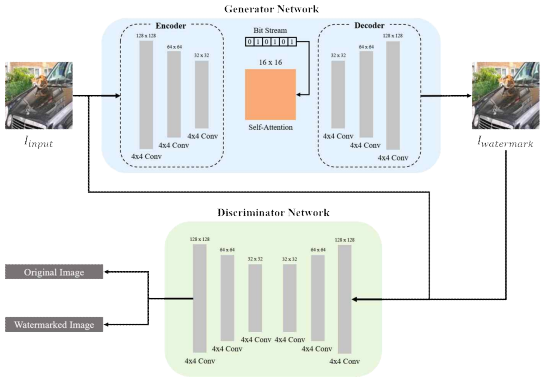

그림 2. 제안하는 모델 구조
Fig. 2. Proposed Model Structure

생성자 네트워크는 입력 이미지와 워터마크 비트 스트림을 받아 워터마크가 삽입된 이미지를 생성한다. 인코더는 4×4 크기의 합성곱 연산을 반복하여 입력 이미지의 저수준 특징을 추출하며, 공간 해상도를 점진적으로 축소한다. 이러한 구조는 이미지의 세부적이고 중요한 특징을 학습하는 데 적합하며, 워터마크 삽입의 기반이 되는 특징맵을 제공한다 [8]. Self-Attention 모듈은 인코더에서 생성된 특징맵을 기반으로 이미지 전체의 상관관계를 학습한다. 이는 각 위치 간의 상관관계를 모델링함으로써, 이미지 전역에 걸쳐 워터마크가 고르게 삽입될 수 있도록 지원한다. 또한, Self-Attention은 공간적 상호작용을 반영하여 워터마킹 삽입 후에도 이미지 품질을 유지하도록 돕는다[4]. 디코더는 Self-Attention에서 학습된 특징맵에 비트 스트림 형식의 워터마크를 삽입한 뒤, 공간적 해상도를 복원하여 워터마크가 삽입된 이미지를 생성한다.

판별자 네트워크는 생성된 이미지와 원본 이

미지를 구별하고 생성자가 시각적 품질을 유지하면서도 공격에 강인한 워터마크를 삽입할 수 있도록 학습한다. 판별자도 4×4 크기의 합성곱 연산을 반복적으로 적용하여 입력 이미지에서 중요한 고수준 특징을 추출한다. 이러한 작은 커널 크기는 세밀한 디테일을 효과적으로 학습하는 데 유리하며, 워터마크가 삽입된 이미지와 원본 이미지의 미세한 차이를 감지하는 데 적합하다.

3.2 손실 함수

생성자와 판별자의 학습을 최적화하기 위해 각각 적합한 손실 함수 구조를 제안한다. 생성자의 손실 함수는 시각적 품질과 강인성을 동시에 향상시키는 데 중점을 두고, 판별자의 손실 함수는 학습 안정성을 유지하면서 생성자와의 상호작용을 최적화하도록 설계되었다. 생성자의 손실 함수는 적대적 손실, L1 손실, 그리고 Perceptual 손실로 구성된다.

- 적대적 손실(L_{adv}): 생성된 워터마크 삽입 이미지가 판별자로부터 원본 이미지로 인식되도록 유도하며, Binary Cross-Entropy(BCE)를 기반으로 정의된다 [9].
- L1 손실(L_{L1}): 원본 이미지와 워터마크 삽입 이미지 간의 픽셀 단위 차이를 최소화하여, 생성된 이미지의 시각적 품질을 보존한다.
- perceptual 손실(L_{prec}): 사전 학습된 VGG-16 네트워크를 활용하여 원본 이미지와 워터마크 삽입 이미지 간의 고수준 특징 유사성을 유지한다. 이를 통해 워터마크 삽입 이미지가 시각적으로 자연스럽고 일관성을 유지하도록 한다.

최종 생성자의 손실 함수(2)는 다음과 같다.

$$L_G = L_{adv} + \lambda_1 L_{L1} + \lambda_2 L_{prec} \qquad (2)$$

판별자의 손실 함수는 진짜 이미지와 생성된 이미지를 구별하는 역할을 하며, BCE 기반의 손실과 R1 Gradient Penalty로 구성된다.

- 진짜 이미지 손실(L_{real}): 판별자가 원본 이미지를 진짜로 올바르게 분류하도록 유도한다.
- 가짜 이미지 손실(L_{fake}): 판별자가 생성된 워터마크 삽입 이미지를 가짜로 정확히 구별하도록 학습한다.
- R1 Gradient Penalty(L_{R1}): 판별자의 학습 안정성을 보장하고 기울기 소실 문제를 방지하기 위해 추가된 손실 요소로, 네트워크가 과도하게 학습되지 않도록 제약을 부여한다.

최종 생성자의 손실 함수(3)는 다음과 같다.

$$L_G = \frac{L_{real} + L_{fake}}{2} + \lambda_3 L_{R1} \qquad (3)$$

4. 구현 세부 사항

4.1 구현을 위한 환경 구축

본 연구는 Google Colab 환경에서 Python과 PyTorch 프레임워크를 사용하여 구현하였다. 하드웨어로는 NVIDIA A100 GPU를 활용하여 모델 학습 속도를 최적화하였다. 학습 데이터셋으로는 MS-COCO 2017을 사용하였으며, 데이터 전처리를 통해 이미지를 128×128 해상도로 리사이즈하였다. 입력 이미지에 삽입되는 워터마크는 임의로 생성된 64비트의 이진 배열로 설정되었다. GAN 기반 워터마킹 모델과 Self-Attention을 도입한 모델은 동일한 환경에서 학습되었으며, 이를 통해 공정한 성능 비교가 이루어지도록 설계하였다.

4.2 성능평가 기준 및 데이터셋

데이터셋은 16,000장의 훈련 이미지와 4,000장

의 검증 이미지를 사용하였다. 학습은 100 에포크 동안 수행되었으며, Adam 옵티마이저를 사용하였다. 학습률은 0.0002로 설정하였고, 배치 크기는 16으로 설정하였다. 검증 데이터는 5 에포크마다 사용하여 모델의 성능을 모니터링하였다. 학습이 완료된 후에는 1,000장의 테스트 이미지를 대상으로 워터마크 삽입 및 복원 성능을 평가하였다.

성능 평가는 다음과 같은 두 가지 주요 기준을 사용하였다:

- 비가시성: 워터마크 삽입 이후의 이미지 품질을 평가하기 위해 PSNR과 SSIM을 사용하였다. PSNR은 높은 값을 가질수록 원본 이미지와 삽입된 이미지의 유사성이 크다는 것을 의미하며, SSIM은 구조적 유사성을 평가한다.
- 강인성: 다양한 공격 환경에서 워터마크를 복원하는 정확도를 평가하기 위해 Bit Accuracy(BA)를 사용하였다. BA는 삽입된 워터마크와 복원된 워터마크 간의 일치율을 나타낸다.

4.3 공격 시나리오

모델의 강인성을 평가하기 위해 다섯 가지 공격 시나리오를 설계하였다.

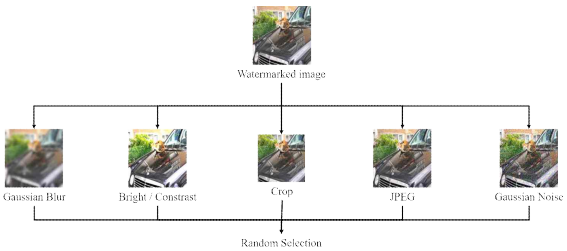

그림 3. 공격 유형
Fig 3. Attack Type

각 공격은 무작위로 선택된 워터마크 삽입 이미지에 적용되었으며, 공격 후 복원된 워터마크의 정확도를 평가하였다. 공격 강도는 JPEG 압축률 50%, 가우시안 노이즈의 표준편차 0.1, 가우시안 블러 커널 크기 5×5, 크롭 영역 10%, 밝기 및 대비 조정 범위 ±20%로 설정하였다.

그림 3은 워터마크 삽입 이미지에 JPEG 압축, 가우시안 노이즈, 가우시안 블러, 밝기 및 대비 조정, 크롭핑 등의 공격을 적용하는 과정을 시각적으로 나타낸다. 실제 환경에서 자주 발생하는 공격으로 선택하였고 다양한 조건에서 공정한 평가를 위해 무작위 샘플링 방식을 사용하였다.

5. 실험 결과 및 분석

5.1 비가시성 평가

비가시성은 워터마크 삽입 이후 이미지 품질을 측정하기 위해 PSNR과 SSIM을 사용하였다. 표 1의 결과에 따르면, Self-Attention을 적용한 모델은 PSNR이 약 1.42dB(5.04%), SSIM이 약 0.06(7.06%) 증가하여 비가시성에서 개선된 결과를 보였다. 이러한 결과는 Self-Attention이 워터마크 삽입으로 인한 품질 저하를 줄였음을 보여준다. 그림 4는 Self-Attention 적용 전후의 워터마크 삽입 이미지를 비교한 것으로, 시각적 품질 차이를 직관적으로 확인할 수 있다.

Method	Without Self-Attention	With Self-Attention
Sample Image		

그림 4. Self-Attention 적용 전후 비교
Fig 4. With vs. Without Self-Attention

Method	Invisibility		Robustness					
	PSNR	SSIM	JPEG(50%)	Gaussian Noise ($\sigma = 0.1$)	Gaussian Blur (5x5)	Crop (10%%)	Brightness/ Contrast (±20%)	Average
Without Self-Attention	28.16	0.85	0.60	0.59	0.63	0.68	0.65	0.63
With Self-Attention	29.58	0.91	0.65	0.67	0.67	0.74	0.71	0.68

표 1. 성능 평가 결과
Table 1. Performance Evaluation Results

5.2 강인성 평가

강인성은 1,000개의 테스트 이미지 중 각 공격 유형에 대해 무작위로 200개의 이미지를 선택하여 공격을 수행하였다. 각 공격에 대한 워터마크 복원 성능은 BA를 기준으로 평가되었다. 표 1에 따르면, Self-Attention을 적용한 모델은 모든 공격 유형에서 Self-Attention을 적용하지 않은 모델보다 높은 BA를 기록하였다.

JPEG 압축(50%) 공격에서는 BA가 0.60에서 0.65로 증가하여 성능이 약 8.3% 향상되었다. 가우시안 노이즈($\sigma = 0.1$) 공격에서는 BA가 0.59에서 0.67로 증가하며 약 13.6%의 성능 향상을 보였고, 가우시안 블러(5x5) 공격에서는 BA가 0.63에서 0.67로 증가해 약 6.3%의 향상을 기록했다. 크롭(10%) 공격에서는 BA가 0.68에서 0.74로 증가하여 약 8.8%의 성능 향상이 나타났으며, 마지막으로 밝기/대비(±20%) 공격에서는 BA가 0.65에서 0.71로 증가하며 성능이 약 9.2% 향상되었다.

이와 같이 모든 공격 유형에서 BA가 향상되었으며, 평균 BA는 0.63에서 0.68로 증가하여 성능이 약 7.94% 향상되었다. 이러한 결과는 Self-Attention 메커니즘이 이미지의 전역적 특징을 효과적으로 학습하여, 공격으로 인한 정보 손실을 보완하고 워터마크 복원의 안정성을 높이는 데 기여했음을 보여준다.

6. 결론 및 향후 연구

본 연구에서는 GAN 기반 워터마킹 모델에 Self-Attention 메커니즘을 도입하여, 워터마크의 비가시성과 강인성에 미치는 영향을 분석하였다. 실험 결과, Self-Attention을 도입한 워터마킹 모델은 비가시성 측면에서 도입하지 않은 모델보다 PSNR이 약 1.42dB, SSIM이 약 0.06 증가하고, 강인성 평가에서는 모든 공격 유형에서 Self-Attention을 도입한 워터마킹 모델이 더 높은 BA를 기록하였다.

특히 JPEG 압축(8.3% 향상), 가우시안 노이즈(13.6% 향상) 및 크롭 공격(8.8% 향상)에서 뛰어난 성능을 보였다. 이는 Self-Attention 메커니즘이 이미지의 전역적 특징을 학습하여 다양한 공격에 강인한 워터마킹 모델을 설계하는 데 기여함을 입증한다.

그러나 본 연구에는 몇 가지 한계점이 존재한다. 첫째, Self-Attention 도입으로 비가시성과 강인성 모두에서 개선을 가져왔다. 그러나 강인성 평가에서 보인 BA 개선 폭에 비하면, 비가시성의 개선 폭은 상대적으로 작게 나타났다. 이는 Self-Attention 메커니즘이 이미지 품질 유지보다는 공격 대응 능력을 더욱 효과적으로 학습한

결과일 가능성이 있다. 둘째, 워터마킹의 강인성을 향상시키는 과정에서 특정 공격 유형에 대한 성능 차이가 발생하여 특정 조건에서 Self-Attention의 전역적 특징 학습이 한계를 가질 수 있음을 보여준다.

이를 보완하기 위해 향후 연구에서는 다음과 같은 방향으로 연구를 확장할 계획이다. 첫째, 비가시성 향상을 위하여 시각적 품질 기반 손실 함수의 새로운 조합을 도입함으로써 워터마크 삽입 후의 이미지 품질을 개선한다. 둘째, Self-Attention 구조를 최적화하여 워터마크 정보 학습을 더욱 정교화한다. 마지막으로 실제 환경에서 발생할 수 있는 복합 공격(예: JPEG 압축과 가우시안 노이즈의 조합)에 대해 성능을 평가함으로써, 모델의 실질적 강인성을 더욱 강화할 것이다.

이러한 방향으로 연구를 확장함으로써 워터마킹 시스템의 비가시성과 강인성 사이의 균형을 최적화하고, 다양한 공격 시나리오에서도 우수한 성능을 발휘하는 솔루션을 제공할 것으로 기대된다.

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2024년도 문화기술 연구개발 사업으로 수행되었음(과제명 : OTT 콘텐츠 저작권 보호기술개발 및적용을 위한 저작권기술(+법) 융합인재양성, 과제번호 : RS-2023-00225267)

참 고 문 헌

[1] Haonan Zhong, Jiamin Chang, Ziyue Yang, Tingmin Wu, Pathum Chamikara Mahawaga Arachchige, Chehara Pathmabandu, Minhui Xue, “Copyright Protection and Accountability of Generative AI: Atack, Watermarking and Atribution”, arXiv preprint arXiv:2303.09272, Mar. 15. 2023. DOI: https://doi.org/10.48550/arXiv.2303.09272

[2] Ray, A., Roy, S., “Recent trends in image watermarking techniques for copyright protection: a survey”, Int J Multimed Info Retr 9, 249 - 270, OCT. 28. 2020. DOI: https://doi.org/10.1007/s13735-020-00197-9

[3] Himanshu Kumar Singh and Amit Kumar Singh, “Comprehensive review of watermarking techniques in deep-learning environments”, Journal of Electronic Imaging 32(03):1-23, Nov. 2023. DOI: 10.1117/1.JEI.32.3.031804

[4] Han Zhang, Ian Goodfellow, Dimitris Metaxas, Augustus Odena, “Self-Attention Generative Adversarial Networks”, arXiv preprint arXiv:1805.08318, May 21, 2018. DOI: https://arxiv.org/abs/1805.08318

[5] K. Hao, G. Feng and X. Zhang, “Robust image watermarking based on generative adversarial network”, in China Communications, vol. 17, no. 11, pp. 131-140, Nov. 23. 2020, DOI: 10.23919/JCC.2020.11.012.

[6] Jiren Zhu, Russell Kaplan, Justin Johnson, Li Fei-Fei, “HiDDeN: Hiding Data With Deep Networks”, arXiv preprint arXiv:1807.09937, Jul 26, 2018. DOI: https://arxiv.org/abs/1807.09937

[7] Artaches Ambartsoumian, Fred Popowich, “Self-Attention: A Better Building Block for Sentiment Analysis Neural Network Classifiers”, arXiv preprint arXiv:1812.07860, Dec. 19, 2018. DOI: https://arxiv.org/abs/1812.07860

[8] A. Vaswani et al., “Attention is All You Need”, Advances in Neural Information Processing Systems (NIPS), vol. 30, 2017. DOI: https://arxiv.org/abs/1706.03762. https://arxiv.org/abs/1511.06434d

[9] Jianwei Fei, Zhihua Xia, Benedetta Tondi, Mauro Barni, “Supervised GAN Watermarking for Intellectual Property Protection”, arXiv preprint arXiv:2209.03466, Sep. 7. 2022. DOI: https://doi.org/10.48550/arXiv.2209.03466

저 자 소 개

이종호(Jong-Ho Lee)

2023.2 숭실대학교 전산원 졸업
2023.9-현재 : 숭실대학교 컴퓨터학과 석사과정
<주관심분야> 인공지능, 빅데이터

이소영(So-Yeong Lee)

2023.2 한양 사이버대학교 졸업
2023.3-현재 : 숭실대학교 컴퓨터학과 석사과정
<주관심분야> 빅데이터, 인공지능

신용태(Yong-Tae Shin)

1985.2 한양대학교 산업공학과 졸업
1990.12 Univ. of Iowa, Computer Science 석사
1994.2 Univ. of Iowa, Computer Science 박사
1995.3-현재 : 숭실대학교 컴퓨터학부 교수
<주관심분야> 컴퓨터네트워크, 분산 컴퓨팅, 인터넷프로토콜, 전자상거래 기술