생성형 AI 사진 조작이 개인 기억 회상에 미치는 위험성 및 대응 전략 연구*

차민정[01], 김나리[2], 이제희[1], 김채영[2], 박상근[1]
[1]경희대학교 소프트웨어융합학과, [2]경희대학교 산업경영공학과
{minjeongcha,kimnarikimnar,juventa23,cdfv,sk.park}@khu.ac.kr

A Study on the Risk of Personal Photo Manipulation Using Generative AI on Autobiographical Memory and Mitigation Strategies

Minjeong Cha[01], Nari Kim[2], Jehui Lee[1], Chaeyeong Kim[2], Sangkeun Park[1]
[1]Department of Software Convergence, Kyung Hee University
[2]Department of Industrial and Management Systems Engineering, Kyung Hee University

요 약

생성형 AI로 정교한 사진 조작이 가능해지면서 개인의 기억이 왜곡될 위험이 커지고 있다. 본 연구는 35명의 참여자를 대상으로 한 실험을 통해, AI로 조작된 개인 사진이 '기억 연쇄'를 방해하고 '허위 기억'을 형성할 수 있음을 실증적으로 밝혔다. 또한 소셜 미디어 공유로 인한 집단 기억 왜곡 위험에 대응하기 위해, CLIP(의미), SIFT(특징점), 픽셀 분석을 결합한 '다중 지표 융합 방식'의 실시간 조작 탐지 시스템을 제안한다. 본 연구는 AI 조작의 심리적 위험을 규명하고 기술적 대응책을 제시했다는 것에 의의가 있다.

1. 서 론

많은 사람들이 개인의 추억을 회상하기 위해 사진을 찍는다. 사진을 통해 기억을 보존하고, 일상을 기록하며, 타인과 그 경험을 공유하기도 한다[1, 2]. 소셜 미디어의 대중화는 이러한 사진의 역할을 변화시켰다. 순수한 기억의 보존 및 회상에서 적극적인 자기 관리(self-curation)로 사진의 역할이 변화하기 시작했다. 사람들은 소셜 미디어에서 다른 사람의 관심을 더 이끌거나[3] 더 이상적인 자신의 모습을 보여주기 위해 사진을 원하는 대로 편집하거나 꾸미기도 한다.

최근 생성형 AI의 등장은 이러한 패러다임에 급진적인 변화를 가져왔다. 이제 스마트폰에도 사진 편집을 도와주는 생성형 AI 도구가 기본으로 장착되기 시작했다. 이를 활용하면 누구나 간단하게 정교하면서도 사실적인 고급 사진 편집을 쉽게 수행할 수 있다[4]. 이는 기존에 굳건히 다져진 디지털 사진의 신뢰성에 근본적인 의문을 제기하게 되었다. 특히, 개인의 추억과 의미가 담긴 자서전적 사진이 본인이나 타인에 의해 미묘하게 조작될 경우, 이는 단순한 이미지 변형을 넘어 개인의 기억 왜곡이나 심지어는 없었던 일을 사실로 믿게 되는 허위 기억 형성의 잠재적 위험으로 이어질 수 있다.

기존 연구는 AI 조작 탐지 정확도가 약 60%에 머무르며 허위 기억 유발을 확인했으나[5, 6], 주로 개인 경험과 무관한 일반 이미지거나[7, 8, 9] 기억이 희미한 과거 사진을 대상으로 했다[10, 11]. 기술적으로도 SIFT[12]는 의미 조작에 취약했고, CLIP[13]은 미세 변조 감지에 약점을 보이는 '사각지대'가 존재했다. 본 연구는 이 한계를 넘어, 최근 개인 사진의 위험성을 탐색하고 의미, 구조, 픽셀을 융합한 시스템을 제안한다.

* "이 논문은 2025년도 정부 재원(과학기술정보통신부 여대학원생 공학연구팀제 지원사업)으로 과학기술정보통신부와 한국여성과학기술인육성재단의 지원을 받아 수행되었습니다. (WISET-2025-125호)"

본 연구는 35명의 참여자를 대상으로 개인의 추억이 담긴 사진에 생성형 AI를 활용한 정교한 조작을 가했을 때 발생할 수 있는 기억 왜곡, 허위 기억 등의 잠재적 위험을 탐색한다. 나아가 이러한 잠재적 위험에 대한 해결 전략으로, 온라인상에서 조작된 사진이 공유될 경우 이를 실시간으로 탐지하고 원본과 비교하여 경고하는 다중 지표 융합(Multi-metric Fusion) 방식의 사진 조작 탐지 시스템을 제안한다.

2. 관련 연구

2.1. AI 조작 이미지에 대한 인간의 인지적 취약성 탐구

인간의 기억은 사건 이후의 다양한 정보에 의해 왜곡되기 쉽다[14]. 이에 사람들은 단순한 기록을 넘어 개인의 추억을 정확하게 남기고 회상하기 위해 사진을 활용한다[1]. 하지만 사진이 조작된다면 개인의 기억이 왜곡될 수 있으며[15], 심지어 발생하지 않은 사건에 대한 생생한 허위 기억을 형성할 수도 있다[10, 11].

AI를 활용해 조작한 이미지에 대해, 인간의 조작 탐지 능력은 이를 따라가지 못하고 있다. 기존 연구들은 인간이 고품질의 AI 생성 또는 조작 이미지를 실제와 구별하는 데 어려움을 겪으며, 그 탐지 정확도는 약 60% 수준에 머무른다는 것을 확인했다[5, 6].

기존의 조작 인지 연구들은 대부분 개인의 경험과 무관한 일반적인 이미지(예: 유명인, 풍경)를 대상으로 수행되었다 [7, 8, 9] 개인 사진을 다룬 소수의 심리학 연구마저도 기억이 희미한 과거(예: 어린 시절)의 사진을 활용하여, 기억 왜곡 가능성이 이미 높은 조건에서 실험을 진행했다는 한계가 있다[10, 11]. 본 연구는 이러한 한계를 넘어, 생성형 AI을 활용해 개인이 최근에 촬영한 사진을 조작하고 그 탐지 정확도 및 이에 기반한 위험 요소를 탐색한다.

2.2. 조작 탐지 기법 연구

생성형 AI로 인한 조작의 위협에 대응하기 위해 다양한 기술적 탐지 기법들이 연구되어 왔다. SIFT, SURF와 같은 전통적인 특징점 기반 방식은 이미지의 회전, 크기, 시점 변화 등 기하학적 변형에도

강한 특징점을 추출하여 이미지의 구조적 동일성을 검증한다 [12, 13, 16]. 하지만 이러한 기술들은 특징점만을 비교하므로 전체적인 맥락을 이해하지 못하며, 객체의 의미적 변경, 삭제, 추가 등 생성형 AI 기반 조작 탐지에는 매우 취약한 한계를 보인다.

CLIP과 같은 최신 딥러닝 기반 의미 분석 방식은 의미와 맥락을 이해하여 배경이 바뀌는 등의 조작에도 강하다는 장점이 있다 [13]. 하지만 특히 CLIP은 전역적 특징에 집중하므로, 미세한 픽셀 변조나 국소적 불일치를 감지하는 데는 약점을 보인다.

이처럼 기존 분석 방식은 명확한 강점과 동시에 특정 조작 유형에 취약한 '사각지대'가 존재한다. 따라서 본 연구는 다양한 조작에 강건하게 대응하기 위해, 각기 다른 강점(의미, 구조, 픽셀)을 가진 모델을 융합하는 하이브리드 접근법, 즉 다중 지표 융합(Multi-metric Fusion) 방식의 시스템을 제안한다.

3. 본론

3.1 생성형 AI를 활용한 개인 사진 조작 인지 실험

그림 1. 원본(좌)과 조작본(우) 예시

본 연구는 생성형 AI를 활용한 사진 조작이 개인의 자서전적 기억에 미치는 영향을 탐구하기 위해 35명의 참가자를 모집했다. 각 참여자는 연구진에게 30개의 개인 사진을 제공했으며, 온라인 인터뷰를 통해 사진마다 육하원칙에 의거한 사진에 대한 설명을 받았다. 연구진은 생성형 AI(Gemini, Adobe Photoshop)을 활용해 각 참여자로부터 받은 원본 사진 30장 중 20장을 조작하여 총 25개의 조작 사진을 생성했다 [그림 1]. 일주일 뒤 참여자들을 다시 불러 개인 사진을 보여주며 해당 사진에 대한 설명을 다시 요청하고, 해당 사진이 원본인지(Positive)인지 조작 사진인지(Negative) 맞추도록 했다. 1인당 원본 25장 + 조작 25장의 사진에 대해 총 50회의 사진 조작 판별 실험을 진행한 결과는 [표 1]과 같다.

표 1. 원본 판별 결과 (총 1,750건 = 35명×50장)

	원본이라고 예측	조작이라고 예측
원본 사진	756 (43.2%)	120 (6.9%)
조작 사진	199 (11.4%)	675 (38.6%)

실험 결과, 참가자들은 개인의 추억과 관련된 사진임에도 불구하고 18.3%의 판단 오류를 보였다. 특히 조작된 사진을 원본으로 인식하는 FP(False Positive) 비율이 11.4% 임을 통해 AI 조작이 개인의 기억에 직접적인 위협이 될 수 있음을 확인했다.

사후 인터뷰를 통해, 생성형 AI를 활용한 정교한 사진 조작으로 발생할 수 있는 잠재적 위험 요소를 발견했다. 가장 큰 위험성으로 '허위 기억 형성' 현상이 발견되었다. 조작된 사진을 원본이라고 판단한 경우, 해당 사진에 대한 원본 사진을 보여줬을 때도 원본 사진이 조작된 것으로 판단하는 사례 총 23건(19명) 존재했다. 또한, 조작된 사진은 기억 회상의 풍부함을 유발하는 '기억 연쇄(Memory Chain)'효과를 단절시켜서, 조작되지 않았다면 더 떠올릴 수 있는 추억 회상을 방해하는 것으로 나타났다.

실험 전후로 일주일의 간격을 두고 실시한 인공지능(AI) 불안 지수(AIA)[17] 설문에서도, 참여자들의 AI로 인한 불안 지수가 본 실험 후에 유의미하게 증가함을 확인했다 ($p<0.001$).

3.2. 생성형 AI 기반 조작 사진 탐지 모델 개발

타인과 함께 촬영한 사진을 타인이 정교하게 조작해서 소셜 미디어에 올리면, 함께 촬영한 다른 사람들이 이 사진을 볼 때 기억이 왜곡될 위험이 있다.

본 연구에서는 다양한 기법을 활용해 앞에서 수집한 원본 1,050장 중 조작된 특정 사진 한 장의 진짜 원본을 찾아내는 테스트를 수행했다. 이를 기반으로, 생성형 AI를 활용해 정교하게 조작된 사진의 원본 사진을 보다 정확하게 찾아낼 수 있는 하이브리드 모델을 개발했다.

(a) (b) (c)

그림 2. 원본(좌)과 조작 사진(우) 예시

3.2.1. CLIP 기반 탐지 (Accuracy 98.5%)

의미 기반 분석 모델인 CLIP[13]을 활용한 원본 이미지 탐지는 사진의 배경이나 스타일 등 대대적인 조작이 가해진 사진의 원본 사진도 매우 높은 정확도로 탐지해냈다. 하지만 이미지의 핵심 객체가 의미적으로 크게 변경된 경우 (예: 그림 2a), 오히려 전체적인 맥락이 유사한 다른 대상을 원본으로 오인하는 오류가 발생하는 한계가 있다.

3.2.2. SIFT 모델 기반 탐지 (Accuracy 97.5%)

특징점 기반 분석 모델인 SIFT[12]는 이미지의 회전, 크기 변화 등 구조적 변형 탐지에 강점을 보였다. 그러나 특징점의 국소적인 모양만을 비교하기 때문에, 조작된 사진에서 객체의 특징이 많이 변형된 경우 (예: 그림 2b) 오히려 의미는 다르지만 구조적으로 유사한 객체 간에 '가짜 매칭'을 생성하여 원본을 오인하는 한계가 있다.

3.2.3. 픽셀 기반 탐지 (Accuracy 97.0%)

픽셀 차이 분석[16]은 일부 객체 추가/삭제와 같이 이미지 대부분이 원본과 동일한 조작 유형 탐지에 효과적이었다. 하지만 조작 과정에서 이미지가 일부 잘려나가거나 크기가 변경되는 등 구조적 변형이 큰 경우 (예: 그림 3c) 오히려 색감이나 질감이 비슷한 다른 사진을 원본으로 오인하는 오류가 발생하는 한계가 있다.

3.2.4. 하이브리드 모델을 활용한 탐지 (Accuracy 99.5%)

앞선 세 모델은 각기 다른 유형의 조작에 취약한 '사각지대'가 존재한다. 이 한계를 극복하기 위해, 세 지표(의미, 특징점, 픽셀)

를 결합한 '다중 지표 융합(Multi-metric Fusion)' 방식을 사용한 하이브리드 모델을 개발했다. 그 결과, 각 모델의 한계를 상호 보완하여 가장 높은 탐지 정확도를 기록했다. 이는 복잡하고 다양한 AI 사진 조작을 탐지하는 데 있어 하이브리드 접근법이 효과적일 수 있음을 나타낸다.

3.3. 탐지 서비스 제안

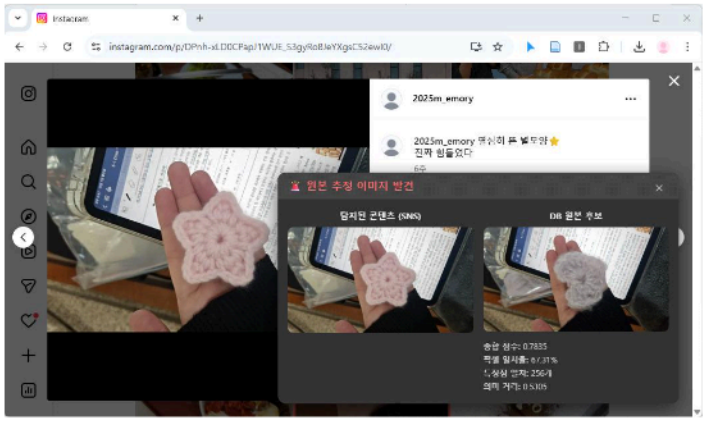

그림 3. 조작 사진 탐지 시스템 동작 화면

본 연구에서 개발한 하이브리드 모델을 기반으로, 소셜 미디어에 공유된 조작된 개인 사진으로부터 사용자를 보호하기 위한 크롬 확장 프로그램을 구현했다 [그림 3]. 이 시스템은 사용자가 인스타그램과 같은 소셜 미디어를 이용할 때 화면 속 이미지의 URL을 자동으로 서버에 전송한다. 서버의 하이브리드 모델은 이 이미지를 사용자의 로컬 사진 데이터베이스와 실시간으로 비교 분석하여, 해당 이미지가 사용자의 원본 사진을 조작한 것으로 판단되면 팝업 경고 창을 통해 그 위험을 즉시 알린다.

4. 결론

본 연구는 생성형 AI가 개인의 자서전적 기억에 미치는 심리적 위험을 실증적으로 규명하고, 이에 대한 기술적 해결책을 함께 제시했다. 사용자 연구(N=35)를 통해, 정교한 사진 조작이 기억 회상의 풍부함을 저해할 뿐만 아니라, 실제 경험을 덮어쓰는 '허위 기억'을 형성할 수 있음을 밝혔다. 이에 대응하기 위해, 본 연구는 CLIP, SIFT, 픽셀 분석을 융합한 '다중 지표 융합'시스템을 개발하여, 사용자가 소셜 미디어에서 접하는 조작 이미지를 실시간으로 탐지하는 기법을 제안했다. 본 연구 결과는 디지털 사진을 더 이상 신뢰하기 어려운 시대를 맞아, 새로운 디지털 리터러시 교육과 기술적 안전장치 마련이 시급함을 강조한다.

참고 문헌

[1] Fawns et al. Cued recall: Using photo-elicitation to examine the distributed processes of remembering with photographs. Memory Studies, 16(2), 264-279. 2023.

[2] Soares et al. Exploring functions of and recollections with photos in the age of smartphone cameras. Memory Studies, 15(2), 287-303. 2022.

[3] Bakhshi et al. Why We Filter Our Photos and How It Impacts Engagement. Proceedings of the International AAAI Conference on Web and Social Media, 9(1), 12-21. 2021.

[4] Jess Weatherbed. Apple is 'concerned' about AI turning real photos into 'fantasy'. The Verge. Oct 23, 2024.

[5] Nightingale et al.. Can people identify original and manipulated photos of real-world scenes?. Cognitive Research: Principles and Implications. 2017.

[6] Schetinger et al. Humans are easily fooled by digital images. Computers & Graphics, 68, 142-151. 2017.

[7] Pataranutapornet al. Synthetic human memories: AI-edited images and videos can implant false memories and distort recollection. CHI '25: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Article 538, 1-20. 2025.

[8] Xu et al. Combating Misinformation in the Era of Generative AI Models. In Proceedings of the 31st ACM International Conference on Multimedia (MM '23), 9291 - 9298. 2023.

[9] Corvi et al. On the detection of synthetic images generated by diffusion models. ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

[10] Wade et al. A picture is worth a thousand lies: Using false photographs to create false childhood memories. Psychonomic Bulletin & Review 9, 597-603. 2002.

[11] Garry et al. When photographs create false memories. Current Directions in Psychological Science 14.6: 321-325. 2005.

[12] Lowe et al. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60, 91-110. 2004.

[13] Radford et al. Learning transferable visual models from natural language supervision. International conference on machine learning. PmLR. 2021.

[14] Loftus et al. When a lie becomes memory's truth: Memory distortion after exposure to misinformation. Current Directions in Psychological Science. 1992.

[15] Loftus et al. Reconstruction of automobile destruction: An example of the interaction between language and memory Journal of Verbal Learning and Verbal Behavior. 1974.

[16] Lukas et al., Digital camera identification from sensor pattern noise, in IEEE Transactions on Information Forensics and Security. vol. 1, no. 2, pp. 205-214. June 2006.

[17] Wang et al. Development and Validation of an Artificial Intelligence Anxiety Scale: An Initial Application in Predicting Motivated Learning Behavior. Interactive Learning Environments. 30(4). 619-634. Taylor & Francis, Abingdon, UK. 2022.