논문 2025-4-6 http://dx.doi.org/10.29056/jsav.2025.12.06

신경망 기반 스타일 변환을 방해할 수 있는 영역별 노이즈 생성 기법

박성환*, 김종성**, 오병훈*, 황요한*, 홍준호***† , 이재우****†

Region-Specific Noise Generation for Untransferable Examples Against Neural Style Transfer

Sunghwan Park*, Jongseong Kim**, Byunghoon Oh*,
Yohan Hwang*, Junho Hong***† , Jaewoo Lee****†

요 약

Neural Style Transfer의 확산은 동의나 보상 없이 독창적인 예술 스타일을 추출하고 복제하는 것에 활용될 수 있기 때문에 디지털 창작자의 지식재산권에 중대한 위협이 된다. 기존의 보호 기법들은 효과적인 보호를 위해 이미지 품질을 크게 저하시키거나, 품질을 보존하는 대신 미미한 보호 성능을 보이는 근본적인 상충관계(trade-off)에 직면해 있다. 본 연구는 이미지를 패치 영역(2%)과 배경 영역(98%)으로 공간적으로 분리하고 각각 적합한 교란 전략을 적용하여 이러한 문제를 해결한다. 패치 영역에는 강력한 교란으로 어텐션 메커니즘을 탈취하고, 배경 영역에는 최소한의 교란으로 전역 통계를 오염시켜 경사도 간섭을 원천적으로 제거한다. 실험 결과 제안 기법은 스타일 변환을 90% 이상 방해(STDR 0.8995)하면서도 높은 시각적 품질(SSIM 0.8624)을 유지하여, 기존 대비 38% 향상된 보호 효율을 달성했다.

Abstract

Neural Style Transfer poses a significant threat to the intellectual property rights of digital artists, as it can extract and replicate unique artistic styles without consent or compensation. Existing protection techniques face a fundamental trade-off: they either degrade image quality for effective protection or offer minimal protection to preserve quality. This paper proposes a novel region-specific noise generation that solves this trade-off by spatially segregating protection objectives. Our method divides an image into two distinct regions and applies different perturbation strategies to each: a small patch region (2% of image area) receives strong perturbations to hijack the attention mechanism, while the background region (98%) is subjected to minimal perturbations to corrupt global feature statistics. The key innovation lies in the spatial separation that fundamentally eliminates gradient interference between different loss objectives. Experimental results demonstrate that our method achieves over 90% reduction in style transfer effectiveness (STDR 0.8995) while maintaining high visual quality (SSIM 0.8624), representing a 38% improvement in protection efficiency compared to existing methods.

한글키워드 : 신경망 스타일 변환, 전송불가능한 예제, 적대적 공격, 영역별 최적화, 저작권 보호

keywords : Neural Style Transfer, Untransferable Examples, Adversarial Attack, Region-specific optimization, Copyright Content Protection

* 중앙대학교 융합보안학과
** 중앙대학교 융합보안학과(공동주저자)
*** 성신여자대학교 융합보안공학과(공동교신저자)
**** 중앙대학교 산업보안학과
† 교신저자: 이재우(email: jaewoolee@cau.ac.kr),
홍준호(email: hjh@sungshin.ac.kr)
접수일자: 2025.10.20. 심사완료: 2025.10.30.
게재확정: 2025.12.20.

1. 서 론

신경망 스타일 변환(Neural Style Transfer, 이하 NST) 기술의 급격한 발전은 예술적 이미지 생성을 대중화하여 누구나 자신의 사진을 유명한 예술가의 스타일을 모방한 예술 작품으로 변환할 수 있게 하였다 [1-6]. NST는 새로운 창작의 가능성을 열었으나, 동시에 지식재산권 및 아티스트의 고유한 시각적 스타일 무단 도용에 대한 심각한 우려를 제기하였다. 특히 현대의 스타일 변환 모델 StyTr^2 [5], S2WAT [6]과 같은 비전 트랜스포머(Vision Transformer, 이하 ViT) 기반 모델들은 어떠한 입력 이미지든 그 독특한 예술적 특징을 추출하고 재현할 수 있다. 따라서 스타일 변환을 위한 이미지의 무단 사용은 콘텐츠 창작자의 고유한 시각적 스타일이 전문적 정체성과 경제적 가치에 중대한 위협을 제기한다 [7, 8]. 스타일 변환이 정확한 픽셀을 복사하는 대신 통계적 패턴을 학습하여 동작하므로, 전통적인 저작권 침해 이슈를 우회한다 [9]. 이러한 규제의 공백은 이미지의 합법적 사용을 위한 시각적 품질은 보존하면서 무단 스타일 추출로부터 시각적 콘텐츠를 선제적으로 보호할 수 있는 기술적 해결책을 필요로 한다.

무단 스타일 변환을 방지하기 위한 기존 접근법 [10-17]들은 보호 효과와 이미지 품질 보존 사이에 근본적인 상충 관계(trade-off)에 직면한다. 이미지 전체에 균일한 교란(perturbation)을 적용하는 방법들은 스타일 변환을 성공적으로 방해할 수 있으나, 의미 있는 보호 성능을 얻기 위해 이미지의 미적, 상업적 가치를 떨어뜨리는 수준의 눈에 띄는 왜곡(artifact)을 감수해야한다. 그림 1은 이러한 한계를 명확히 보여준다.

이미지의 시각적 품질을 정량적으로 측정하는 SSIM 지표를 기준으로 볼 때, 기존 기법인 NSP [16]는 원본 이미지에 비해 품질이 약 38% 저하

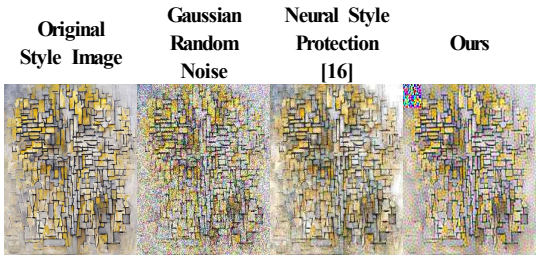


그림 1. 제안한 방법과 기존 보호 기법 간의 시각적 품질 비교 (SSIM 지표로 측정됨)
Fig. 1. Visual quality comparison between the proposed method and existing protection techniques, measured by the SSIM metric.

되는 것을 확인할 수 있다. 이러한 심각한 품질 저하는 저작권 보호의 목적 자체를 무의미하게 만든다. 아티스트는 합법적인 전시 및 판매를 위해 자신의 작품이 시각적으로 매력적인 상태를 유지해야 하기 때문이다. 결국 기존 기법들은 약한 보호와 심각한 품질 저하 사이에서 비효율적인 선택을 강요한다. 이러한 문제를 해결하기 위해, 본 논문은 무단 스타일 변환으로부터 이미지를 보호하면서도 보호 성능과 시각적 품질 간의 효율적인 균형을 달성하는 새로운 영역별 노이즈 생성 기법을 제안한다. 본 연구의 핵심 아이디어는 “이미지의 시각적 품질을 실용적인 수준으로 유지하면서, 모델이 일관된 스타일 표현을 추출하는 능력을 효과적으로 손상시키는 공간적으로 표적화된 교란”을 통해 스타일 변환 모델을 공략하는 것이다. 이미지 전체에 균일한 노이즈를 적용하는 대신, 이미지를 전략적으로 두 개의 영역으로 분할한다. 하나는 강력한 교란으로 모델의 어텐션 메커니즘을 탈취(hijack)하는 작은 패치(patch) 영역이며, 다른 하나는 특징 통계(feature statistics)를 미묘하게 오염시키는 배경 영역이다. 본 방법론은 현대 스타일 변환 아키텍처의 다음의 두 가지 근본적인 취약점을 이용한다. **1) 어**

텐션 취약점(Attention Vulnerability): ViT는 셀프-어텐션(self-attention) 메커니즘을 통해 높은 현저성(salience)을 갖는 영역을 우선시한다. 작지만 매우 활성화된 패치를 생성함으로써, 모델의 어텐션을 독점하여 실제 예술 스타일을 적절히 분석하지 못하도록 방해할 수 있다 [18]. **2) 특징 통계 손상(Feature Statistics Corruption):** 스타일 변환은 다양한 스케일(scale)에 걸쳐 특징 통계(평균 및 표준편차)를 일치시키는 것에 의존한다. 배경 특징이 적대적 패치와 상관관계를 갖도록 강제함으로써, 지역적 시각적 일관성은 유지하면서 이러한 전역 통계를 손상시킬 수 있다 [19].

본 논문의 주요 기여는 다음과 같다.

1. **저작권 보호 프레임워크:** 디지털 콘텐츠 보호의 중요한 공백을 해결하며, 무단 스타일 변환으로부터 예술 이미지의 창작성 등을 보호하기 위해 설계된 영역별 노이즈 생성 기법을 제시한다.
2. **비간섭 최적화 전략(Non-Interfering Optimization Strategy):** 서로 다른 손실 목표를 별도의 영역에 할당하여 경사도 간섭을 제거하고, 최소한의 가시적 교란으로 더 효과적인 보호를 달성하는 공간적 경사도 분리 기법을 도입한다.
3. **효율적인 보호 균형(Efficient Protection Balance):** 스타일 변환 품질을 80% 이상 저하시키면서도, 보호된 이미지가 상업적 사용 및 예술적 표시에 적합한 수준의 시각적 품질을 유지하도록 보호 성능과 품질 간의 효율적인 상충 관계를 달성한다.

2. 선행 연구

본 장에서는 제안하는 보호 기법의 기반이 되는 선행 연구들을 고찰한다. NST 기술과 취약점, 기존 보호 기법의 한계, 그리고 본 연구의 이론적 토대가 되는 경사도 간섭 문제를 중심으로 논의한다.

2.1 신경 스타일 변환과 취약점

NST는 Gatys et al. [1]이 컨볼루션 신경망(CNN: Convolutional Neural Network, 이하 CNN)이 콘텐츠와 스타일 표현을 분리하고 재결합할 수 있음을 처음으로 입증한 이래로 상당한 발전을 이루었다. 그들의 선구적인 연구는 스타일 정보가 CNN의 여러 계층에 걸친 특징 상관관계(Gram Matrix)에 인코딩되는 반면, 콘텐츠는 특징 맵 자체에 보존된다는 것을 밝혔다. 이 발견은 빠른 피드-포워드(feed-forward) 방식 [2], 임의 스타일 변환 [3], 그리고 어텐션 기반 접근법 [4] 등 수많은 개선을 촉발시켰다.

최근 ViT를 활용한 발전은 스타일 변환 품질을 한층 더 향상시켰다. StyTr^2 [5]은 트랜스포머 아키텍처가 스타일 패턴의 장거리 의존성(long-range dependencies)을 더 잘 모델링할 수 있음을 보여주었다. 유사하게 S2WAT 모델 [6]은 계층적 스트립 윈도우 어텐션(hierarchical strip window attention)을 사용하여 다중 스케일(multi-scale)의 스타일 패턴을 포착함으로써 스타일 변환을 충실히 수행할 수 있음을 보였다. 이러한 모델들의 정교한 어텐션 메커니즘은 품질을 향상시키는 동시에 새로운 취약점을 야기하는데, 특히 현저한(salient) 적대적 패턴에 의해 어텐션이 독점될 수 있는 민감성이 두드러진다.

이러한 NST 모델로부터 이미지를 보호하려는 초기 연구들 [11, 14]부터 이미지의 고주파수 성분(high-frequency domain)을 조작하는 기법 [12, 15], 의미론적인 특징을 변경하는 기법 [13, 17] 등이 제시되었으나 이들은 주로 전역 교란(Global Perturbation) 방식에 의존해왔다. 이 접근법

들은 이미지 전체에 균일한 노이즈를 추가하여 스타일 특징 추출을 방해하지만, 효과적인 보호를 위해서는 필연적으로 심각한 시각적 품질 저하를 감수해야 한다는 명백한 한계를 가진다. 이는 보호의 실용성을 저해하는 근본적인 상충 관계(trade-off)이다.

기존 보호 기법들이 상충 관계 문제에 직면하는 이론적 원인은 경사도 간섭(gradient interference)에서 찾을 수 있다[20]. 다수의 손실 함수(예: 보호 성능 손실, 품질 보존 손실)가 동일한 이미지 픽셀에 대해 동시에 최적화를 시도할 때, 각 손실에서 계산된 경사도가 서로 충돌하여 노이즈 생성을 방해하고 차선의 결과(suboptimal solution)를 야기한다.

본 연구는 이 문제를 해결하기 위해 목표를 공간적으로 분리(spatially segregating objectives)하는 새로운 접근법을 취한다. 즉, 서로 다른 손실 함수가 이미지의 서로 다른 영역에만 배타적으로 작용하도록 설계하여 경사도 간섭을 원천적으로 제거한다. 이러한 공간적 분리 전략은 본 논문에서 제안하는 기법의 핵심적인 이론적 토대이며, 기존의 스타일 변환 보호 연구에서는 시도된 바 없다.

2.2 인공지능 시대의 저작권 보호

인공지능과 저작권법의 교차점은 기술적 해결책이 반드시 다루어야 할 전례 없는 과제를 제시한다 [21-22]. 정확한 복제를 위해 설계된 전통적인 저작권 체계는 픽셀을 복사하기보다 패턴을 학습하고 재현하는 인공지능 시스템에 대처하는 데 어려움을 겪는다. 주요국 등에서는 아직 인공지능을 통한 스타일 도용에 대한 명확한 선례를 확립하지 못했으며, 이는 아티스트와 콘텐츠 창작자에게 불확실성을 야기한다.

기술적 보호 메커니즘은 법적 체계를 보완하는 중요한 요소로 부상했다. Glaze [23]는 확산 모델(diffusion model)을 위한 스타일 은닉(style cloaking) 기술의 선구자로서, 모델이 예술 스타일을 잘못 학습하도록 유도하는 교란을 추가했다. Mist [24]는 이 개념을 확장하여 여러 생성 모델로부터 동시에 보호하는 기술을 선보였다. 그러나 이러한 기법들은 주로 스타일 변환보다는 텍스트-이미지(text-to-image) 모델에 초점을 맞추고 교란이 종종 모델에 특화되어 있어 새로운 모델이 등장할 때마다 업데이트가 필요하다.

데이터 포이즈닝(data poisoning) 접근법 [25]은 학습 데이터를 보호하는 데 유망함을 보였으나, 종종 퓨샷(few-shot) 또는 단일 이미지 설정에서 작동하는 스타일 변환에는 적용하기 어렵다. 워터마킹(watermarking) 기술 [26]은 스타일 변환 결과물을 추적할 수는 있지만, 최초의 도용 자체를 막지는 못한다. 본 방법론은 스타일 변환 시나리오를 위해 특별히 설계된 선제적 보호를 제공함으로써 중요한 공백을 메운다.

3. 제안한 기법

본 장에서는 무단 스타일 변환으로부터 이미지를 보호하기 위한 우리의 영역별 노이즈 생성 기법을 제시한다. 먼저 공식적인 문제 정의를 시작으로 아키텍처 개요, 손실 함수에 대한 상세 설명, 그리고 영역별 최적화 절차를 순서대로 기술한다.

3.1 문제 정의

무단 스타일 변환으로부터 보호가 필요한 예술 스타일 이미지를 $I_s \in \mathbb{R}^{H \times W \times 3}$라 하자. 우리의 목표는 그림 2에서 나타나듯이 스타일 변환을 방해하는 이미지 $I_u = I_s + \delta$를 생성하는 것이며, 여기서 δ는 시각적 품질을 유지하면서 스타일

추출을 방지하는 적대적 노이즈를 나타낸다. 콘텐츠 이미지 I_c와 스타일 이미지 I_s를 결합하여 결과물 I_{cs}를 생성하는 스타일 변환 모델 가 주어졌을 때, 보호 목표는 Eq. (1)과 같이 정의된다.

$$\min_{\delta} Q(T(I_c, I_s + \delta))$$
$$\text{s.t. } \|\delta\|_\infty \leq \epsilon \text{ and } P(I_s, I_s + \delta) \geq \tau \quad (1)$$

여기서 $Q(\cdot)$는 스타일 변환 품질을 측정하며 (낮을수록 좋음), $P(\cdot)$는 인지적 유사성을 측정하고 (높을수록 좋음), ϵ은 교란 예산이며, τ는 최소한의 허용 가능한 시각적 품질 임계값이다.

본 접근법의 핵심은 Eq. (2)에서 볼 수 있듯이 노이즈를 패치 영역과 배경 영역으로 공간적으로 분해하는 데 있다.

$$\delta = \delta_{patch} \oplus \delta_{bg} \quad (2)$$

여기서 δ_{patch}는 작은 패치 영역 R_{patch}(이미지 면적의 2%)에서의 교란을, δ_{bg}는 배경 영역 R_{bg}(이미지 면적의 98%)에서의 교란을 나타내며, $\oplus$는 공간적 결합(spatial concatenation)을 의미한다.

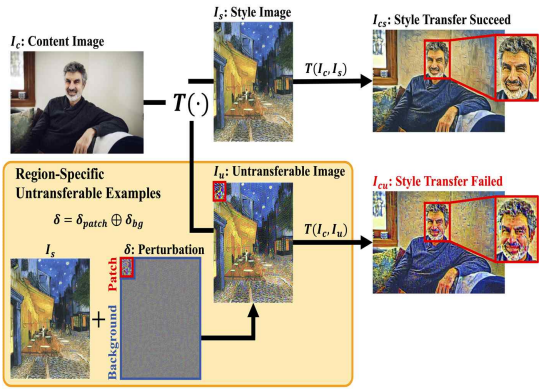


그림 2. 영역별 노이즈 생성 기법 개요
Fig. 2. Region-Specific Untransferable Examples Overview.

3.2 영역별 노이즈 생성 기법 개요

우리의 프레임워크는 세 가지 주요 구성 요소로 이루어진 S2WAT [6] 스타일 변환 파이프라인 상에서 작동한다.

1. **계층적 인코더(Hierarchical Encoder)** E: 스트립 윈도우 어텐션(Strip Window Attention)을 갖춘 ViT로, 다중 스케일 특징을 추출한다.
2. **변환 모듈(Transfer Module)** M: 콘텐츠와 스타일 특징을 결합하는 트랜스포머 디코더이다.
3. **디코더(Decoder)** D: 스타일화된 이미지를 재구성하는 미러링된 VGG 네트워크이다.

이때 스타일이 변환된 이미지 I_{cs}를 얻는 과정은 $I_{cs} = D(M(E(I_c), E(I_s)))$으로 표현될 수 있다. 우리의 보호 메커니즘은 이 파이프라인의 두 가지 취약점, 즉 E의 어텐션 메커니즘과 스타일 표현에 사용되는 특징 통계를 표적으로 한다.

3.3 특정 영역에 특화된 손실 함수 설계

우리는 상호 보완적인 두 개의 손실 함수를 사용하며, 각각은 특정 공간 영역에 할당된다. 그림 3은 본 논문에서 사용하는 두 손실 함수에 의해 생성되는 노이즈가 이미지의 어떤 부분을 변화시키는지 시각적으로 보여준다.

3.3.1 어텐션 하이재킹을 위한 패치 영역 대상 활성 손실 설계

패치 활성화 손실은 트랜스포머의 어텐션을 특정 영역에 과도하게 집중시켜 전반적인 스타일 분석을 방해할 목적으로, 패치 영역에만 독점적으로 작용한다. 이를 위해 Eq. (3)와 같이 패치 영역의 특징 활성화는 극대화하는 동시에 배경 영역의 활성화는 억제한다.

$$L_{act} = -\frac{1}{|F_{patch}|}\sum_{f \in F_{patch}} \|f\|_2 + \lambda_{reg}\frac{1}{|F_{bg}|}\sum_{f \in F_{bg}} \|f\|_2 \quad (3)$$

여기서 F_{patch}와 F_{bg}는 각각 패치 및 배경 영역에서 추출된 인코더 특징들이며, $\lambda_{reg} = 0.5$는 정규화 가중치이다.

첫 번째 항은 패치 영역의 활성화 크기를 최대화하여 극도로 현저한(salient) 영역을 생성한다. 두 번째 항은 배경 활성화를 억제하는 정규화(regularization) 역할을 수행하며, 이는 어텐션 탈취(attention hijacking)의 효율성을 높이는 데 필수적이다. 만약 이 정규화 항이 없다면, 최적화 과정에서 패치뿐만 아니라 배경의 활성화까지 전반적으로 증가하여 패치와 배경 간의 활성화 대비(activation contrast)가 줄어들게 된다. 따라서 이 항은 모델의 어텐션을 소수의 패치에 효과적으로 고정시키는 '대조적 최적화(contrastive optimization)'를 보장한다.

3.3.2 통계 교란을 위한 배경 영역 대상 특징 상관 손실 함수

상관관계 손실 L_{corr}은 Eq. (4)의 식에 따라 배경 영역의 특징 통계를 오염시켜 일관된 스타일 정보 추출을 방해할 목적으로 배경 영역 R_{bg}에만 독점적으로 작용한다. 이 손실 함수의 핵심 메커니즘은 정상적인 배경 영역의 특징 통계를 L_{act}에 의해 인위적으로 왜곡된 패치 영역의 특징 통계에 강제로 동기화(synchronization)시키는 것이다.

$$L_{corr} = \sum_{f}^{F_{VGG}} [\|\mu_{bg}^{f} - \mu_{patch}^{f}\|_2^2 + \|\sigma_{bg}^{f} - \sigma_{patch}^{f}\|_2^2] \quad (4)$$

여기서 μ_{bg}^{f}, σ_{bg}^{f}는 배경 영역에 대한 VGG 특징 레이어 $f \in F_{VGG}$의 평균 및 표준편차이며, μ_{patch}^{f}, σ_{patch}^{f}는 패치 영역의 해당 통계이다.

결과적으로 이미지 전반의 특징 통계가 소수의 왜곡된 패치 통계 프로필로 오염(contaminate)된다. 이는 NST 모델이 이미지 전체에서 일관되고 의미 있는 스타일 특징(Gram Matrix)을 계산하는 과정을 근본적으로 교란하여 스타일 전송에 실패하도록 유도한다.

3.4 공간적 분리를 통한 기울기 최적화 과정의 간섭 최소화

본 방법론의 핵심은 최적화 과정에서의 최적화하는 경사도를 공간적으로 분리하는 것에 있다. 최적화 과정을 이미지 전역에서 수행하는 전략은 Eq. (5)과 같이 진행한다.

$$\delta^{t+1} = \delta^{t} - \alpha \nabla_{\delta}(\omega_{act} L_{act} + \omega_{corr} L_{corr}) \quad (5)$$

이는 $\nabla_{\delta} L_{act}$와 $\nabla_{\delta} L_{corr}$가 동일 픽셀에 대해 상충하는 방향을 가질 때 경사도 간섭을 유발한다. 반면에 우리의 영역별 접근법은 Eq. (6)과 같이 패치 영역과 배경 영역으로 공간적으로 분리하여 최적화함으로써 경사도 간섭의 영향을 줄인다.

$$\nabla_{\delta} = M_{patch} \odot \nabla_{\delta} L_{act} + M_{bg} \odot \nabla_{\delta} L_{corr} \quad (6)$$

여기서 M_{patch}와 M_{bg}는 각각 패치 및 배경 영역에 대한 이진 마스크(binary masks)이며, $\odot$는 요소별 곱셈(element-wise multiplication)을 나타낸다.

4. 실험 결과

본 장에서는 우리의 영역별 노이즈 생성 기법의 효율성을 평가하는 포괄적인 실험을 제시한다. 최신 스타일 변환 모델에 대한 보호 성능과 시각적 품질 보존을 평가하며 제안하는 접근법을 기존 보호 기법들과 비교한다.

4.1 실험 설정

4.1.1 실험 환경

실험환경: 모든 실험은 NVIDIA RTX 4090 GPU 1대, AMD Ryzen 3950x CPU를 사용하였으며, PyTorch 2.5.1 프레임워크에서 구현되었다. 영역별 노이즈 생성을 위한 최적화는 Adam optimizer (lr=0.01, batch size=8)를 사용하여 200 iteration 동안 수행되었다.

4.1.2 대상 모델

본 연구의 보호 성능은 최신 ViT 기반의 NST 모델인 S2WAT [6]을 대상으로 평가하였다. S2WAT은 정교한 어텐션 메커니즘을 통해 높은 품질의 스타일 변환을 수행하므로, 이를 효과적으로 방해할 수 있다면 다른 모델에도 일반화된 보호 성능을 기대할 수 있다.

4.1.3 평가 지표

우리는 보호 효과와 시각적 품질을 포괄적으로 평가하기 위해 아래 지표들을 사용한다.

1. **스타일 변환 저하율($STDR$):** 스타일 변환 품질의 저하율을 측정하는 핵심 지표. 보호 적용 후 스타일 손실 $L_{style} = L_{\mu} + L_{\sigma}$의 변화를 기반으로 $(Q_{I_{cs}} - Q_{I_{cu}})/Q_{Ics}$로 계산되며, 값이 높을수록 보호 성능이 우수함을 의미한다. 여기서 $Q(\cdot)$는 Eq. (7)로 계산되는 스타일 변환 품질 점수이다.

$$Q = -\frac{1}{F}\sum_{f}^{F}(L_{\mu}^{f} + L_{\sigma}^{f}) \tag{7}$$

여기서 스타일 손실 구성 요소 중 L_{μ}은 $MSE(\mu(I_{cs}),\mu(I_s))$을 통해 계산하며, L_{σ}는 $MSE(\sigma(Ics),\sigma(Is))$을 통해 계산한다.

2. **특징 상관관계 거리(FCD):** 원본 스타일 변환 결과 I_{cs}와 보호된 스타일 변환 결과 I_{cu} 사이의 Gram Matrix 차이를 L2 거리로 측정한다. 스타일 특징 통계가 얼마나 효과적으로 손상되었는지를 정량화하며, 높은 값은 성공적인 특징 교란을 나타낸다.

$$FCD = \frac{1}{F_{VGG}}\sum_{f}^{F_{VGG}}|G_f^{I_{cs}} - G_f^{I_{cu}}|_2 \tag{8}$$

여기서 $G^{I_{cs}}$와 $G^{I_{cu}}$는 각각 I_{cs}와 I_{cu}에 대한 그람 행렬(Gram Matrices)을 나타낸다.

3. **어텐션 엔트로피 감소량(AED):** 트랜스포머의 어텐션 맵(attention map)에 대한 엔트로피를 계산하여 어텐션 분포의 무질서도를 측정한다. 패치 영역으로 어텐션을 집중시켜 엔트로피를 낮추는 것이 목표이므로, 높은 AED 값은 성공적인 어텐션 탈취(hijacking)를 의미한다. 이 때 AED는 Eq. (9)로 계산한다.

$$AED = H(I_{cs}) - H(I_{cu}) \tag{9}$$

여기서 $H(\cdot)$는 공간적 어텐션 확률 분포의 새넌 엔트로피(Shannon entropy)를 나타낸다.

4. **구조적 유사성 지수($SSIM$):** 원본 스타일 이미지 I_s와 보호된 이미지 I_u간의 구조적 유사성을 측정하는 지표. 1에 가까울수록 원본과의 시각적 차이가 적음을 의미하여 품질 보존성이 높다고 평가된다. 이 때 SSIM은 Eq. (10)으로 계산한다.

$$SSIM(I_s, I_u) = \frac{(2\mu_{I_s}\mu_{I_u} + C_1)(2\sigma_{I_sI_u} + C_2)}{(\mu_{I_s}^2 + \mu_{I_u}^2 + C_1)(\sigma_{I_s}^2 + \sigma_{I_u}^2 + C_2)} \tag{10}$$

여기서 C_1과 C_2는 수치적 안정성을 위한 작은 상수들을 나타낸다.

4.1.4 비교 기법

우리는 최신 보호기법 Neural Style Protection 및 Random Noise 기반 perturbation 생성 기법을 대상으로 제안하는 방법론의 성능을 검증한다. 아래는 각 기법에 대한 설명이다.

- **Uniform Random Noise:** 이미지 전체에 균일 분포를 따르는 무작위 노이즈를 추가하는 가장 기본적인 교란 방식.
- **Gaussian Random Noise:** 정규 분포를 따르는 무작위 노이즈를 추가하는 방식.
- **Neural Style Protection(NSP) [16]:** 스타일 변환을 방해할 목적으로 설계된 최신(State-Of-The-Art) 적대적 공격 기법. 이미지 전체 픽셀에 대해 스타일 손실을 최대화하는 방향으로 교란을 생성한다.

4.2. 주요 결과

4.2.1 보호 효과

Table 1은 제안하는 영역별 노이즈 생성 기법이 모든 비교 대상을 압도하는 보호 성능을 달성했음을 보여준다. 제안 기법의 $STDR$은 0.8995를 기록하여 스타일 변환 품질을 약 90% 저하시켰다. 이미지의 손상을 감수하고 높은 ϵ을 사용했음에도 스타일 변환 품질을 최대 약 60% (Gaussian Random Noise) 저하시키는 것만 가능했다. 이는 NSP 및 Random Noise 방식들과 비교할 때, 제안된 보호 메커니즘이 실질적인 스타일 변환에 방어 효과가 있음을 입증한다.

이러한 강력한 보호 성능의 원인은 FCD와 AED 지표를 통해 분석할 수 있다. 먼저 FCD 값 2.8466은 제안 기법이 스타일 변환의 근간이 되는 특징 통계를 매우 효과적으로 손상시켰음을 정량적으로 나타낸다. 이는 패치 영역에 적용된 L_{act}와 배경 영역에 적용된 L_{corr}이 성공적으로 스타일 변환을 방해했음을 의미한다. 다만 Gaussian Random Noise에서 가장 높은 손상 정도($FCD = 4.4563$)를 보였음에도 그 결과가 스타일 보호 성능으로 직결되지 않았다. 즉, 특징 통계를 손상시키는 것만으로는 스타일 보호를 성공적으로 수행하지 못함을 시사한다. 다음으로 AED 값은 L_{act}이 S2WAT 모델의 어텐션 메커니즘을 의도대로 특정 패치 영역에 과도하게 집중시켜, 모델이 이미지 전반의 스타일 정보를 정상적으로 분석하지 못하도록 만드는 '어텐션 탈취' 전략이 성공했음을 보여준다. 결론적으로 제안하는 영역별 최적화 전략은 어텐션 탈취와 특징 통계 오염이라는 두 가지 목표를 독립적이고 효율적으로 달성함으로써, 기존의 교란 방식으로는 도달할 수 없었던 수준의 강력한 보호 성능을 제공한다.

표 1. 다양한 스타일 변환 모델의 보호 효과 비교
Table 1. Protection effectiveness against different style transfer models. Higher $STDR$, FCD, and AED values indicate better protection. Best results in **bold.**

Method	ϵ	STDR ↑	FCD ↑	AED ↑
Uniform Random Noise	16/255	0.0085	0.1614	-0.1654
	64/255	0.1429	1.2142	-0.3954
	128/255	0.2974	2.7342	-0.4131
Gaussian Random Noise	16/255	0.0438	0.3546	-0.1486
	64/255	0.2539	2.2979	-0.1315
	128/255	0.5964	**4.4563**	-0.4691
NSP	16/255	0.0262	0.0635	-0.0493
	64/255	0.1900	0.5749	0.1852
	128/255	0.1293	0.5080	-0.2071
Ours	16/255	**0.8995**	2.8466	**1.0579**

더불어 스타일 변환 방어를 목적으로 설계된 NSP가 Random Noise 생성 기법보다 낮은 성능을 보인 것은 다음과 같은 이유로 추측된다. NSP는 경사도 기반 최적화로 특징 통계의 차이(feature distance)를 최대화하려 하지만, 트랜스포머의 LayerNorm이 경사도를 왜곡하여 비효율적인 방향으로 교란을 유도한다. 반면 Random Noise는 모든 방향에 균등하게 노이즈를 추가하므로, 일부가 우연히 LayerNorm을 우회하는 효과적인 방향으로 작용하여 평균적으로 더 큰 영향을 미칠 수 있다. 조금 더 부연하자면, 트랜스포머의 LayerNorm 연산은 입력 특징 x의 평균 μ와 분산 σ를 정규화하는 과정에서 경사도 값의 스케일(scale)을 재조정한다. LayerNorm 연산이 $L_N(x) = \gamma(x-\mu)/\sigma + \beta$와 같이 표현될 때, 역전파 시 경사도 ∇_x는 σ에 반비례하는 항을 포함하게 된다. 이로 인해 NSP가 의도한 특정 방향의 경사도(예: 스타일 손실 최대화)가 LayerNorm을 통과하며 그 크기나 방향성이 왜곡되어, 최적화가 비효율적인 방향으로 유도될 수 있다.

4.2.2 시각적 품질 유지

Table 2는 제안하는 기법의 시각적 품질 보존 성능과 보호 성능 간의 상충 관계(trade-off)를 보여준다. 본 연구에서 '합법적 사용을 위한 실용적 수준의 시각적 품질'은 SSIM 지표 0.85 이상을 기준으로 삼았다. 이는 원본과 미세한 차이는 존재할 수 있으나, 이미지의 주요 구조와 미적 가치가 보존되어 상업적 전시나 합법적 감상에 무리가 없는 수준을 의미한다.

표 2. 제안하는 기법의 시각적 품질 손상 및 시각적 품질 손상 대비 보호효과의 정도 측정
Table 2. Visual quality metrics comparing protected images to original images. Best results in **bold.**

Method	ϵ	SSIM ↑	STDR/ (1-SSIM) ↑
Uniform Random Noise	16/255	0.9422	0.1471
	128/255	0.3419	0.4519
Gaussian Random Noise	16/255	0.8652	0.3249
	128/255	0.1797	0.7271
NSP	16/255	**0.9658**	0.7661
	128/255	0.4044	0.2171
Ours	16/255	0.8624	**6.5371**

우리의 방법은 0.8624의 $SSIM$을 기록했으며, 이는 NSP(0.9658)나 Uniform random (0.9422) 방식보다 다소 낮은 수치이다. 이는 패치와 배경 영역에 걸친 미세한 교란이 이미지의 구조적 정보에 일부 영향을 미쳤음을 의미한다.

그러나 $STDR/(1-SSIM)$ 지표는 단위 시각적 품질 손실 당 얻을 수 있는 보호 성능의 효율을 나타낸다. 우리의 기법은 이 지표에서 6.5371이라는 압도적으로 높은 값을 기록했으며, 이는 다른 기법들보다 훨씬 효율적으로 스타일 전송을 방해함을 의미한다. Random Noise 전략에서 시각적 품질 손상을 감수하고 ϵ을 크게 증가시킨 경우 해당 지표가 $\times 2.2-3.1$ 상승하는 모습을 보였으나, 제안하는 기법 대비 약 11% 수준에 머물렀다. 즉, 제안 기법은 약간의 시각적 품질을 희생하는 대신, 강력한 보호 효과를 제공하여 가장 실용적인 균형점을 달성했음을 시사한다.

4.2.3 손실 함수별 기여도 분석

Table 3은 제안하는 두 가지 손실 함수의 개별적 및 통합적 기여도를 분석한 결과이다. 분석 결과 L_{act}만 사용했을 경우에도 AED가 1.2831로 매우 높게 나타나며 어텐션 탈취가 성공적으로 이루어졌음을 명확히 보여준다. 반면에 L_{corr}만 적용한 경우 모든 경우에서 L_{act}보다 낮은 성능을 보인다.

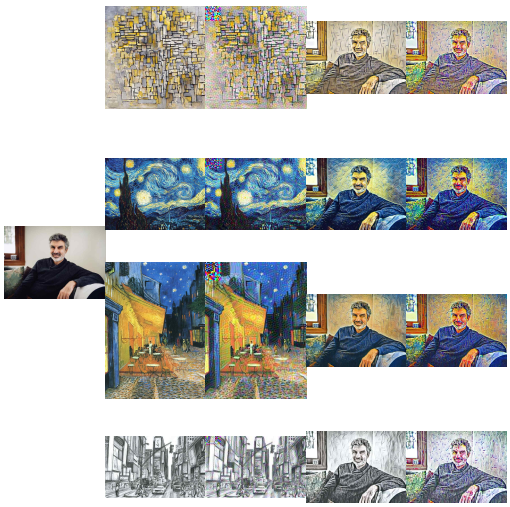
표 3. 제안하는 손실 함수별 기여도 측정
Table 3. Individual contribution of loss components. Best results in **bold.**

Configuration	STDR ↑	FCD ↑	AED ↑
L_{act} only	0.7536	1.8147	**1.2831**
L_{corr} only	0.5214	0.9690	0.3444
$L_{act}+L_{corr}$	**0.8902**	**3.1509**	1.2483

하지만 여기서 주목할 점은 영역별 할당을 통한 그들의 결합은 $STDR$ 점수 0.8902을 달성하며 L_{act}만 적용한 경우보다 약 18% 향상된 성능을 달성했다. 이는 두 손실함수의 시너지가 더 효과적인 이미지 보호를 수행할 수 있음을 의미한다. 즉, 어텐션 탈취와 특징 통계 오염이라는 두 가지 서로 다른 메커니즘이 공간적으로 분리되어 상호 보완적으로 작용하며 시너지를 창출했음을 의미한다.

4.3. 정성적 분석

그림 3은 우리의 기법이 원본 이미지(b)와 비교하여 시각적으로 거의 차이가 없는 보호 이미지(c)를 생성하면서도, 스타일 전송 결과(e)를 효과적으로 방해했음을 보여준다. 특히 원본 스타일이 가진 복잡한 질감(texture) 및 세밀한 선 표현이 결과물에서 대부분 소실되어, 스타일 이미지의 색상 정보만 전송되었음을 확인할 수 있다.



(a) (b) (c) (d) (e)

그림 3. 제안하는 보호기법의 시각적 비교. (a) 콘텐츠 이미지 I_c, (b) 스타일 이미지 I_s, (c) 보호된 스타일 이미지 I_u, (d) 보호 없이 스타일이 변환된 이미지 I_{cs}, (e) 제안 기법을 통해 보호된 스타일 변환 이미지 I_{cu}.

Fig. 3. Visual comparison of protection methods. (a) Content Image I_c, (b) Original Style Image I_s, (c) Untransferable Image I_u, (d) Original Transfer Result I_{cs}, (e) Disrupt Transfer Result I_{cu}.

이미지의 전반적인 색상 정보를 전송하는 데에는 비교적 약한 보호 효과를 보이는 이유에 대해 다음과 같은 가설을 제시할 수 있다. 우리의 방법론은 주로 VGG 네트워크의 중간 및 상위 계층에서 추출되는 특징 맵의 통계(Gram Matrix)를 교란하는 데 집중되어 있다. 이 특징들은 주로 객체의 형태, 질감과 같은 복잡하고 구조적인 스타일 정보를 인코딩한다. 반면, 이미지의 전반적인 색상 분포와 같은 저수준(low-level) 정보는 하위 계층에서 처리되거나 다른 메커니즘에 의해 전달될 수 있다. 따라서 우리의 공격이 텍스처와 스타일에 대해서는 매우 효과적이었으나, 색상 정보에는 상대적으로 영향을 덜 미쳤을 가능성이 있다. 이는 향후 연구에서 보호 범위를 색상 정보까지 확장하는 방향으로 개선될 수 있는 지점이다.

5. 논의

본 장에서는 제안하는 기법의 한계와 발전 방향에 대해 논의하고 저작권 보호를 위한 영역별 노이즈 생성의 광범위한 함의에 대해 논의한다.

5.1. 한계점 및 향후 연구 과제

본 연구는 경사도 간섭을 피하기 위해 보호 메커니즘을 공간적으로 분리하는 새로운 패러다임을 제시했다. 이 접근법은 적은 시각적 품질 손실로 높은 보호 효율을 달성하여, 기존 기법들이 가진 상충 관계 문제에 대한 실용적인 해결책을 제공한다. 이는 아티스트에게 법적 회색지대에 놓인 AI 스타일 도용에 대응할 수 있는 선제적인 기술적 도구를 제공한다는 점에서 중요한 의의를 가진다. 다만, 본 기법은 다음과 같은 한계와 잠재적인 발전 방향을 가진다.

1. **고정된 패치 위치:** 현재 구현은 패치 위치가 고정되어 있어, 적대자가 이 영역을 특정하여 무시하는 적응형 공격에 취약할 수 있다. 향후

연구에서는 이미지 콘텐츠에 기반한 동적 패치 배치 전략을 통해 강건성을 향상할 수 있을 것이다.

2. **탐지 가능성:** 패치 영역의 높은 활성화 값은 특화된 필터에 의해 탐지될 가능성이 있다. 이는 보호 기법의 존재를 노출시킬 수 있다.
3. **스타일 변환에 국한된 보호:** 본 연구는 NST에 초점을 맞추고 있다. 확산 모델(diffusion models) 등 다른 생성 AI를 통한 도용을 막기 위해서는 각 모델의 특성에 맞는 별도의 보호 전략이 필요하며, 본 연구의 영역별 최적화 원칙이 이러한 연구에 확장 적용될 수 있을 것이다.

5.2. 저작권 보호에 관한 법적 · 윤리적 시사점

본 연구의 스타일 보호를 위한 기술적 해결책은 인공지능과 저작권법 간의 관계에 중요한 고려사항을 제기한다. 스타일 추출을 법적으로 금지하기보다는 기술적으로 불가능하게 만듦으로써, 집행 메커니즘을 사후 법적 조치에서 선제적 기술 보호로 전환시킬 수 있음을 시사한다.

본 연구의 접근법은 다른 미디어의 디지털 저작권 관리(DRM)와 유사하지만 결정적인 차이점이 있다. 우리의 보호 기술은 인간의 시각적 접근은 보존하면서 특정 인공지능 기반의 악용만을 방지한다. 이러한 선택적 보호는 자동화된 도용은 차단하면서 인간의 감정과 영감은 허용하므로 공정 이용(fair use) 원칙과 일치한다. 그러나 우리는 이 해결책이 갖는 군비 경쟁과 같은 양상을 인지하고 있다. 스타일 변환 모델이 진화함에 따라 보호 기법도 그에 맞춰 적응해야 한다. 즉, 이와 같은 기법의 지속적인 연구를 잠재적으로 요구한다.

또한 우리의 연구는 인공지능 시대에 인간의 창의성을 보존하는 더 넓은 목표에 기여한다. 아티스트에게 자신의 독특한 스타일을 보호할 도구를 제공함으로써 우리는 독창적인 예술적 표현의 경제적, 문화적 가치를 유지하는 데 도움을 준다. 이러한 보호는 전문 예술가 경력을 유지하고 예술 분야의 지속적인 혁신을 장려하는 데 필수적이다.

6. 결론

본 논문은 NST를 통한 예술 스타일의 무단 도용이라는 저작권 위협에 대응하기 위한 새로운 보호 기법을 제시했다. 기존 보호 기법들은 보호 성능과 시각적 품질 간의 비효율적인 상충 관계라는 근본적 한계에 직면해왔다. 본 연구의 핵심 기여는 이러한 한계의 원인이 다중 목표 최적화 과정에서 발생하는 경사도 간섭에 있음을 통찰하고, 보호 목표의 공간적 분리라는 혁신적인 방법으로 이 문제를 해결한 것이다.

본 연구의 핵심 기여는 다음과 같이 요약할 수 있다. 첫째, 무단 스타일 변환을 방해하는 새로운 저작권 보호 프레임워크로서 **영역별 노이즈 생성 기법**을 제안했다. 둘째, 어텐션 탈취를 위한 활성화 손실과 특징 통계 오염을 위한 상관관계 손실을 각각 패치와 배경이라는 분리된 영역에 배타적으로 할당하는 **비간섭 최적화 전략**을 도입했다. 이를 통해 경사도 경쟁을 원천적으로 제거하여, 각 보호 메커니즘이 최대의 효율로 작동하도록 보장한다. 셋째, 실험을 통해 제안 기법이 **실용적인 수준의 시각적 품질(SSIM > 0.85)을 유지하면서도 스타일 변환을 약 90% 방해하는 효율적인 보호 균형**을 달성하여 기존의 상충 관계를 성공적으로 극복했음을 입증했다.

결론적으로 본 연구는 NST에 대한 스타일 전송 방어에 대한 강력한 패러다임으로서 영역별 최적화 전략을 정립한다. 이 원칙은 향후 다른 생성형 인공지능 모델에 대한 보호 기술 개발에

도 영감을 줄 수 있을 것으로 기대한다. 뿐만 아니라 본 연구가 제시하는 접근법이 인공지능 시대의 창작자 권리 보호를 위한 기술적 초석이 되기를 희망한다.

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-정보통신방송혁신인재양성사업의 지원을 받아 수행된 연구임 (IITP-2025-RS-2023-00266605, 40%); 본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2025년도 문화체육관광 연구개발사업으로 수행되었음(과제명: 블록체인 기술 기반 SW 저작권 보호를 위한 유통 · 관리 플랫폼 기술개발 및 인재양성, 과제번호: RS- 2023-00228867, 30%);본 연구는 과학기술정보통신부 및 정보통신기획평가원의 ICT 혁신인재4.0 사업의 연구결과로 수행되었음 (IITP-2022-RS-2022-00156310, 30%).

참 고 문 헌

[1] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. “Image style transfer using convolutional neural networks”, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2414-2423, 2016, DOI : http://dx.doi.org/10.1109/CVPR.2016.265

[2] Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. “Perceptual losses for real-time style transfer and super-resolution”, In European conference on computer vision, pp. 694-711, 2016, DOI : https://doi.org/10.48550/arXiv.1603.08155

[3] Huang, Xun, and Serge Belongie. “Arbitrary style transfer in real-time with adaptive instance normalization”, In Proceedings of the IEEE international conference on computer vision, pp. 1501-1510, 2017, DOI : https://doi.org/10.48550/arXiv.1703.06868

[4] Park, Dae Young, and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 5880 - 5888, 2019, DOI : https://doi.org/10.48550/arXiv.1812.02342

[5] Deng, Yingying, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. “Stytr2: Image style transfer with transformers”, In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11326-11336, 2022, DOI : https://doi.org/10.48550/arXiv.2105.14576

[6] Zhang, Chiyu, Xiaogang Xu, Lei Wang, Zaiyan Dai, and Jun Yang. “S2wat: Image style transfer via hierarchical vision transformer using strips window attention”, In Proceedings of the AAAI conference on artificial intelligence, vol. 38, no. 7, pp. 7024-7032, 2024, DOI : https://doi.org/10.48550/arXiv.2210.12381

[7] Jafari, Reza, and Mahsa Noori Sarcheshme. “Legal Frameworks for Protecting Digital Art and NFTs: Navigating Copyright and Ownership Rights in Virtual Spaces”, Legal Studies in Digital Age 2, no. 3 (2023): 25-36, DOI : https://doi.org/10.31941/pj.v22i3.5067

[8] Foerster, Hanna, Sasha Behrouzi, Phillip Rieger, Murtuza Jadliwala, and Ahmad-Reza Sadeghi. “{LightShed}: Defeating Perturbation-based Image Copyright Protections”, In 34th USENIX Security Symposium (USENIX Security 25), pp. 7271-7290. 2025.

[9] Zhang, Dawen, Boming Xia, Yue Liu, Xiwei Xu, Thong Hoang, Zhenchang Xing, Mark Staples, Qinghua Lu, and Liming

Zhu. “Privacy and copyright protection in generative AI: A lifecycle perspective”, In Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI, pp. 92-97, 2024, DOI : https://doi.org/10.1145/3644815.3644952

[10] Gatys, Leon A., Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. “Preserving color in neural artistic style transfer”, arXiv preprint arXiv:1606.05897, 2016, DOI : https://doi.org/10.48550/arXiv.1606.05897

[11] Addepalli, Sravanti, Samyak Jain, Gaurang Sriramanan, and R. Venkatesh Babu. “Scaling adversarial training to large perturbation bounds”, In European Conference on Computer Vision, pp. 301-316, 2022, DOI : https://doi.org/10.1007/978-3-031-20065-6_18

[12] Dewi, V., R. Njatrijani, and B. Rahmanda. “Legal protection of digital art copyrights on social media”, In Proceedings of the International Conference on Sustainability in Technological, Environmental, Law, Management, Social and Economic Matters, ICOSTELM, pp. 4-5, 2022, DOI : http://dx.doi.org/10.4108/eai.4-11-2022.2329343

[13] NARASIMHAN, Sharan; DEY, Suvodip; DESARKAR, Maunendra Sankar. Towards robust and semantically organised latent representations for unsupervised text style transfer. arXiv preprint arXiv:2205.02309, 2022, DOI : https://doi.org/10.48550/arXiv.2205.02309

[14] Luo, Xin, Wei Chen, Zhengfa Liang, Chen Li, and Yusong Tan. “Adversarial style discrepancy minimization for unsupervised domain adaptation”, Neural Networks, Vol. 157, pp. 216-225, 2023, DOI : https://doi.org/10.1016/j.neunet.2022.10.015

[15] Guo, Zhongliang, Junhao Dong, Yifei Qian, Kaixuan Wang, Weiye Li, Ziheng Guo, Yuheng Wang, Yanli Li, Ognjen Arandjelović, and Lei Fang. “Artwork protection against neural style transfer using locally adaptive adversarial color attack”, arXiv preprint arXiv:2401.09673, 2024, DOI : https://doi.org/10.3233/FAIA240643

[16] Li, Yaxin, Jie Ren, Han Xu, and Hui Liu. “Neural style protection: Counteracting unauthorized neural style transfer”, In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 3966-3975. 2024, DOI : https://doi.org/10.1109/WACV57701.2024.00392

[17] Bala, Aniruddha, Rohit Chowdhury, Rohan Jaiswal, and Siddharth Roheda. “Dct-shield: A robust frequency domain defense against malicious image editing”, arXiv preprint arXiv:2504.17894, 2025, DOI : https://doi.org/10.48550/arXiv.2504.17894

[18] Kang, Xu, and Bin Song. “Rect-ViT: Rectified attention via feature attribution can improve the adversarial robustness of Vision Transformers”, Neural Networks, Vol. 190, 2025, DOI : https://doi.org/10.1016/j.neunet.2025.107666

[19] Li, Yanghao, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. “Demystifying neural style transfer”, In Proceedings of the 26th International Joint Conference on Artificial Intelligence, pp. 2230-2236, 2017, DOI : https://doi.org/10.48550/arXiv.1701.01036

[20] Gardner, Steven, Oleg Golovidov, Joshua Griffin, Patrick Koch, Wayne Thompson, Brett Wujek, and Yan Xu. “Constrained multi-objective optimization for automated machine learning”, In 2019 IEEE International conference on data science and advanced analytics (DSAA), pp. 364-373, 2019, DOI : https://doi.org/10.48550/arXiv.1908.04909

[21] Neel, Seth, and Peter Chang. “Privacy

issues in large language models: A survey", arXiv preprint arXiv:2312.06717, 2023, DOI : https://doi.org/10.48550/arXiv.2312.06717
[22] Watiktinnakorn, Chawinthorn, Jirawat Seesai, and Chutisant Kerdvibulvech. "Blurring the lines: how AI is redefining artistic ownership and copyright", Discover Artificial Intelligence, Vol. 3, No. 37, 2023, DOI : http://dx.doi.org/10.1007/s44163-023-00088-y
[23] Shan, Shawn, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. "Glaze: Protecting artists from style mimicry by {Text-to-Image} models", In 32nd USENIX Security Symposium (USENIX Security 23), pp. 2187-2204, 2023, DOI : https://doi.org/10.48550/arXiv.2302.04222
[24] Li, Jiacheng, Ninghui Li, and Bruno Ribeiro. "{MIST}: Defending against membership inference attacks through {Membership-Invariant} subspace training", In 33rd USENIX Security Symposium (USENIX Security 24), pp. 2387-2404. 2024.
[25] Patrick K. Lin, Can This Data Poisoning Tool Help Artists Protect Their Work from AI Scraping?, Center for Art Lawhttps://itsartlaw.org/art-law/can-this-data-poisoning-tool-help-artists-protect-their-work-from-ai-scraping/.
[26] Zhang, Yunming, Dengpan Ye, Sipeng Shen, and Jun Wang. "StyleMark: A Robust Watermarking Method for Art Style Images Against Black-Box Arbitrary Style Transfer", arXiv preprint arXiv:2412.07129, 2024, DOI : https://doi.org/10.48550/arXiv.2412.07129

저 자 소 개

박성환(Sunghwan Park)

2019.2 중앙대학교 융합공학부 학사
2021.2 중앙대학교 융합보안학과 석사
2021.3-현재 : 중앙대학교 융합보안학과 박사과정
<주관심분야> 연합학습, 인공지능과 개인정보보호, 사이버-물리 시스템

김종성(Jongseong Kim)

2021.2 중앙대학교 융합보안학과 석사
2024.2 중앙대학교 융합보안학과 박사
2020.11-2025.1 한국저작권보호원 주임
2025.2-현재 : 중앙대학교 융합보안학과 연구교수
<주관심분야> 정보보호, 디지털저작권, 개인정보 보호, 산업보안법

오병훈(Byunghoon Oh)

2025.2 명지대학교 수학과 학사
2025.3-현재 : 중앙대학교 융합보안학과 석사과정
<주관심분야> AI 보안, 적대적 학습, 데이터 분석

황요한(Yohan Hwang)

2023.2 경기대학교 지식재산학과 학사
2024.9-현재 : 중앙대학교 융합보안학과 석사과정
<주관심분야> 저작권법, 산업보안법, 디지털포렌식

홍준호(Junho Hong)

2012.2 단국대학교 법학과 학사
2014.2 단국대학교 법학과 석사
2018.2 단국대학교 법학과 박사
2014.4-2024.8 한국정보보호산업협회 한국정보보호교육원 원장
2024.9-현재 : 성신여자대학교 융합보안공학과 조교수
<주관심분야> 정보보호, 디지털저작권, 개인정보보호법

이재우(Jaewoo Lee)

2006.2 서울대학교 컴퓨터공학 학사
2008.2 서울대학교 컴퓨터공학 석사
2017.2 University of Pennsylvania Computer and Information Science 박사
2018.3-현재 : 중앙대학교 산업보안학과 부교수
<주관심분야> 사이버-물리 시스템, 실시간 임베디드 시스템, 정보시스템 보안