특집논문 (Special Paper)
방송공학회논문지 제30권 제4호, 2025년 7월 (JBE Vol.30, No.4, July 2025)
https://doi.org/10.5909/JBE.2025.30.4.546
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

생성형 AI 모델의 저작권 침해 방지 기법 연구 동향 : 스타일 모방 기법과 방지 기법을 중심으로

한 주 아[a], 박 주 원[a], 차 민 희[a], 조 성 인[a)‡]

Research Trends in Copyright Infringement Prevention Techniques for Generative AI Models: Focusing on Style Imitation and Prevention Methods

Jua Han[a], Joo Won Park[a], Min Hee Cha[a], and Sung In Cho[a)‡]

요 약

Stable Diffusion, DALL·E와 같은 텍스트 기반 이미지 생성 모델의 발전은 창의적인 시각 콘텐츠 제작의 새로운 가능성을 열었지만, 동시에 예술가의 고유한 화풍이 무단으로 모방되어 저작권 침해로 이어질 수 있다는 우려를 불러일으키고 있다. 이에 따라, 원작자의 독창적인 시각적 스타일이 생성 모델에 무분별하게 학습되는 것을 막기 위한 기술적 대응책인 '스타일 모방 방지 기법'의 필요성이 대두되고 있다. 본 논문은 먼저 스타일 모방 기법인 DreamBooth, Textual Inversion, Custom Diffusion, StyleDrop을 기술적 관점에서 정리하고, 이에 대응하는 Glaze, Mist, Anti-DreamBooth 등 최신 스타일 모방 방어 기법들의 구조와 작동 원리를 심층적으로 비교·분석한다. 또한 현존하는 스타일 모방 방어 기법들의 한계, 예를 들어 시각적 품질 저하와 모델 구조에 대한 종속성 등의 문제점을 명확히 지적하고, 비가시적 교란 요소 삽입 및 워터마크 기반 보호 방식 등 향후 기술적 발전의 방향성을 제시한다. 이러한 분석을 바탕으로, 본 논문은 생성형 AI 시대에 창작자의 권리를 보호하기 위한 기술적 프레임워크 구축의 필요성과 가능성을 조망한다.

Abstract

The advancement of text-to-image generative models such as Stable Diffusion and DALL·E has opened new possibilities for creative visual content production. However, these developments have simultaneously raised concerns about copyright infringement, as artists' unique styles can be imitated without consent. In response, there is a growing need for technical countermeasures—referred to as style mimicry protection methods—to prevent generative models from indiscriminately learning and reproducing original visual styles. This study first provides a technical overview of representative style mimicry methods including DreamBooth, Textual Inversion, Custom Diffusion, and StyleDrop. It then presents an in-depth comparative analysis of the structure and operational principles of recent defense techniques such as Glaze, Mist, and Anti-DreamBooth. Furthermore, the study identifies key limitations of current protection methods, such as degradation of visual quality and dependency on specific model architectures. It concludes by suggesting future research directions, including the application of imperceptible perturbations and watermark-based protection schemes. Ultimately, this paper highlights the necessity and potential of establishing a robust technical framework for protecting creators' rights in the era of generative AI.

Keyword : Text-to-image generation, Adversarial attack, Diffusion model, Style mimicry, Copyright protection

Ⅰ. 서 론

최근 몇 년간 텍스트를 조건으로 하는 이미지 생성(text-to-image generation) 모델의 발전은 디지털 콘텐츠 생태계에 급격한 변화를 가져왔다[24,25,26]. 특히 Stable diffusion[1], DALL·E[2]와 같은 모델은 단 몇 줄의 텍스트 프롬프트만으로도 사실적이거나 예술적인 이미지를 생성할 수 있는 능력을 갖추었다. 이는 예술, 디자인, 게임 등 다양한 분야에서 창작 도구로 활용되고 있다.

그러나 이러한 비약적인 기술의 발전은 저작권 침해 우려를 불러일으켰다[27,28]. 대부분의 대규모 생성 모델은 공개 웹사이트에서 수집한 이미지들을 바탕으로 학습된다. 이 과정에서 특정 작가의 화풍(style)이 함께 학습되기도 한다. 그 결과, 그림 1의 오른쪽과 같이 사용자들이 입력한 'a drawing of [작가 이름] style'과 같은 프롬프트만으로도 생성 모델은 해당 작가의 고유한 화풍을 매우 높은 유사도로 재현할 수 있다. 하지만 이는 창작자의 의사와 무관하게 그 정체성과 표현 양식이 무단으로 복제되는 결과를 초래할 수 있으며, 나아가 저작권 침해의 소지가 있다. 예컨대, ChatGPT[3]에 '지브리 스타일의 이미지 생성해줘'라는 프롬프트를 입력할 경우, 실제 지브리 스튜디오에서 만든 이미지가 아님에도 불구하고, 마치 지브리 스튜디오에서 제작한 것과 같은 이미지를 생성한다. Diffusion 기반 생성 모델[1]은 지브리 스튜디오가 제작한 이미지에서 추출한 색감, 구도, 인물 형태, 배경 연출 등을 학습하고, 이를 바탕으로 지브리 스튜디오의 스타일처럼 보이는 완전히 새로운 이미지를 창출할 수 있다. 이는 법적으로 저작권의 핵심 보호 대상인 '인간의 사상 또는 감정을 표현한 창작물' 자체가 무단 학습되어 재생산되는 것으로 해석될 수 있으며, 궁극적으로는 원작자의 경제적 이익 침해 문제로 이어질 수 있다.

생성형 AI로 인한 저작권 침해는 다양한 사례가 있다. 첫째로 작가 및 예술가의 저작권 침해 소송 사례가 있다. 여러 명의 시각 예술가들이 이미지 생성 AI (Stable Diffusion[1], Midjourney[29], DeviantArt[30] 등)를 상대로 저작권 침해 소송을 제기하였다. 이는 자신의 작품이 무단으로 AI 모델 학습에 사용되었고, 생성된 이미지가 원작과 유사하거나 스타일을 모방했다고 주장한다. 법원이 원고의 일부 주장을 받아들여, AI가 생성한 이미지가 원작과 실질적으로 유사할 경우 저작권 침해가 성립할 수 있음을 인정하였다[31]. 그리고 세계적인 이미지 에이전시인 Getty Images는 Stability AI가 수백만 장의 자사 이미지를 무단으로 학습 데이터로 사용했다며 대규모 저작권 소송을 제기하였다. 이 사건은 AI 모델 학습에 사용된 데이터의 출처와 허가 여부가 중요한 쟁점임을 보여준다[32].

둘째로 언론 및 콘텐츠 기업의 집단 소송 사례가 있다. 뉴욕 타임즈는 OpenAI가 자사 기사를 무단으로 학습 데이터로 활용했다고 주장하며 소송을 제기하였다. 이 사건은 AI가 생성한 텍스트가 원본 기사와 유사하거나 일부를 그대로 재현하는 등, 실제 저작권 침해 피해가 발생할 수 있음을 보여준다[33]. 유사하게 인도 최대 뉴스통신사 ANI 역시 OpenAI가 자사 뉴스를 무단으로 학습에 사용했다고 주장하며 소송을 제기했다. 이 사례는 비영어권 미디어도 AI 저작권 침해 문제에 적극 대응하고 있음을 보여준다.

셋째로 브랜드 및 캐릭터 스타일 침해 사례가 있다. Microsoft의 Bing AI 이미지 생성기가 '픽사 스타일' 이미지를 생성하면서, 실제 디즈니-픽사 로고와 유사한 디자인이 포함된 결과물이 생성되어 논란이 되었다. 이는 AI가 특정 브랜드나 캐릭터의 고유 스타일을 모방하여 상표권 및 저작권 침해 위험을 높인 대표적 사례이다[34].

넷째로 작가 스타일 모방 및 2차 저작물 문제가 있다. AI가 특정 작가의 이름을 프롬프트에 입력하면, 해당 작가의 고유한 화풍이나 스타일을 모방한 작품을 생성할 수 있다. 실제로 여러 아티스트 들이 자신의 이름이 프롬프트로 사용되어 모방 작품이 대량 생성되는 현상에 대해 항의하고 있다. 이러한 피해는 작가의 경제적, 명예적 손실로 이어질 수 있다[35].

마지막으로 AI 모델의 훈련 데이터 추출 공격 및 저작권

a) 동국대학교 컴퓨터·AI 학부(Department of AI Software convergence, Dongguk University)
‡ Corresponding Author : 조성인(Sung In Cho)
E-mail: csi2267@dongguk.edu
Tel: +82-2-2250-3336
ORCID: https://orcid.org/0000-0003-4251-7131
※ 이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 한국연구재단 지원 (RS-2023-00208763) 정보통신기획평가원 - 학·석사연계ICT핵심인재양성 지원을 받아 수행된 연구임 (IITP-2025-RS-2023-00260248)
· Manuscript May 19, 2025; Revised July 8, 2025; Accepted July 9, 2025.

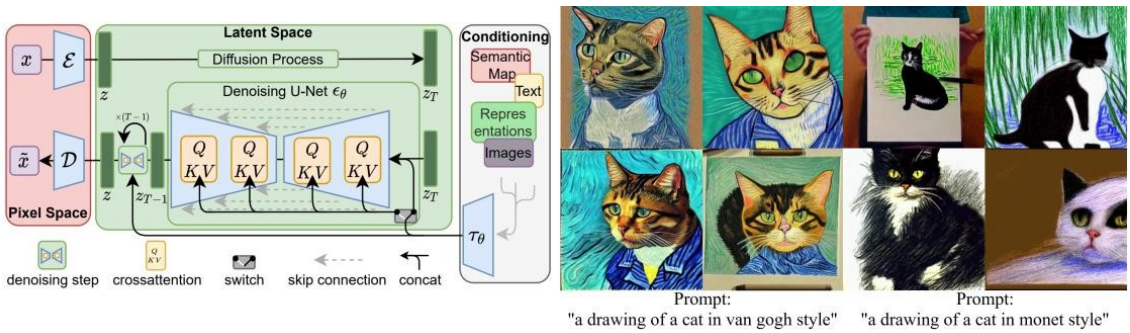

그림 1. Latent Diffusion Model (LDM)[1]의 구조와 입력 프롬프트에 따른 결과 영상의 스타일 변화 예시
Fig. 1. Architecture of the LDM[1] and generated images conditioned on different prompts

노출 사례가 있다. 연구자들은 AI 모델에 특정 프롬프트를 입력해 훈련 데이터에 포함된 저작권 이미지를 거의 그대로 복원해내는 사례를 다수 확인하였다. 이는 단순한 스타일 모방을 넘어, 원본 데이터의 직접적인 유출과 저작권 침해로 이어질 수 있음을 보여준다[36].

이러한 text-to-image 생성 모델에 의해 발생할 수 있는 예술가 고유 화풍의 무단 모방 및 저작권 침해 문제에 대응하기 위해, 해당 스타일의 학습 및 재현을 방해하는 기술적 방어 기법 (style mimicry protection)의 도입이 활발히 논의되고 있다. 본 논문에서는 스타일 모방 가능성이 확인된 주요 생성 모델들을 분석하고, 예술가의 시각적 저작물을 보호하기 위한 효과적인 기술적 방어 메커니즘의 원리와 구현 방식을 고찰한다.

II. 배경 지식

1. Text-to-image generative diffusion models

Latent Diffusion Model (LDM)[1]은 대표적인 text-to-image 생성 모델로, 이는 고차원의 이미지 픽셀 공간 대신 더 낮은 차원의 잠재 공간 (latent space)에서 확산 과정을 수행함으로써 semantic한 정보 생성을 가능하게 한다.

LDM[1]은 크게 두 가지 주요 구성 요소로 이루어진다. 먼저, autoencoder는 입력 이미지를 latent space로 인코딩한 후 다시 원본 이미지로 디코딩하는 구조를 통해 latent representation을 학습한다. 다음으로, denoiser는 확산 과정 중 노이즈가 점진적으로 추가된 latent representation으로부터 원래의 latent variable을 복원하는 역할을 한다. 이 denoiser는 텍스트 프롬프트를 조건으로 활용하여 학습되며, 이를 통해 그림 1의 오른쪽과 같이 다양한 문장에 대응하는 이미지 생성이 가능해진다.

특히, 이러한 조건부 학습은 사전 학습된 text-image 멀티모달 모델인 Contrastive Language Image Pretraining (CLIP) 모델[4]과의 결합을 통해 더욱 강화된다. CLIP[4]은 텍스트와 이미지 간의 semantic alignment를 학습한 모델로, 텍스트 프롬프트를 벡터로 인코딩한 뒤, 이를 이미지 생성 과정에서 조건으로 제공한다.

그림 1은 직접 생성한 이미지이며 좌측에서 확인할 수 있듯, 텍스트 인코더 τ_θ는 입력 텍스트를 임베딩하여 U-Net[5] 내부의 cross-attention 메커니즘에 조건 정보로 사용한다. 이때, style을 표현한 텍스트가 입력될 때 이는 이미지의 스타일 구성에 직접적인 영향을 미칠 수 있다. 그림 1의 우측에서는 입력 프롬프트에 따른 스타일 변화 예시를 보여준다. 이러한 설계 덕분에 LDM[1] 기반 모델은 자연어 기반의 정밀하고 의미 있는 이미지 생성을 가능하게 한다.

2. 스타일 모방 기법 (Style mimicry methods)

스타일 모방 기법은 생성 모델을 활용해 특정 예술 스타일과 일치하는 이미지를 생성하는 기법[6,7,8,9]으로, 최신 기법들은 특정 예술 스타일을 나타내는 소수의 참조 이미지 (reference images)만으로도 스타일을 효과적으로 학습하고, 이를 다양한 대상 및 조건에 적용할 수 있도록 발전하고

있다. 이러한 스타일 모방 기법으로는 대표적으로 DreamBooth[6], Textual Inversion[7], Custom Diffusion[8], 그리고 StyleDrop[9]이 있다.

먼저, DreamBooth[6]는 소량의 이미지-텍스트 쌍을 이용해 사전 학습된 모델을 미세 조정 (fine-tuning)함으로써, 특정 주제 (subject)나 화풍 (style)을 반영하는 개인화된 이미지 생성을 가능하게 하는 기법이다. Textual Inversion[7]은 하나의 개념 또는 스타일을 단일 토큰 벡터로 임베딩하고, 이를 텍스트 프롬프트에 삽입하여 해당 스타일에 맞는 이미지를 생성할 수 있도록 한다. Custom Diffusion[8]은 다중 개념을 동시에 학습 가능한 구조를 제공하는 동시에, 스타일 간의 독립적인 제어를 지원한다. 마지막으로, StyleDrop[9]은 프롬프트에 포함된 스타일 키워드를 기반으로 이미지의 시각적 속성을 조절할 수 있도록 설계되어 있다. 그러나 이러한 스타일 모방 기술의 발전에 대해, 일부 예술가들은 이러한 스타일 모방 기술이 무단으로 자신의 작품을 복제하거나, 여러 방면으로 악용될 수 있다는 점에 대한 우려를 표하고 있다 (Heikkila, 2022)[10].

3. 스타일 모방 방지 기술 (Style mimicry protection methods)

앞서 설명한 바와 같이, 스타일 모방 기술의 발전에 대한 예술가들의 우려를 해소하기 위한 방안으로써 이미지에 미세한 교란 (perturbation)을 삽입하여 스타일이 모방되는 것을 방해하는 스타일 모방 방지 기술들이 제안되고 있다. 이러한 기법들은 예술가가 작품을 공개하기 전에 본인의 작품에 적용할 수 있어, 사후 대응이 어려운 저작권 침해 문제에 선제적으로 대응할 수 있도록 한다.

대표적인 스타일 모방 방지 기술로는 Glaze[11], Mist[12], 그리고 Anti-DreamBooth[13]가 있다. Glaze[11]와 Mist[12]는 LDM의 인코더를 주요 공격 표적으로 삼는 기술로, 이미지의 latent representation에 교란을 가함으로써 latent variable을 디코딩 했을 때 모방하고자 하는 스타일이 아닌 방향으로 재구성되도록 유도한다.

반면, Anti-DreamBooth[13]는 LDM의 denoiser를 주요 공격 표적으로 삼는 기술이다. 이 기법은 latent space 내에서 perturbation을 추가해 denoiser가 해당 이미지를 조건부로 재구성하는 과정을 교란시킨다. 구체적으로는, 교란된 이미지가 주어진 텍스트 조건에 대해 정확한 latent representation을 복원하지 못하도록 유도함으로써, 스타일이 학습되지 않도록 유도한다.

Ⅲ. 스타일 모방 기법의 구체화 및 이해

특정 스타일을 학습하고 재현하는 생성 모델들은 각기 다른 다양한 방식으로 작동한다. 이에 대한 자세한 분석을

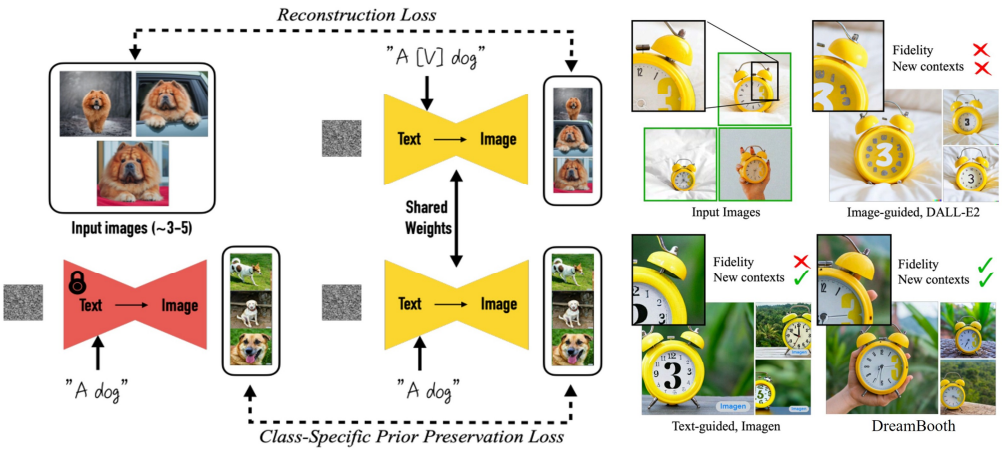

그림 2. DreamBooth[6]의 전체 구조도
Fig. 2. Overall architecture of DreamBooth[6]

위해, 본 장에서는 대표적인 스타일 모방 기법으로 알려진 DreamBooth[6], Textual Inversion[7], Custom diffusion[8], StyleDrop[9]을 중심으로, 각 기법이 어떤 방식으로 스타일 정보를 주입하고 모방하는지 비교한다.

1. DreamBooth[6]

DreamBooth[6]는 Google research와 Boston university가 공동으로 개발한 개인화 text-to-image generation 기법으로, 그림 2와 같은 구조를 갖고 있다. 소수의 이미지 샘플만으로도 특정 객체나 스타일을 모델이 생성할 수 있다는 특징을 가지고 있다. 사용자는 3~5장의 스타일 이미지와 함께 이를 설명하는 텍스트를 제공하며, 해당 데이터는 사전 학습된 stable diffusion 모델의 U-Net[5] 네트워크를 fine-tuning하는 데 사용된다. 이 과정에서 "[V]"와 같은 사용자 정의 토큰을 프롬프트에 삽입함으로써, 새로운 개체나 스타일이 모델 내 latent 표현 공간에 고정(anchor)된다.

이렇듯 이미지를 설명하는 텍스트를 활용한 DreamBooth[6]의 구조는 특정 작가의 색채, 질감, 구도 등 복합적인 시각 특성을 이미지 샘플과 유사하게 재현할 수 있도록 하며, 이를 통해 높은 수준의 스타일 모방을 가능하게 한다. 이는 작가의 고유한 시각적 표현이 단 몇 장의 샘플만으로도 모델에 내재화될 수 있음을 의미한다.

2. Textual Inversion[7]

Textual Inversion[7]은 DreamBooth[6]와 달리 diffusion 모델의 파라미터를 조정하지 않고, 단일 텍스트 임베딩 벡터를 학습하는 방식이다. 그림 3에서와 같이, 사용자는 같은 스타일을 공유하는 이미지들과 "[S_*]"와 같은 새로운 pseudo-token을 정의한다. 이후, 학습 과정에서 해당 토큰이 주어진 이미지들의 특성을 반영하도록 CLIP[4] 임베딩 공간 내에서 벡터가 조정된다.

이러한 방식은 단일 임베딩 벡터만을 학습하기 때문에 상대적으로 연산량이 적고, 기존 모델 구조를 변경하지 않아 안정적인 학습이 가능하다는 장점이 있다. 그리고 이 장점을 바탕으로, Textual Inversion[7]은 단순한 스타일 모방이나 키워드 기반의 표현 확장에서 유용하게 활용된다. 그러나 스타일 재현의 정밀도는 DremaBooth[6]와 같은 파라미터 조정 방식에 비해 낮으며, 복잡한 스타일의 경우 효과가 제한적일 수 있다는 단점이 존재한다.

3. Custom Diffusion[8]

그림 4의 Custom Diffusion[8]은 DreamBooth[6]와 Textual Inversion[7]의 절충형 접근 기법으로, 모델의 일부 파라미터만 선택적으로 학습한다. 이를 위해 U-Net[5]의 중간 레이어 일부를 fine-tuning하며, 필요에 따라 Low-Rank Adaptation

그림 3. Textual Inversion[7]의 pseudo-token 예시
Fig. 3. Example of pseudo-tokens in Textual Inversion[7]

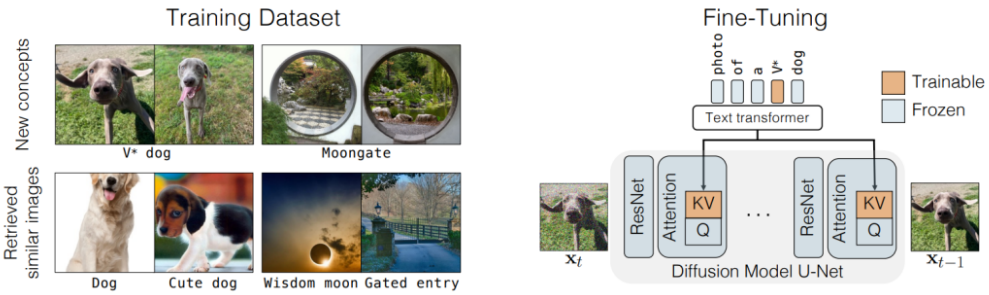

그림 4. Custom diffusion[8]의 fine-tuning 과정
Fig. 4. Fine-tuning process of Custom Diffusion[8]

of large language models (LoRA)[14] 등을 함께 적용하기도 한다. 이 기법은 전체 모델을 수정하지 않으면서도 앞선 기법들과 비교했을 때 상대적으로 높은 수준의 스타일 반영이 가능하다는 점에서 실용적이라는 장점이 있다.

Custom Diffusion[8]은 제한된 자원 (GPU memory, 시간 등) 내에서도 학습이 가능하므로, 개인 사용자가 사용하기에 용이하다. 그림 4는 이러한 selective fine-tuning 구조를 시각적으로 설명하고 있으며, 이는 해당 기법이 효율적인 학습을 지향함을 보여준다.

4. StyleDrop[9]

StyleDrop[9]은 Google이 제안한 사실적인 이미지를 생성하는 모델인 imagen[15]에 적용한 스타일 적응 시스템으로, 하나의 스타일을 다양한 콘텐츠에 적용할 수 있도록 설계되었다. 이 기법은 대략 30장 이상의 스타일 이미지로부터 스타일 임베딩을 추출하고, 이를 생성 모델의 조건(condition)으로 삽입하여 다양한 입력 프롬프트에 일관되게 스타일을 적용한다.

또한, 이 기법은 단일 개체 또는 캐릭터뿐만 아니라 보다 복합적인 시각 표현 (예: 텍스트 타이포그래피, 인쇄체 등)에 대한 스타일링에도 효과적이다. 그림 5는 논문에서 발췌한 것으로 reference image가 주어졌을 때 StyleDrop, DreamBooth, Textual inversion의 스타일 모방 성능을 비교한 실험으로, StyleDrop[9]은 가장 reference image와 유사한 스타일로 이미지를 생성한 반면, Textual Inversion[7]은 색감과 질감 등 모든 스타일 요소 측면에서 유사도가 떨어짐을 확인할 수 있다.

Ⅳ. 스타일 모방 방어 기법 구체화 및 한계

본 장에서는 2.3장에서 서술한 내용을 바탕으로, Glaze[11], Mist[12], Anti-DreamBooth[13]와 같은 대표적인 스타일 모방 방어 기법들에 대해 각 기법이 어떤 방식으로 스타일 모방을 방어하는지 비교한다.

1. 인코더 기반 스타일 방어 기법

최근 스타일 모방 공격 (style mimicry attack)이 증가함

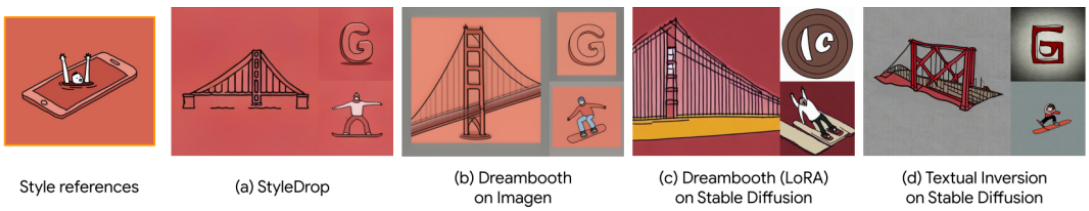

그림 5. Style reference image를 기반으로 한 다양한 text-to-image 생성 기법의 스타일 모방 성능 비교
Fig. 5. Comparative analysis of style mimicry in text-to-image models using style reference images

에 따라, 이미지의 latent representation을 교란하여 저작권 스타일로 학습되는 것을 방지하는 인코더 방어 기법이 제안되고 있다. 이 기법은 일반적으로 이미지 x에 소량의 적대적 교란 (perturbation) δ_x를 추가함으로써, 인코더 ε_ϕ가 해당 이미지를 잠재 표현으로 인코딩할 때 타겟 잠재 표현 $\mathbf{t_x}$으로 매핑되도록 유도한다. 이때의 목적함수는 다음과 같이 정의된다:

$$\min_{\boldsymbol{\delta_x}} d_{Lat}(\varepsilon_\phi(x+\boldsymbol{\delta_x}), \boldsymbol{t_x}) \text{ subject to } d_{Img}(x+\boldsymbol{\delta_x}, x) \leq p. \quad (1)$$

여기서 d_{Lat}은 latent space 상에서 이미지 간 차이를 측정하는 평가 방식이며, d_{Img}는 시각적으로 인지 가능한 표현에서의 이미지 간 차이를 측정하는 평가 방식이다. 위 제약 조건은 이미지의 외향적 왜곡이 인간이 인지할 수 있는 수준 이하로 유지되도록 제한한다.

1.1 Glaze[11]

Glaze[11]는 위 목적식 (1)을 구현한 대표적인 인코더 방어 기법으로, 이미지의 latent representation이 학습하고자 하는 스타일이 아닌 타겟 스타일로 인코딩되도록 유도하는 방식이다. 이를 위해 Glaze는 먼저 방어 대상 스타일 S 대신 학습자가 모방하도록 유도한 적대적 타겟 스타일 S_{adv}를 선택한다. 이후 적대적 타겟 스타일 S_{adv}을 방어 대상 스타일 S를 가지는 원본 이미지 x에 적용한 xS_{adv}를 생성하고, 생성한 latent representation $\varepsilon_\phi(xS_{adv})$를 방어 대상 이미지의 타겟 표현 $\mathbf{t_x}$로 설정한다.

타겟 스타일 S_{adv}는 총 50개의 스타일 집합으로부터 선택되며, CLIP[4] 임베딩 기반 거리 계산을 통해 원본 스타일과 의미론적으로 유사하지 않은 스타일을 선별한다. 구체적으로, 이미지 X의 평균 임베딩과 각 스타일 프롬프트 Ps'간의 거리를 측정하여, 이 중 상위 50%~70%에 해당하는 스타일을 무작위로 샘플링한다. 이러한 방식은 학습자가 의도치 않은 스타일을 학습하도록 유도함으로써 방어하고자 하는 스타일의 모방을 방지한다. Glaze[11]는 다음 목적함수를 최소화함으로써 교란 $\boldsymbol{\delta_x}$를 최적화한다:

$$\min_{\boldsymbol{\delta_x}} \|\varepsilon_\phi(x+\boldsymbol{\delta_x}), \boldsymbol{t_x}\|_2^2 + \alpha \cdot max\,(LPIPS(x+\boldsymbol{\delta_x}, x) - p, 0) \quad (2)$$

여기서 첫 번째 항은 latent representation과 타켓 표현 $\mathbf{t_x}$의 유사도를 측정하고, 두 번째 항은 교란을 적용한 후의 이미지 $(x+\boldsymbol{\delta_x})$가 원본 이미지인 x와 시각적으로 얼마나 유사한지 평가한다. LPIPIS[16]는 인간의 지각을 기반으로 이미지의 유사도를 측정하는 평가 기법으로, 이미지에 교란을 적용했을 때의 시각적인 차이가 임계값 p를 넘지 않도록 제한하며, α는 두 항간의 균형을 조절하는 하이퍼파라미터이다.

1.2 Mist[12]

Mist[12]는 예술가의 스타일 무단 모방을 방지하기 위한 인코더 방어 기법이다. 본 기법은 Projected Gradient Descent (PGD)[17]를 이용하여, 아티스트 이미지들의 latent representation을 특정 무관한 타겟 이미지 $\boldsymbol{z_{target}}$ 표현에 가까워지도록 하는 perturbation을 최적화한다. 구체적으로는 다음의 목적함수를 최소화한다:

$$\delta := \arg\min_{\boldsymbol{\delta}} \|\varepsilon_\phi(\boldsymbol{z_{target}}) - \varepsilon_\phi(x_i+\boldsymbol{\delta_i})\|_2^2. \quad (3)$$

여기서 x_i는 style image, δ_i는 x_i에 적용되는 perturbation이다. 그리고 $\boldsymbol{z_{target}}$은 style image와 무관한 타겟 이미지이고 ε_ϕ는 인코더이다. 이 목적함수는 style image들의 representation을 타겟 representation에 강제로 유사하게 만듦으로써, style image 간의 일관성과 분산을 무너뜨리는 효과를 유도한다. 이를 통해 모델은 아티스트의 고유한 스타일을 제대로 학습하지 못하게 유도하는 perturbation을 학습하게 된다.

2. Denoiser 기반 스타일 방어 기법

Denoiser 기반 스타일 방어 기법은 스타일 모방 여부를 정량화하기 위해 reconstruction 에러를 활용하며, 이 값이 클수록 denoiser가 원본 이미지를 정확히 복원하지 못해 스타일 학습이 실패한다고 가정한다. 이를 토대로 denoiser ϵ_θ가 출력하는 reconstruction 예측 값과 실제 노이즈 간의 오차를 모방률로 간주하여, 이를 증가시키는 방향으로 교란을 설계한다. 이 접근법은 에러 값이 클수록 이미지 re-

construction 품질이 떨어지고, 결과적으로 스타일 특유의 표현이 제대로 재현된다는 가정에 기반한다.

2.1 Anti-DreamBooth[13]

Anti-DreamBooth[13]는 DreamBooth[6] 기반의 text-to-image 생성 모델이 악의적으로 사용되어 특정 개인의 이미지가 가지는 외형이나 스타일이 무단으로 재현되는 것을 방지하기 위해 제안된 대표적인 denoiser 기반 스타일 방어 기법이다. 그림 6은 기법의 전체 구조도로, 이 기법은 diffusion 모델의 denoising 과정에서 예측되는 노이즈 값과 실제 노이즈 간의 오차를 증가시키는 방식으로 perturbation을 설계한다. 이를 통해 diffusion 모델이 원본 이미지를 정확하게 복원하지 못하도록 유도하며, 결과적으로 DreamBooth[6] 모델의 개인 스타일 학습 및 재현 능력을 저하시킨다.

Anti-DramBooth[13]는 공격자가 사용할 수 있는 DreamBooth[6] 모델의 내부를 직접적으로 알 수 없다는 현실적인 제약을 고려하여, surrogate model (대리 모델)을 도입한다. Surrogate 모델이란, 실제 스타일 모방 기법 모델과 유사한 학습 구조를 갖도록 훈련된 대체 모델로, 이를 이용해 교란을 간접적으로 최적화한다. Anti-DreamBooth[13]는 이러한 surrogate 모델을 이용해 노이즈 예측 오차를 최대화하는 방향으로 perturbation을 학습함으로써, 실제 공격자 모델에서도 유사한 효과를 얻을 수 있다고 기대한다.

이러한 구조 아래에서 Anti-DreamBooth는 두 가지 주요 변형 기법을 제안한다. 첫 번째는 Fully-trained Surrogate Model Guidance (FSMG)로, 원본 이미지에 대해 사전에 DreamBooth[6] 파인튜닝을 완료한 고정된 surrogate 모델을 사용한다. 이후에는 해당 모델을 참조하여 각 방어 대상 이미지에 대한 적대적 교란을 설계한다. 이 방식은 효율적이라는 장점이 있지만, surrogate 모델과 스타일 모방 기법 모델 간 학습 설정 또는 데이터 차이로 인하여 (예: 다른 reference 이미지로 학습되어) 방어 성능이 저하될 수 있다는 단점이 존재한다.

두 번째는 Alternating Surrogate and Perturbation Learning (ASPL)으로, surrogate 모델을 고정하지 않고 반복적으로 갱신하는 동적 구조를 갖는다. 각 반복 (iteration)마다 surrogate 모델을 클린 이미지로 파인튜닝한 후, 이 모델을 기반으로 현재 perturbed image에 대해 PGD를 적용하여 교란을 최적화한다. 이후 이 perturbed images로 surrogate 모델을 다시 학습하는 과정을 반복하며, 이를 통해 실제 공격자의 DreamBooth[6] 학습 과정을 더 정밀하게 모사할 수 있다.

그림 6은 Anti-DreamBooth의 두 가지 변형 방식인 FSMG와 ASPL의 구조적 차이를 구체적으로 보여준다. 두

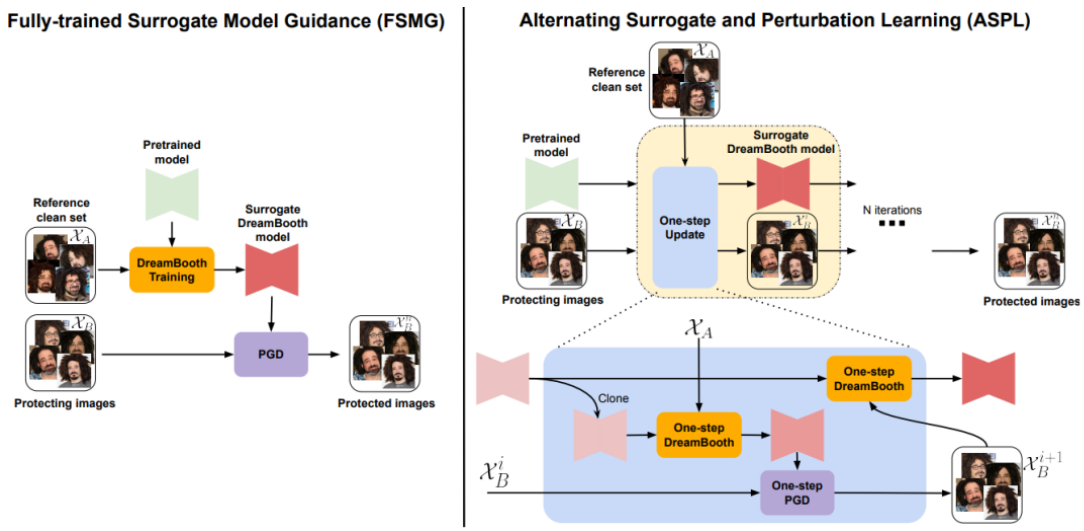

그림 6. Anti-DreamBooth[13]의 전체 구조도
Fig. 6. Overall architecture of Anti-DreamBooth[13]

그림 7. 좌측: 원본 이미지 / 우측: perturbation 삽입 이미지
Fig. 7. Left: Original image / Right: Image with inserted perturbation

방식 모두 PGD를 활용하여 적대적 교란 δ를 생성하며, 이는 surrogate model의 reconstruction 손실함수인 L_{cond}를 최대화하는 방향으로 최적화된다. FSMG는 고정된 surrogate 모델을 기반으로 모든 방어 이미지에 대해 동일한 모델을 참조하여 perturbation을 생성하는 반면, ASPL은 surrogate 모델을 각 반복마다 업데이트하여, perturbation 생성과 모델 학습을 교차적으로 수행한다.

표 1은 CelebA-HQ[18]와 VGGFace2[19] 두 얼굴 데이터셋을 대상으로, 제안된 네 가지 방어 기법 (FSMG, ASPL, T-FSMG, T-ASPL)의 방어 성능을 비교한 결과이다. 이때 FSMG와 ASPL은 비타겟 방식이며, T-FSMG와 T-ASPL은 특정 방향으로 교란을 유도하는 타겟 방식의 변형 기법이다. 이때 타겟 이미지는 클린 이미지 (reference image)와는 무관한, 사용자가 임의로 지정한 별도의 이미지이다.

실험 평가 지표로는 Fake-to-Detection Failure Rate (FDFR)[20], Identity Similarity Metric (ISM)[21], Style Error Rate – Face Quality Assessment (SER-FQA)[22], Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE)[23]을 사용하였다. 먼저 FDFR는 방어된 이미지가 탐지 모델에 의해 원본 이미지로 오인되는 비율로, 값이 클수록 방어 이미지가 원본과 유사한 외형을 유지하고 있음을 의미한다. ISM은 원본 이미지와 생성 이미지 간의 신원 유사도를 나타내며, 일반적으로 얼굴 임베딩 간의 코사인 유사도 등을 통해 계산된다. ISM 값이 작을수록 신원 정보가 효과적으로 은폐되었음을 의미한다. SER-FQA는 방어 이미지의 얼굴 품질 저하 정도를 평가하는 지표로, 값이 작

표 1. Anti-DreamBooth[13]의 FSMG 및 ASPL 방어 성능을 비교한 결과
Table 1. Comparison of defense performance between FSMG and ASPL in Anti-DreamBooth[13]

Dataset	Method	"a photo of sks person"				"a dslr portrait of sks person"			
		FDFR ↑	ISM ↓	SER-FQA ↓	BRI-SQUE ↑	FDFR ↑	ISM ↓	SER-FQA ↓	BRI-SQUE ↑
VGG-Face2[18]	No Defense	0.07	0.63	0.73	15.61	0.21	0.48	0.71	9.64
	FSMG	0.56	0.33	0.31	36.61	0.62	0.29	0.37	38.22
	ASPL	0.63	0.33	0.31	36.42	0.76	0.28	0.30	39.00
	T-FSMG	0.07	0.58	0.74	36.42	0.28	0.44	0.71	17.29
	T-ASPL	0.07	0.57	0.72	15.36	0.39	0.44	0.70	20.06
Celeb-A-HQ[19]	No Defense	0.10	0.68	0.72	17.06	0.26	0.44	0.72	7.30
	FSMG	0.34	0.48	0.56	36.13	0.35	0.36	0.66	33.60
	ASPL	0.31	0.50	0.55	38.57	0.34	0.39	0.63	34.89
	T-FSMG	0.06	0.64	0.73	25.75	0.24	0.45	0.73	8.04
	T-ASPL	0.06	0.64	0.73	20.58	0.26	0.46	0.72	5.36

을수록 스타일 손실이 적고 이미지 품질이 우수함을 의미한다. 마지막으로, BRI-SQUE는 참조 이미지 없이 이미지 품질을 정량적으로 평가하는 지표로, 값이 클수록 구조적 왜곡이 적고 시각적으로 자연스러운 이미지임을 의미한다.

실험에는 학습 시 사용된 프롬프트 (“a photo of sks person”)와 학습 과정에서 사용되지 않은 새로운 프롬프트 (“a dslr portrait of sks person”)가 모두 포함되었다. 그 결과 ASPL이 모든 지표에서 가장 안정적이고 우수한 방어 성능을 보였다. 반면, 타겟 방식 (T-FSMG, T-ASPL)은 비타겟 방식에 비해 오히려 방어 성능이 낮은 경향을 보였는데, 이는 노이즈가 특정 방향으로 수렴하면서 전체적인 파인튜닝 방해 효과가 분산되었기 때문으로 해석된다. 이러한 결과는 ASPL이 DreamBooth[6] 학습 과정을 가장 현실적으로 반영하고 있으며, 강건한 일반화 성능을 갖춘 효과적인 방어 기법임을 시사한다.

3. Watermark-embedded Adversarial Examples for Copyright Protection against Diffusion Models[37]

이 논문은 diffusion 모델에 의한 무단 창작물 모방 및 이로 인한 저작권 침해 문제에 대응하기 위한 새로운 프레임워크를 제안한다. 기존 워터마킹 기법의 경우 이미지에 보이지 않는 메시지를 삽입하여 저작권을 식별하는 방식이지만, 주로 diffusion 모델 자체의 저작권을 보호하거나 생성된 이미지와 자연 이미지를 구별하는 데 중점을 둔다. 또한, 생성 과정에 워터마크를 삽입하려면 diffusion 모델을 재훈련하거나 fine-tuning 해야하며, 워터마크 추출을 위한 후처리 과정이 필요하다. 또한, Mist[12]와 같은 기존 적대적 예제 기반 방법은 diffusion 모델의 재훈련이나 생성된 이미지에 대한 후처리가 필요 없다는 장점이 있지만 저작권 관련 정보가 내장되지 않아 원본 이미지의 저작권 추적이 불가능하다. 또한 각 적대적 예제를 확산 모델에 대해 개별적으로 최적화해야 하므로 생성 시간이 오래 소요된다. 본 논문은 이러한 기존 방법의 한계를 해결하기 위해 워터마크가 내장된 적대적 예제 생성 프레임워크를 제안한다.

핵심 아이디어는 개인 워터마크를 적대적 예제 생성에 직접 포함하며, diffusion 모델이 워터마크가 선명하게 보이는 이미지를 생성하도록 하며 동시에 perturbation도 추가될 수 있도록 한다. 이는 저작권 침해를 경고하는 더 직관적인 방법을 제공한다. 조건부 GAN 아키텍처를 기반으로 한 생성기를 훈련하여 워터마크가 내장된 적대적 예제를 생성한다. 생성기는 원본 이미지와 워터마크를 입력으로 받아 미묘한 perturbation을 생성한다. 훈련된 생성기는 매우 빠르게 (이미지 당 0.2초) 적대적 예제를 생성할 수 있다. 이는 단 5-10개의 샘플만으로 개인 워터마크를 위한 생성기를 2~3분 내에 훈련할 수 있어 효율적이다.

손실함수는 적대적 손실, GAN 손실, 가중치 교란 손실로 총 3가지를 사용한다. 적대적 손실은 확산 모델이 생성한 이미지가 워터마크에 가깝게 생성되도록 한다. 특히 LDM [1]의 잠재 표현을 기반으로 손실을 설계하여 다양한 작업 및 모델에 대한 전이성을 높인다. GAN 손실은 생성된 적대적 예제가 원본 이미지와 시각적으로 유사하도록 강제하여 교란이 사람의 눈에 보이지 않게 한다. 가중치 교란 손실은 perturbation의 크기를 제한하고, 특히 워터마크 영역의 교란이 눈에 띄지 않도록 더 큰 가중치를 부여하여 워터마크가 적대적 예제 자체에서는 보이지 않게 한다.

제안된 방법은 반복 최적화 방식과 달리 훈련된 생성기를 직접 사용하므로 이미지 당 0.2초로 훨씬 빠르게 적대적 예제를 생성한다. 또한 생성된 적대적 예제는 사람의 눈에 거의 보이지 않는 미묘한 perturbation만을 도입한다. 이 예제는 JPEG 압축, randomized smoothing 등 다양한 공격이 적용된 후에도 워터마크가 가시적으로 남아있는 강건성을 가지며 훈련에 사용되지 않는 다른 생성 모델 (예: Stable Diffusion, Runway AI)에 대해서도 성공적으로 방어가 전이됨을 입증하여, 다양한 이미지 생성 모델로부터 저작권 침해를 방지할 수 있음을 보였다.

4. I2VGuard: Safeguarding Images against Misuse in Diffusion-based Image-to-Video Models[38]

I2VGuard[38]는 이미지 기반 비디오 생성 모델에서 발생할 수 있는 저작권 침해 문제를 해결하기 위한 기법이다. 해당 기법은 이미지에 perturbation을 추가하여 적대적 예제를 생성한 후 이를 기반으로 생성되는 비디오의 품질을

의도적으로 낮추는 adversarial attack 방법이다. 공간적, 시간적, 확산 단계 총 3가지의 공격 스텝으로 이루어져 있다. 공간적 공격은 개별 프레임의 이미지 품질을 저하시킨다. 이때 encoder를 조작하여 이미지의 latent vector를 저품질로 변형한다.

구체적으로 공간적 공격은 개별 프레임의 품질을 저하시키고 부자연스러운 텍스처를 유발하여 공간적 품질을 떨어뜨린다. 예를 들어, 이미지 속 개 주변에 비정상적인 텍스처를 생성할 수 있다. 시간적 공격은 프레임 간 일관성을 직접적으로 저해하여 부자연스러운 움직임 변화를 유도한다. 예를 들어, 개의 머리 부분에 흐릿함을 유발하여 움직임의 불일치를 나타낼 수 있다. 확산 공격은 공간적, 시간적 측면에 모두 동시에 영향을 미친다. 이 세 가지 공격 방식은 각각 다른 영향을 미치며, 모든 공격이 함께 적용될 때 최적의 보호 효과를 달성한다. 특히, 공간적 공격 없이는 보호 효과가 상대적으로 약해지며, 추가된 노이즈가 인코더에 의해 부분적으로 필터링 될 수 있다.

실험에 따르면 I2VGuard[38]는 JPEG 압축 (60% 품질)과 가우시안 노이즈 추가에도 불구하고 보호 효과는 실질적으로 유지되어 움직임 생성을 계속 방해하였다. 또한 NAFNet[39]과 같은 denoising 모듈이 훈련 파이프라인에 통합되어 있더라도, I2VGuard[38]는 효과적으로 작동하여 비디오 생성 모델에 영향을 미쳤다. 그리고 해당 기법이 적용된 이미지에서 생성된 비디오는 원본 이미지에서 생성된 비디오에 비해 주제 일관성, 움직임 부드러움, 그리고 이미지 품질 등 모든 평가 지표에서 수치가 감소했다. 이는 해당 기법이 생성된 비디오의 공간적 및 시각적 품질을 손상시키는 데 효과적임을 보여준다.

대부분의 상황에서 효과적이지만, 원본 비디오에 중요한 움직임이 부족할 경우 해당 기법은 의미 있는 방해를 도입하기 어렵다. 이러한 경우, 보호된 버전의 비디오는 움직임이 정지되거나 원본 동역학을 재현 또는 방해하지 못할 수 있다. 때로는 해당 기법이 움직임을 너무 심하게 손상시켜 장면이 정지된 것처럼 보일 수 있다. 하지만 비록 움직임 방해에 실패하는 경우에도, 공간적 품질 공격은 일반적으로 생성된 콘텐츠의 시각적 충실도를 저하시키는 데 효과적이다. 이는 비정상적인 텍스처와 아티팩트를 유발하여 전반적인 시각적 품질을 떨어뜨린다.

5. 스타일 모방 방지 기법의 한계

현재까지 제안된 스타일 모방 방지 기법들은 이미지 생성 모델이 저작권 스타일을 무단으로 학습하는 것을 방지하기 위한 효과적인 시도이지만, 몇 가지 본질적인 한계를 내포하고 있다.

첫째, 시각적 품질 저하 문제이다. 대부분의 방어 기법은 원본 이미지에 일정 수준의 perturbation을 인위적으로 삽입하여 스타일 모방 기법의 학습 효율을 저하시키고, 그 결과로 스타일 모방을 억제한다. 그러나 이러한 방식은 이미지의 시각적 품질을 훼손시키며, 때로는 눈에 띄는 잡음이

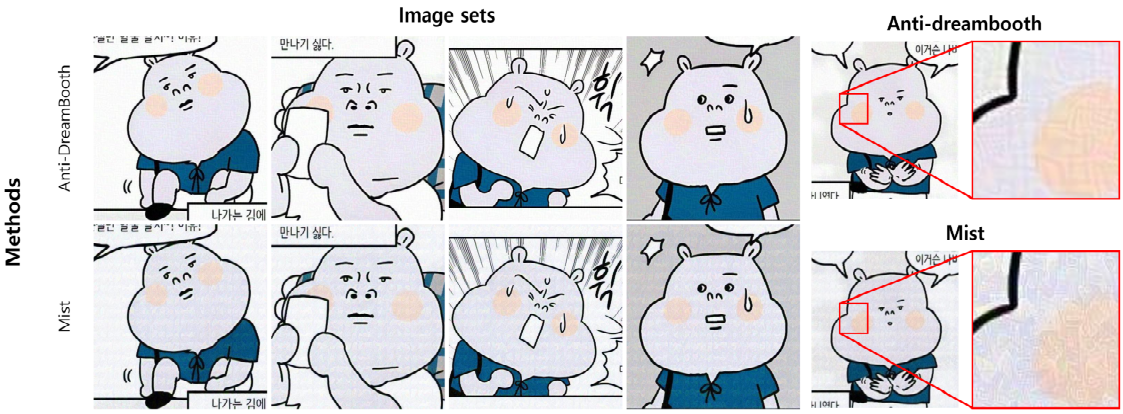

그림 8. 웹툰 이미지를 활용한 기존 기법 이미지 품질 실험 결과
Fig. 8. Experimental results on image quality using webtoon images in existing methods

나 왜곡을 유발할 수 있다. 그림 7의 좌측은 원본 이미지, 우측은 원본 이미지에 perturbation을 삽입한 이미지로, 우측 이미지에서 시각적 품질 저하가 발생했음을 확인할 수 있다. 예술가나 디자이너와 같은 창작자들은 자신의 포트폴리오나 SNS 게시물에 고품질의 이미지를 게시하는 것이 중요한데, 이미지가 손상된 상태로 게시되어야 한다면 해당 방어 기법의 실사용 가능성은 크게 저하된다.

둘째, 모델 구조 및 버전 종속성 문제이다. 대부분의 방어 기법은 특정한 이미지 생성 모델 구조, 예컨대 stable diffusion v1.4와 같은 특정 버전에 맞춰 설계되고 최적화되어 있다. 이러한 기법은 해당 버전에서는 효과적으로 작동할 수 있지만, 구조적으로 상이한 다른 버전의 생성 모델에서는 방어 효과가 현저히 감소하거나 무효화된다. 이는 공격자가 방어된 이미지를 수집하여 다른 모델 아키텍처에서 학습을 수행하는 경우 방어 실효성이 상실될 수 있음을 의미한다. 따라서 방어 기법이 특정 모델 구조에 의존적이라는 점은 일반화 측면에서 중대한 제약으로 작용한다.

추가적으로, 본 논문에서는 웹툰 이미지에 기존 기법을 적용하여 성능을 실험하였다. 기존 기법들은 주로 사실적인 사진을 대상으로 평가되어 왔으나, 웹툰 이미지는 평면적이고 단순한 특성을 지니기 때문에 perturbation의 영향이 더욱 뚜렷하게 나타날 것으로 예상하였다. 이에 따라 실험을 진행한 결과, 그림 8과 같이 perturbation 삽입 시 이미지 품질 저하가 명확하게 관찰되었다. 이는 복잡한 실제 사진과 달리, 단순한 웹툰 이미지 도메인에서는 기존 기법들이 이미지 품질 저하라는 한계를 가지는 것을 보여준다. 특히, 웹툰 작가들은 자신의 작품을 온라인에 고화질로 공개하는 것이 중요한데, 기존 기법들은 이미지 품질 저하를 초래하여 실질적인 활용에 한계가 있음을 시사한다.

V. 향후 연구 방향 및 결론

본 논문에서는 딥러닝 기반 이미지 생성 모델에서 발생할 수 있는 스타일 모방 문제를 인식하고, 현재까지 제안된 스타일 모방 방지 기법들에 대한 소개와 더불어 해당 기법들의 한계를 분석하였다. 대표적인 방어 방식으로 활용되고 있는 이미지 perturbation 삽입 기법은 생성 모델이 특정 스타일에 대한 학습을 방해하는 데 일정 수준의 효과를 보였으나, 시각적 품질 저하와 모델 구조 및 버전의 종속성에 대한 구조적 한계를 가지고 있었다. 이러한 한계를 극복하기 위해, 향후 연구에서는 다음과 같은 방향으로의 발전이 필요할 것으로 보인다.

첫째, 시각적 품질을 보존하는 방어 기법에 대한 연구가 필요하다. 인간이 인지할 수 없는 수준의 고차원적 특징 공간에서의 방어 perturbation 삽입 또는 주파수 도메인 기반의 비가시적 변형 방식 등을 고려할 수 있다.

둘째, 생성 모델의 구조에 대한 의존성이 낮은 일반화된 기법 연구가 필요하다. 생성 모델의 구조에 의존하지 않고, 다양한 모델 및 버전에서 방어 효과를 유지할 수 있는 일반화된 방어 메커니즘 개발이 요구된다.

셋째, adversarial robustness가 아닌, privacy-preserving generation 또는 워터마크 삽입 및 사후 추적 기법 개발로의 전환이 필요하다. 기존의 스타일 모방 방지 기법이 adversarial noise에 기반했다면, 향후에는 모델 내부에서 학습이 되지 않도록 하는 원천적 접근, 혹은 복원 가능한 워터마크 삽입을 통한 사후 추적 기반의 protection이 필요할 것이다.

종합하면 스타일 모방 방지 기법은 단순한 시각적 변형을 넘어서, 학습 과정 전반에 영향을 미칠 수 있는 구조적이고 일반화된 접근이 필요하다. 특히 다양한 생성 모델 환경에서도 일관된 방어 효과를 보장할 수 있는 일반화된 메커니즘의 방어 전략 개발이 주요한 과제로 남을 것이다. 더불어, 시각적 품질을 보존하면서도 효과적인 방어를 달성할 수 있는 새로운 방식에 대한 탐색이 요구된다.

참 고 문 헌 (References)

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695, 2022.
doi: https://doi.org/10.48550/arXiv.2112.10752

[2] A. Ramesh, M. Pavlov, G. Goh, et al., “Zero-shot text-to-image generation,” Proc. Int. Conf. Machine Learning (ICML), PMLR, pp. 8821–8831, 2021.
doi: https://doi.org/10.48550/arXiv.2102.12092

[3] OpenAI, “ChatGPT overview,” https://openai.com/chatgpt/overview/ (accessed May 19, 2025).

[4] A. Radford, J. W. Kim, C. Hallacy, et al., “Learning transferable visual models from natural language supervision,” Proc. Int. Conf. Machine Learning (ICML), PMLR, pp. 8748 – 8763, 2021.
doi: https://doi.org/10.48550/arXiv.2103.00020

[5] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Vol. 9351, Springer, pp. 234 – 241, 2015.
doi: https://doi.org/10.48550/arXiv.1505.04597

[6] N. Ruiz, Y. Li, C. Liao, et al., “Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation,” Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 28517 – 28527, 2023.
doi: https://doi.org/10.48550/arXiv.2208.12242

[7] R. Gal, O. Alaluf, Y. Nadav, et al., “An image is worth one word: Personalizing text-to-image generation using textual inversion,” arXiv preprint arXiv:2208.01618, 2022.
doi: https://doi.org/10.48550/arXiv.2208.01618

[8] N. Kumari, Y. Li, R. Salakhutdinov, et al., “Multi-concept customization of text-to-image diffusion,” Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 28528 – 28538, 2023.
doi: https://doi.org/10.48550/arXiv.2212.04488

[9] K. Sohn, S. Gidaris, R. Zhang, et al., “Styledrop: Text-to-image generation in any style,” arXiv preprint arXiv:2306.00983, 2023.
doi: https://doi.org/10.48550/arXiv.2306.00983

[10] M. Heikkilä, “This artist is dominating AI-generated art. And he’s not happy about it,” MIT Technology Review, 2022.

[11] S. Shan, E. Zou, J. Zhu, et al., “Glaze: Protecting artists from style mimicry by text-to-image models,” Proc. 32nd USENIX Security Symposium (USENIX Security 23), 2023.
doi: https://doi.org/10.48550/arXiv.2302.04222

[12] C. Liang and X. Wu, “Mist: Towards improved adversarial examples for diffusion models,” arXiv preprint arXiv:2305.12683, 2023.
doi: https://doi.org/10.48550/arXiv.2305.12683

[13] T. V. Le, X. Xu, X. Zhu, et al., “Anti-dreambooth: Protecting users from personalized text-to-image synthesis,” Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), pp. 16310 – 16319, 2023.
doi: https://doi.org/10.48550/arXiv.2303.15433

[14] E. J. Hu, Y. Shen, P. Wallis, et al., “LoRA: Low-rank adaptation of large language models,” Proc. Int. Conf. Learning Representations (ICLR), pp. 1 – 18, 2022.
doi: https://doi.org/10.48550/arXiv.2106.09685

[15] C. Saharia, W. Chan, S. Saxena, et al., “Photorealistic text-to-image diffusion models with deep language understanding,” Advances in Neural Information Processing Systems (NeurIPS), Vol. 35, pp. 36479 – 36494, 2022.
doi: https://doi.org/10.48550/arXiv.2205.11487

[16] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 586 – 595, 2018.
doi: https://doi.org/10.48550/arXiv.1801.03924

[17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” arXiv preprint arXiv:1706.06083, 2017.
doi: https://doi.org/10.48550/arXiv.1706.06083

[18] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “VGGFace2: A dataset for recognising faces across pose and age,” Proc. IEEE Int. Conf. Automatic Face & Gesture Recognition (FG), pp. 67 – 74, 2018.
doi: https://doi.org/10.48550/arXiv.1710.08092

[19] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” arXiv preprint arXiv:1710.10196, 2017.
doi: https://doi.org/10.48550/arXiv.1710.10196

[20] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, “RetinaFace: Single-shot multilevel face localisation in the wild,” Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 5203 – 5212, 2020.
doi: https://doi.org/10.48550/arXiv.1905.00641

[21] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 4690 – 4699, 2019.
doi: https://doi.org/10.1109/TPAMI.2021.3087709

[22] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” Proc. Int. Conf. Learning Representations (ICLR), 2021.
doi: https://doi.org/10.48550/arXiv.2010.02502

[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” Proc. Int. Conf. Learning Representations (ICLR), 2018.
doi: https://doi.org/10.48550/arXiv.1706.06083

[24] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al., “Cogview: “Mastering text-to-image generation via transformers,” Advances in neural information processing systems, 34:19822 – 19835, 2021.
doi: https://doi.org/10.48550/arXiv.2105.13290

[25] MingDing, WendiZheng, and Wenyi Hong,andJie Tang, “Cogview2: Faster and better text-to-image generation via hierarchical transformers,” Advances in Neural Information Processing Systems, 35:16890 – 16902, 2022.
doi: https://doi.org/10.48550/arXiv.2204.14217

[26] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al., “Photorealistic text-to-image diffusion models with deep language understanding,” Advances in neural information processing systems, 35:36479 – 36494, 2022.
doi: https://doi.org/10.48550/arXiv.2205.11487

[27] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima, “Uncurated image-text datasets: Shedding light on demographic bias,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6957 – 6966, 2023.

doi: https://doi.org/10.48550/arXiv.2304.02828
[28] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein, “Diffusion art or digital forgery? investigating data replication in diffusion models,” In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6048–6058, 2023.
doi: https://doi.org/10.48550/arXiv.2212.03860
[29] https://en.wikipedia.org/wiki/Midjourney
[30] https://www.deviantart.com/
[31] MacCarthy, M. (2024, January 24). Generative AI and copyright issues globally: ANI Media v. OpenAI. Tech Policy Press, https://www.techpolicy.press/generative-ai-and-copyright-issues-globally-ani-media-v-openai/
[32] Silicon Republic. (2023). New Glaze app looks to protect art from AI mimicry. Silicon Republic, https://www.siliconrepublic.com/machines/glaze-ai-art-protection-copyright-mimicry
[33] Wired Staff. (2024). AI copyright case tracker. Wired, https://www.wired.com/story/ai-copyright-case-tracker/
[34] BrandShield. (2024). Generative AI and copyright infringement: What brands need to know. BrandShield, https://www.brandshield.com/blog/generative-ai-and-copyright-infringement-what-brands-need-to-know/
[35] Ars Technica Staff. (2024). Glaze, a tool protecting artists from AI, bypassed by attack as demand spikes. Ars Technica, https://arstechnica.com/tech-policy/2024/07/glaze-a-tool-protecting-artists-from-ai-bypassed-by-attack-as-demand-spikes/
[36] Freethink Staff. (2024). AI copyright violations. Freethink, https://www.freethink.com/robots-ai/ai-copyright-violations
[37] Zhu, Peifei, Tsubasa Takahashi, and Hirokatsu Kataoka. “Watermark-embedded adversarial examples for copyright protection against diffusion models.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
doi: https://doi.org/10.48550/arXiv.2404.09401
[38] Gui, Dongnan, et al. “I2VGuard: Safeguarding Images against Misuse in Diffusion-based Image-to-Video Models.” Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.
[39] Chen, Liangyu, et al. “Simple baselines for image restoration.” European conference on computer vision. Cham: Springer Nature Switzerland, 2022.
doi: https://doi.org/10.48550/arXiv.2204.04676

저 자 소 개

한 주 아

- 현재 : 동국대학교 공과대학 멀티미디어공학과 학사과정
- ORCID : https://orcid.org/0009-0008-6957-6944
- 주관심분야 : Computer vision, preventing personalized concept from diffusion model

박 주 원

- 동국대학교 AI융합대학 AI소프트웨어융합학부 멀티미디어소프트웨어공학전공 학사
- 현재 : 동국대학교 일반대학원 컴퓨터AI학과 멀티미디어소프트웨어전공 석사과정
- ORCID : https://orcid.org/0009-0008-6606-5663
- 주관심분야 : Computer vision, 3D human mesh recovery

저 자 소 개

차 민 희

- 동국대학교 공과대학 멀티미디어공학과 학사
- 현재 : 동국대학교 일반대학원 컴퓨터AI학과 멀티미디어소프트웨어전공 석사과정
- ORCID : https://orcid.org/0000-0003-3496-2227
- 주관심분야 : Computer vision, preventing personalized concept from diffusion model

조 성 인

- 서강대학교 공과대학 전자공학과 학사
- 포항공과대학교 전기전자공학과 박사
- 현재 : 동국대학교 첨단융합대학 컴퓨터AI학부 부교수
- ORCID : https://orcid.org/0000-0003-4251-7131
- 주관심분야 : Computer vision, image analysis and enhancement, video processing, deep learning