산업 가시적 워터마크의 취약성과 출처 표시 체계의 보완 필요성

▶ 생성형 AI 콘텐츠가 확산되면서 딥페이크 등 실제와 구분하기 어려운 콘텐츠에 대한 우려가 커지고 있다. 이에 한국, 미국, EU 등 주요국은 AI 산출물에 워터마크 표시를 의무화하거나 관련 기준을 도입하고 있으나, 동시에 워터마크 제거 도구의 상업화와 운영체제 기본 기능을 활용한 제거 사례도 늘어나며 제도의 실효성에 의문이 제기되고 있다. 가시적 워터마크는 제거 기술의 고도화와 낮은 이용 장벽으로 쉽게 삭제될 수 있는 반면, C2PA 메타데이터는 제거 이후에도 대부분 유지돼 출처 추적에 활용될 수 있다. 다만 일부 편집 도구가 내보내기 과정에서 해당 메타데이터를 삭제하는 사례도 확인돼, 편집·유통 전 과정에서 이를 안정적으로 보존할 수 있는 기술적 지원과 제도적 보완이 함께 필요하다는 지적이 나온다.

산업 엣지 워터마킹과 스트리밍 불법복제 대응

▶ 라이브 스포츠 중계를 중심으로 한 불법복제 피해가 산업 전반의 구조적 위협으로 확산되면서, 기존의 DRM 방식만으로는 유출 이후의 경로를 추적하기 어렵다는 한계가 부각되고 있다. 이에 시나미디어는 2026년 3월 콘텐츠 배포의 최종 계층인 CDN 엣지에서 이용자별 식별자를 실시간으로 삽입하는 엣지 워터마킹 솔루션을 공개했다. 이는 기존 A/B 워터마킹의 이중 스트림 비용 구조와 엣지 환경의 기술적 제약을 극복한 사례로, 배포 최종 구간에서 탐지와 집행이 실시간으로 연계될 수 있는 기술적 기반을 마련했다는 사례로 평가된다. 다만 탐지 기술의 고도화만으로는 실효적 보호에 한계가 있으며, 집행 체계와의 연계 및 국제 공조가 함께 갖춰져야 한다는 과제가 함께 제기되고 있다.

산업 사운드온의 변형 음원 탐지 도구 도입과 음원 유통 플랫폼의 검증 체계 강화

▶ 음원의 박자·음정 등을 변형해 별개의 곡처럼 등록하는 사례가 늘면서, 일각에서는 원곡에 대한 저작권 침해 우려가 제기되고 있다. 2024년, 틱톡과 유니버설뮤직그룹의 계약에 변형 음원 보호 조치가 포함되고, 멀린과의 협상이 중단된 것이 대표적이다. 이에 틱톡의 음원 유통 플랫폼 사운드온은 ACR클라우드의 변형 음원 탐지 도구를 도입해, 음원 등록 시점에 원곡과의 유사 여부를 점검하고 있다. 여기에 틱톡의 콘텐츠 스캔 인프라를 활용한 원곡 무단 활용에 대한 검증, 등록자 신원 확인, 담당자 직접 검토를 결합한 다층적 검증 체계도 운영 중이다. 유통 플랫폼이 음원의 적법성까지 검증·관리하는 역량을 요구받는 흐름이 나타나고 있다.

저작권 이슈 브리프

SUMMARY | 산업/기업 | 기술

산업 AI 음악 생성 플랫폼 수노의 저작권 필터 취약성 논란

▶ 2026년 4월 미국 기술 매체 더 버지가 AI 음악 생성 플랫폼 수노의 저작권 필터 취약성을 보도하며 논란이 확산되었다. 수노는 저작권이 있는 곡이나 가사의 사용을 금지하고, 해당 곡이나 가사가 업로드되면 이를 인식해 차단하는 저작권 필터를 이용하고 있으나, 더 버지는 간단한 조작으로 필터를 우회하거나 필터가 아예 작동하지 않는 사례를 확인하였다. 수노는 음원 업로드 시점에만 필터를 적용하고 생성된 음원을 재검사하지 않아, 수노에서 무단으로 제작된 AI 커버가 스트리밍 플랫폼에 유입되어 수익화될 수 있는 경로가 열려 있다. 주요 스트리밍 플랫폼은 AI 커버에 대응하는 조치를 시행 중이나 이를 위한 기술적 한계를 인정하고 있으며, 더 버지는 수노 뿐 아니라 배포·스트리밍 플랫폼을 포함한 전체 시스템에 문제가 있다고 지적하였다.

산업 이르데토, 대규모 OTT 환경을 위한 포렌식 워터마킹 기술 공개

▶ 네덜란드의 디지털 보안 솔루션 기업 이르데토는 대규모 OTT 환경을 위한 포렌식 워터마킹 기술 트레이스마크를 공개했다. 이 기술은 영상에 비가시적 식별 정보를 삽입하여 유출 경로를 추적하며, 재인코딩, 화면 캡처 등 불법복제 과정의 변형에도 워터마크가 유지된다. 핵심 기술인 균일 전환 식별자 방식은 기존 A/B 워터마킹의 한계를 극복하여 일부 비트가 손실되어도 식별 정확도가 균등하게 유지되며, 영상 일부만 확보된 상황이나 콜루전 공격에도 안정적으로 대응한다. 트레이스마크는 단순 워터마크 삽입을 넘어 유출 콘텐츠 모니터링, 유출자 식별, 법적 조치까지 연계하는 통합 체계를 구축했다. 미국 방송 인프라 기업 하모닉이 자사 스포츠 스트리밍 솔루션에 이 기술을 통합한 사례는 실전급 워터마킹 기술이 OTT 산업 표준으로 자리잡을 가능성을 보여준다.

기술 주간 기술 동향

▶ 생성형 AI 기술의 발전으로 딥페이크가 급증하면서 개인의 디지털 정체성 보호가 중요한 과제로 떠올랐다. 기존 워터마킹 기술은 단일 AI 도구에만 적용 가능하고 편집 과정에서 쉽게 손상되며, AI 생성 여부만 판별할 뿐 구체적인 출처를 추적하지 못한다. SAiW는 콘텐츠 생성 단계부터 출처 정보를 삽입하는 사전 방어 기술이다. 어떤 AI 도구가 언제 만들었는지 구체적인 정보를 이미지에 기록하며, 육안으로 구별이 불가능할 정도로 비가시성을 유지한다. JPEG 압축, 가우시안 노이즈, 필터 적용 등 다양한 공격에도 98% 이상의 추출 정확도를 보이며, 단일 시스템으로 여러 AI 모델의 출처를 동시에 식별할 수 있다. 다만 적대적 공격 방어, 실시간 처리 속도 개선, 멀티모달 확장 등이 향후 과제로 남아 있다.

가시적 워터마크의 취약성과 출처 표시 체계의 보완 필요성

워터마크 표시 확대와 제거 도구의 동시 확산

• AI 산출물에 대한 워터마크 표시 도입의 확대

- 생성형 AI로 제작된 영상과 이미지가 확산되면서 딥페이크 등 실제와 구분이 어려운 콘텐츠에 대한 우려가 커지고 있음. 이에 대응해 AI 산출물의 출처를 식별하기 위한 워터마크 표시 제도가 주요국에서 도입되고 있음
- 한국은 2026년 1월 시행된 AI 기본법을 통해 AI 사업자에게 워터마크 표시를 의무화했고, 미국은 C2PA*를 기술 표준으로 제시하면서 오픈AI(OpenAI), 구글(Google) 등 주요 사업자가 이를 채택하고 있음. EU도 2026년 8월부터 워터마크 표시 의무를 포함한 투명성 규정을 시행할 예정임
- 이들 제도는 공통적으로 AI 사업자가 콘텐츠 생성 단계에서 워터마크를 표시하도록 하는 구조를 취하고 있으나, 생성 이후 편집·재가공·유통되는 과정에서도 워터마크가 잘 유지되도록 할 장치는 아직 충분히 마련되지 않은 상태임

* C2PA(Coalition for Content Provenance and Authenticity): 디지털 콘텐츠에 비가시적 메타데이터를 삽입해 AI 생성 여부와 편집 이력을 추적하는 개방형 기술 표준

• 워터마크 제거 도구의 상업화와 접근성 확대

- 이러한 구조적 공백은 제도 시행 초기부터 드러났는데, 한국에서는 AI 기본법 시행 보름 만인 2026년 2월, 유튜브(YouTube) 등 소셜 미디어를 중심으로 오픈AI의 영상 생성 도구 소라(Sora)로 생성한 콘텐츠의 워터마크를 제거하는 방법을 소개하는 게시물이 확산됨
- 제거 수단도 빠르게 다양화되고 있음. 2026년 3월에는 클린비디오AI(CleanVideoAI)가 소라, 비오(Veo) 등 AI 영상 플랫폼용 워터마크 전용 제거 엔진을 출시하는 등 전문 도구의 상업화가 진행되고 있음
- 한편 전문 도구 외에 별도 소프트웨어 없이 운영체제에 내장된 기본 편집 기능만으로도 손쉽게 제거가 가능한 것으로 나타남
- 이는 생성 단계에서 워터마크를 표시하는 방식만으로는 유통 과정에서 출처 식별을 안정적으로 유지하기 어렵다는 점을 보여줌

가시적 워터마크 제거 기술의 발전

• AI 기반 프레임 복원 기술의 작동 원리

- 최근 등장한 워터마크 제거 도구는 해당 영역을 흐리게 처리하거나 잘라내던 기존 방식과 달리, AI가 워터마크 주변의 색상·윤곽·배경을 분석한 뒤 해당 영역을 자동으로 복원하여 프레임을 재구성함
- 기술이 발전하면서 정지 이미지뿐 아니라 동영상에서도 활용 범위가 확대됨. 동영상의 경우에는 장면 전환 흐름을 분석해, 워터마크가 있던 구간을 전후 프레임과 자연스럽게 연결되도록 복원할 수 있음
- 이 기술은 원래 영화 복원 분야에서 개발된 인페인팅(inpainting)* 기법이 AI 기반으로 고도화된 것으로, 최근 워터마크 제거 용도로 활용 범위가 확대되고 있음

* 인페인팅(inpainting): 이미지나 영상에서 손상되거나 제거가 필요한 영역을 주변 정보에 기반하여 자동 복원하는 기술

• 제거 수단의 낮은 진입 장벽

- 가시적 워터마크의 취약성은 제거 기술의 고도화에서만 비롯되는 것이 아니라, 전문 지식 없이도 손쉽게 표시를 없앨 수 있을 만큼 제거 수단의 접근성이 높아졌다는 점과도 관련이 있음
- 실제로 윈도우(Windows)의 '생성형 지우기(Generative Erase)'와 맥(macOS)의 '클린업(Clean Up)' 등 운영체제에 내장된 기본 편집 기능만으로도 워터마크를 제거한 사례가 확인됨
- 전문 제거 도구 역시 브라우저에서 자동으로 처리하는 방식, 이용자가 제거 영역을 직접 지정하는 온라인 편집기, 대량의 영상을 일괄 처리하는 API 방식 등으로 다양하게 제공됨
- 해외 온라인 커뮤니티인 레딧(Reddit) 등에서는 매직이레이저(MagicEraser), 바일로(Bylo.ai) 등 유·무료 제거 도구가 공유되고 있으며, 화면 잘라내기나 화면 재구성을 통해 워터마크가 포함된 부분을 제거하는 단순 편집 방식도 함께 활용됨

[표1] 가시적 워터마크 제거 방식 유형별 비교

구분	브라우저 기반 AI 인페인팅 도구	온라인 편집기	API 기반 자동화
처리 속도	빠름(1~10분)	중간(가변적)	빠름(자동화)
출력 품질	정적 배경에서 양호	조정 가능	일정한 품질
파일 크기 제한	50~100MB	일부 500MB 이상	요금제에 따라 상이
출력물 상태	워터마크 없음(무료 등급 기준)	워터마크 없음	워터마크 없음
적합 용도	단건 SNS 클립	정밀 제어가 필요한 작업	대량 배치 처리

출처: WaveSpeed, 'Sora Watermark Remover Online: No Download Needed', 2026.03.19., https://wavespeed.ai/blog/posts/sora-watermark-remover-online/

비가시적 식별 수단을 활용한 출처 표시 체계의 보완 방향

• C2PA 메타데이터의 기술적 역할과 보존 한계

- 앞서 살펴본 것처럼 눈에 보이는 워터마크는 다양한 방법으로 제거할 수 있음. 그러나 C2PA 메타데이터는 콘텐츠에 눈에 보이지 않는 형태로 삽입되어 있어, 워터마크를 지우거나 영상을 다시 저장하더라도 대부분 그대로 유지됨

- 또한 C2PA는 메타데이터를 삭제하거나 변경하려는 시도 자체가 이력으로 기록되므로, 가시적 워터마크가 제거된 이후에도 해당 콘텐츠가 AI로 생성된 것인지 여부와 편집 이력을 확인할 수 있음
- 다만 일부 온라인 편집기가 내보내기 과정에서 C2PA 메타데이터를 경고 없이 삭제하는 것으로 확인되었으며, 이는 비가시적 출처 표시 수단 역시 유통 경로에 따라 보존 여부가 달라질 수 있음을 보여줌

• 생성 단계와 유통 단계를 연결하는 출처 표시 설계의 필요성

- 가시적 워터마크와 C2PA 메타데이터를 함께 적용하면, 한쪽이 제거되더라도 출처 추적 가능성을 일부 유지할 수 있어 단일 수단보다 효과적임
- 다만 이러한 이중 구조가 실제로 작동하려면 콘텐츠의 편집, 내보내기, 유통 전 과정에서 메타데이터가 보존될 수 있도록 편집 도구와 유통 플랫폼 차원의 기술적 지원이 병행돼야 함
- 아울러 전문가들은 워터마크 삽입 의무화만으로는 유통 단계에서의 출처 식별을 보장하기 어렵다고 지적하며, 제거·우회 행위에 대한 규율 도입 필요성을 제기하고 있음. 다만 제재 대상과 적용 범위를 어떻게 설정할 것인지에 대해서는 업계와 정책 당국 간 추가 논의가 필요한 상황임

참고문헌

• CleanVideoAI, "CleanVideoAI Launches AI Engine to Remove Watermarks from Sora, Veo, and CapCut Videos", GlobeNewswire, 2026.03.19., https://www.globenewswire.com/news-release/2026/03/19/3259081/0/en/CleanVideoAI-Launches-AI-Engine-to-Remove-Watermarks-from-Sora-Veo-and-CapCut-Videos.html
• 손슬기, "'AI 기본법' 시행 보름 만에…워터마크 제거법 SNS 확산", 디지털투데이, 2026.02.06., https://www.digitaltoday.co.kr/news/articleView.html?idxno=627511
• WaveSpeed, "Sora Watermark Remover Online: No Download Needed", WaveSpeed, 2026.03.19., https://wavespeed.ai/blog/posts/sora-watermark-remover-online/

엣지 워터마킹과 스트리밍 불법복제 대응

불법복제 확산과 기존 대응 방식의 한계

• 불법복제의 조직화와 사후 단속 중심 대응

- 스포츠 중계·방송 업계의 연간 불법복제 피해 규모는 약 280억 달러(원화 약 42조 1,904억 원)[1]에 달하는 것으로 추산되며[2], 이는 단순한 권리 침해를 넘어 중계권 가치 하락과 구단 수익 감소로 이어지는 구조적 위협으로 인식됨
- 단편적 구독 서비스 구조와 높은 이용 비용이 불법복제의 주요 유인으로 지목되는 가운데, 단순 계도 중심의 대응만으로는 실질적인 억제 효과가 제한적이라는 비판이 제기됨
- 특히 방송 시점에 수익 가치가 집중되는 스트리밍 콘텐츠의 특성상, 불법복제물을 신속하게 차단하더라도 이미 불법 영상 유통과 수익화가 진행된 뒤 집행된다는 한계가 존재함
- 또한, 불법 스트리밍 플랫폼은 도메인 교체와 인프라 재구성만으로 단기간 내 재가동할 수 있어, 단속 이후에도 유사 서비스가 반복적으로 등장하는 구조가 형성됨. 이에 따라 기존 대응 방식은 사후 차단을 반복하는 구조에 머무른다는 비판이 제기됨

• 기존 보호 수단의 한계와 다층 보호 체계로의 변화

- 불법복제가 사후 단속만으로 억제되지 않는 원인 중 하나는, 콘텐츠 보호의 핵심 수단인 DRM (Digital Rights Management)*이 구조적으로 '유출 이후'의 경로를 추적하는 기능을 갖추고 있지 않기 때문임
- 이에 따라 스포츠 중계·방송 업계에서는 유출 이후의 경로 추적을 중심으로 ▲ 포렌식 워터마킹 (Forensic Watermarking)** ▲ 실시간 모니터링 ▲ 이상 트래픽 사전 감지 ▲ 수사기관 연계 등을 결합한 다층적 대응 체계로 변화하는 흐름이 나타나고 있음
- 일부 주요 스트리밍 사업자는 관련 전담 조직과 기술을 결합한 통합 대응 구조를 이미 운영하고 있으며, 단일 기술만으로는 스트리밍 콘텐츠 보호에 한계가 있다는 인식이 업계 내에서 확산되는 추세임

* DRM(Digital Rights Management): 디지털 콘텐츠의 무단 복제·배포를 방지하기 위해 암호화 등의 방법으로 접근을 통제하는 기술 및 관리 체계

** 포렌식 워터마킹(Forensic Watermarking): 콘텐츠에 세션·사용자별 고유 식별자를 비가시적으로 삽입해, 유출 발생 시 누출 경로와 책임 주체를 추적·식별할 수 있도록 하는 기술

1) 1달러=1,508.80원(KEB 하나은행 매매기준율 적용, 2026.04.01., 이하 동일)
2) Office of the United States Trade Representative(USTR), 2025 Review of Notorious Markets for Counterfeiting and Piracy, 2026.03.03., https://ustr.gov/sites/default/files/files/Press/Releases/2026/2025%20Notorious%20Markets%20List%20(final).pdf

엣지 워터마킹의 등장과 기술적 차별점

• 기존 워터마킹 방식의 한계와 엣지 워터마킹의 등장

- 전통적인 포렌식 워터마킹 방식인 A/B 워터마킹*은 동일 콘텐츠를 두 가지 버전으로 제작해 CDN**에 저장한 뒤, 사용자 세션별로 이를 조합해 제공하는 방식임. 이 과정에서 추가적인 대역폭과 저장 자원이 필요하며, 인프라 비용이 지속적으로 발생하는 한계가 존재함
- 이에 대한 대안으로 콘텐츠 배포의 최종 구간인 엣지(edge)*** 서버에서 사용자 식별자를 직접 삽입하는 엣지 기반 워터마킹 방식이 제안되었으나, 엣지 서버의 제한적인 성능과 DRM 암호화 키가 엣지에 존재하지 않는 구조적 문제로 인해 실제 구현이 어려웠음
- 시나미디어(Synamedia)는 헤드엔드(Headend)****에서 콘텐츠를 한 번 인코딩한 뒤 최종 배포 단계에서 사용자 세션별 식별자를 실시간으로 삽입하는 방식으로 이러한 한계를 개선한 엣지 워터마킹 기술을 구현함

* A/B 워터마킹(A/B watermarking): 동일한 콘텐츠에 서로 다른 식별 정보를 삽입한 두 가지 버전(A/B)을 생성한 뒤, 사용자별로 조합해 제공함으로써 유출 시 해당 콘텐츠의 배포 경로 또는 이용자를 추적할 수 있도록 하는 포렌식 워터마킹 기법

** CDN(Content Delivery Network): 콘텐츠를 사용자와 가까운 서버에 분산·배포해 전송 속도와 안정성을 높이는 네트워크 인프라

*** 엣지(edge): CDN을 구성하는 서버 중 최종 이용자와 가장 가까운 위치에 배치된 서버 계층으로, 콘텐츠를 중앙 서버에서 직접 전송하는 대신 이용자 인근의 엣지 서버를 통해 제공함으로써 지연 시간을 최소화함

**** 헤드엔드(Headend): 방송·스트리밍 시스템에서 콘텐츠를 수신·처리·인코딩해 배포망으로 송출하는 중앙 처리 시설

• 엣지 워터마킹의 성능과 콘텐츠 유형별 효과

- 시나미디어의 엣지 워터마킹 기술을 적용한 콘텐츠는 워터마크 삽입 및 추출 시간이 A/B 방식 대비 절반 수준으로 단축되며, 공개된 사례 기준으로 콘텐츠 유출 발생부터 해당 세션 차단까지 약 5분 이내 대응이 가능한 것으로 소개됨
- 이를 통해 두 개의 파일을 별도로 운용할 필요 없이 단일 파일로 처리할 수 있어 서버 간 데이터 전송량과 저장 부담이 줄어들며, 이는 인프라 운용 비용 절감으로 이어질 수 있음
- 라이브 방송에서는 식별 정보를 더 촘촘하게 삽입하여 짧은 구간 분석만으로도 유출자를 빠르게 특정할 수 있으며, 다른 CDN 사업자와도 연동할 수 있어 특정 인프라에 종속되지 않는 구조를 갖춤

• 스트리밍 콘텐츠 보호에서 엣지 워터마킹이 갖는 의의

- DRM이 비인가 이용자의 접근을 차단하는 데 집중하는 반면, 엣지 워터마킹과 같은 포렌식 워터마킹은 유출 주체를 식별하고 경로를 추적하는 기능을 담당하여 상호 보완적으로 작동함
- 이번 기술의 의의는 콘텐츠가 이용자에게 전달되는 최종 구간에서 이용자 단위 식별이 실시간으로 이루어진다는 데 있으며, 이를 통해 탐지 결과가 세션 차단 등 후속 집행 조치와 즉시 연결될 수 있는 기술적 기반이 마련됨
- 이는 기존에 예방·탐지·집행을 분리하여 운영하던 구조에서 벗어나, 배포 인프라 내에서 세 기능이 하나의 흐름으로 연계되는 실시간 보호 체계로의 이동을 보여주는 사례로 평가됨

시사점: 탐지 기술과 집행 체계의 연계 및 국제 공조 필요성

• 탐지 기술의 고도화와 실효성 확보 과제

- 스트리밍 콘텐츠의 특성상 불법복제로 인한 저작권자의 수익 기회가 소멸될 수 있으므로, 침해 대응 또한 사후 삭제 중심에서 사전 차단을 중심으로 대응 체계를 수립할 필요성이 제기됨
- 현재 차단 요청의 약 9%만 실제 집행되는 상황에서[3], 엣지 워터마킹과 같은 탐지 기술의 고도화가 실질적 보호 효과로 이어지기 위해서는 차단 명령의 집행과 중개자(intermediary)* 협조 체계 확보가 병행되어야 함
- 특히 불법 스트리밍 플랫폼은 단속에 적발된 이후에도 도메인 교체만으로 재가동이 가능한 구조로, 기술 대응과 집행 체계가 연결되지 않으면 반복적 피해가 발생할 수 있다는 점도 함께 지적됨
- 이러한 기술적 진보가 실질적인 보호 효과로 이어지기 위해서는 수사 및 차단 등 집행 체계와의 연계가 병행되어야 한다는 점도 함께 제기됨

* 중개자(intermediary): 콘텐츠 유통 과정에서 플랫폼·인터넷서비스제공자(ISP)·도메인 등록기관 등 권리자와 최종 이용자 사이에 위치해 콘텐츠 전송·접근을 매개하는 사업자를 통칭함

• 불법복제 대응을 위한 공조 필요성

- 미국 무역대표부(Office of the United States Trade Representative, USTR)*는 현행 통지-삭제(notice-and-takedown) 방식이 라이브 콘텐츠 보호에 한계가 있다고 밝히며, 기존 저작권법 체계가 불법복제 실시간 유통구조를 충분히 반영하지 못하고 있다고 지적함[4]
- 이는 불법복제 대응이 개별 국가나 단일 플랫폼 차원의 조치만으로는 한계가 있으며, 권리자·플랫폼·수사기관 간 실시간 대응 체계와 국제 공조를 전제로 설계되어야 함을 의미함
- 나아가 콘텐츠 보호는 개별 기술의 고도화만으로는 완성되지 않으며, 탐지·차단·집행이 결합된 다층적 보호 체계를 구축하고 이를 유기적으로 연결하는 접근이 병행되어야 함을 시사함

* 미국 무역대표부(Office of the United States Trade Representative, USTR): 미국의 무역 정책을 총괄하는 행정부 기관으로, 매년 저작권 침해·위조 관련 '악명 높은 시장(Notorious Markets)' 보고서를 발간해 글로벌 지식재산권 보호 현황을 점검함

참고문헌

- Brad Watts, "Synamedia Launches Edge Watermarking Piracy Solution", Content+Technology, 2026.03.26., https://content-technology.com/nabshow/synamedia-launches-edge-watermarking-piracy-solution/
- mburns, "Piracy in live sports: How broadcasters, leagues, platforms and federations are fighting back", SVG Europe, 2025.10.07., https://www.svgeurope.org/blog/headlines/piracy-in-live-sports-how-broadcasters-leagues-platforms-and-federations-are-fighting-back/
- Office of the United States Trade Representative(USTR), 2025 Review of Notorious Markets for Counterfeiting and Piracy, 2026.03.03., https://ustr.gov/sites/default/files/files/Press/Releases/2026/2025%20Notorious%20Markets%20List%20(final).pdf
- Maria Mascha Malinkowitsch, "Why Even Major Piracy Takedowns Show the Industry Must Put Greater Emphasis on Proactive Protection Efforts", TV Technology, 2026.04.07., https://www.tvtechnology.com/infrastructure/security/why-even-major-piracy-takedowns-show-the-industry-must-put-greater-emphasis-on-proactive-protection-efforts
- Verimatrix, "Why Piracy Is Outpacing Streaming Regulation", 2026.01.13., https://www.verimatrix.com/anti-piracy/anti-piracy-insights/piracy-is-evolving-faster-than-legislation-what-streaming-platforms-must-do-now/

3) Verimatrix, "Why Piracy Is Outpacing Streaming Regulation", 2026.01.13., https://www.verimatrix.com/anti-piracy/anti-piracy-insights/piracy-is-evolving-faster-than-legislation-what-streaming-platforms-must-do-now/
4) Office of the United States Trade Representative(USTR), 2025 Review of Notorious Markets for Counterfeiting and Piracy, 2026.03.03., https://ustr.gov/sites/default/files/files/Press/Releases/2026/2025%20Notorious%20Markets%20List%20(final).pdf

사운드온의 변형 음원 탐지 도구 도입과 음원 유통 플랫폼의 검증 체계 강화

사운드온의 음원 적법성 검증 체계와 변형 음원 탐지 도구 도입

- **사운드온, 음원 유통의 신뢰성 확보를 위해 변형된 음원을 탐지할 수 있는 도구 도입**

- 틱톡(TikTok)이 자사 음원 유통 플랫폼 사운드온(SoundOn)*에 변형된 음원을 탐지할 수 있는 도구를 도입함
- 이 도구는 기존 곡의 박자나 음정 등을 변형하여 별개의 곡처럼 등록하는 음원을 탐지하는 데 쓰이며, 기존에 사운드온에 음원 식별 기술을 제공해 온 ACR클라우드(ACRCloud)에서 개발함
- 또한 이 도구는 음원의 고유한 음향적 특징(주파수 패턴, 리듬 등)을 추출하여 사전에 구축된 데이터베이스와 비교함으로써 동일하거나 유사한 음원을 식별하는 기술인 오디오 핑거프린팅(audio fingerprinting) 기술을 기반으로 함
- 해당 도구는 무단으로 사용될 가능성이 있는 콘텐츠의 유통을 최소화하여, 아티스트와 음반사가 신뢰할 수 있는 음원 유통 환경을 구축하기 위해 도입됨

* 사운드온(SoundOn): 아티스트가 음원을 업로드하면 이를 검토·관리한 뒤 스포티파이(Spotify), 애플뮤직(Apple Music) 등 주요 디지털 음원 서비스에 전달하는 음원 유통 플랫폼

- **변형 음원 탐지 도구 외에도 음원 유통 적법성을 검증하는 다층적 체계를 운영 중**

- 사운드온은 ① 음원 자체의 적법성 검토, ② 기존 저작물을 무단으로 사용한 것으로 의심되는 음원에 대한 추가 심사, ③ 음원 등록자의 신원 확인을 결합한 다층적 검증 체계를 운영하고 있음
- 구체적으로, 음원을 검증할 때는 틱톡의 콘텐츠 스캔 인프라를 활용하여 해당 음원이 기존 저작물을 무단으로 사용한 것인지를 검토함
- 콘텐츠 스캔 과정에서 의심 콘텐츠로 분류된 음원은 추가 검토 대상이 되며, 필요한 경우 담당자가 직접 확인하는 절차를 거쳐 최종 유통 여부가 결정됨
- 음원 등록자의 신원을 확인하는 절차도 별도로 운영되며, 음원을 등록할 때 사진이 부착된 신분증으로 등록자 본인 여부를 확인함

[표1] 사운드온의 검증 요소별 역할 비교

검증 요소	검증 대상	작동 방식
변형 음원 탐지(ACRCloud)	등록 음원	오디오 핑거프린팅 기반 유사성 점검
틱톡 콘텐츠 스캔 인프라	등록 음원	기존 저작물 무단 사용 여부 자동 검토
등록자 신원 확인	등록 주체	신분증 사진 기반 신원 확인
담당자 직접 검토	의심 콘텐츠	자동 분류된 음원에 대한 사람의 최종 판단

출처: 참고문헌 종합하여 재구성

변형 음원 유통에 따른 저작권 침해 우려와 주요 음반사 간 갈등

- **변형 음원 유통이 원곡 권리자의 저작권 침해로 이어질 수 있다는 우려 확산**

- 사운드온이 음원 검증 체계를 강화한 것은, 원곡의 박자나 음정 등을 변형하여 만든 음원이 유통될 경우, 원곡 권리자의 저작권을 침해할 수 있다는 업계의 우려에 대응하기 위함임
- 이러한 우려는 틱톡과 주요 음반사 간의 계약에도 반영되었음
- 일례로, 2024년 5월 사운드온의 모기업인 틱톡과 세계 최대 음반사인 유니버셜뮤직그룹(Universal Music Group)은 유니버셜뮤직그룹 소속 아티스트의 음원을 틱톡 플랫폼에서 이용할 수 있도록 하는 계약을 체결함
- 해당 계약에는 AI로 제작되거나 변형된 음원으로부터 유니버셜뮤직그룹 소속 아티스트를 보호하기 위한 조치가 포함됨
- 변형 음원 문제는 계약 체결뿐만 아니라 협상 결렬의 원인이 되기도 함. 같은 해 10월, 틱톡은 글로벌 라이선싱 조직인 멀린(Merlin)과의 협상을 중단함
- 이 협상이 중단된 배경에는, 멀린 소속 일부 아티스트가 기존 곡의 템포·음정 등을 변형하여 다수의 음원을 업로드하고 수익을 창출했다는 틱톡 측의 우려가 작용한 것으로 알려짐[5)]
- 이러한 상황에서 유통 플랫폼은 단순히 음원을 전달·유통하는 데 그치지 않고, 음원의 적법성을 검증·관리하는 역량까지 요구받고 있음

참고문헌

- Mandy Dalugdug, “TikTok’s distro service SoundOn cracks down on manipulated audio via ACRCloud partnership to intercept unauthorized tracks”, MusicBusinessWorldwide, 2026.04.02., https://www.musicbusinessworldwide.com/tiktoks-distro-service-soundon-cracks-down-on-manipulated-audio-via-acrcloud-partnership-to-intercept-unauthorized-tracks/
- Soundwave, “How Does TikTok SoundOn Boost Your Music Carrer?”, 2026.04.08. 접속기준., https://www.soundon.global/forum/tiktok-soundon-music-distribution?lang=en

5) Mandy Dalugdug, “TikTok’s distro service SoundOn cracks down on manipulated audio via ACRCloud partnership to intercept unauthorized tracks”, MusicBusinessWorldwide, 2026.04.02., https://www.musicbusinessworldwide.com/tiktoks-distro-service-soundon-cracks-down-on-manipulated-audio-via-acrcloud-partnership-to-intercept-unauthorized-tracks/

AI 음악 생성 플랫폼 수노의 저작권 필터 취약성 논란

AI 커버 무단 업로드로 인한 피해 확산과 수노의 저작권 필터 논란

• AI 커버 무단 업로드로 인한 저작권 분쟁 확산

- 최근 원곡자의 동의 없이 AI가 생성한 커버곡이 스트리밍 플랫폼에 무단으로 업로드되며 원곡자의 수익이 줄어드는 피해 사례가 확산되고 있음
- 실험음악 작곡가 윌리엄 바신스키(William Basinski), 인디 록 그룹 킹 기자드 앤 더 리저드 위저드(King Gizzard and The Lizard Wizard)는 자신의 곡을 커버한 AI 생성 음원이 스트리밍 플랫폼에 무단으로 업로드되어 조회수가 분산되는 피해를 입음
- 포크 아티스트 머피 캠벨(Murphy Campbell)은 자신의 곡을 AI가 커버한 음원이 스포티파이(Spotify)에 무단 업로드된 후, 배포사로부터 역으로 저작권 침해 신고를 당해 저작권료 지급이 일시적으로 보류되는 피해를 입음
- 원곡자의 동의 없이 생성·업로드된 AI 커버는 스트리밍 플랫폼에서 원곡의 스트리밍 횟수를 분산시키며, 최소 1,000회 이상 스트리밍되어야 수익이 지급되는 스포티파이 같은 플랫폼에서는 원곡자의 수익이 감소하는 결과로 이어질 수 있음
- 이러한 피해는 AI 음악 생성 단계에서의 저작권 필터링 실패가 유통 단계의 분쟁으로 이어질 수 있음을 보여줌

• 수노의 저작권 필터를 우회하는 AI 커버 생성 논란

- 이러한 상황에서 2026년 4월 미국 기술 매체 더 버지(The Verge)가 AI 음악 생성 플랫폼 수노(Suno)*에서 작동하는 저작권 필터의 취약성을 보도하면서 논란이 확산됨[1)]
- 수노에서는 저작권이 있는 곡이나 가사 사용은 금지되어 있으며, 이용자가 저작권이 있는 곡이나 가사를 업로드하면 이를 인식해 차단하는 저작권 필터가 작동함
- 그러나 더 버지는 간단한 조작으로 저작권 필터를 우회할 수 있거나, 저작권 필터가 아예 작동하지 않는 사례를 확인하고, 이러한 취약성을 지적함
- 이로 인해 저작권이 있는 곡의 AI 커버를 수노에서 무단으로 생성할 수 있는 것으로 나타나, 이러한 취약성이 문제로 지적됨

* 수노(Suno): 이용자가 본인의 곡을 업로드해 리믹스하거나, 직접 작성한 가사로 음원을 생성할 수 있는 AI 음악 플랫폼

1) Terrence O'Brien, "Suno is a music copyright nightmare", The Verge, 2026.04.06., https://www.theverge.com/ai-artificial-intelligence/906896/suno-copyright-ai-music-covers

수노 저작권 필터의 취약점과 무단 수익화 경로

• 수노의 저작권 필터를 우회할 수 있는 기법

- 더 버지에 의하면, 수노의 저작권 필터를 우회하는 방법으로 음원 속도를 의도적으로 변경하거나 음원에 노이즈를 추가하는 방법이 확인됨
- 오다시티(Audacity)*와 같은 무료 소프트웨어로 음원 속도를 절반이나 두 배로 변경하면 수노의 저작권 필터가 해당 곡을 인식하지 못하고 통과시킴. 음원의 시작과 끝에 화이트 노이즈를 추가하면 우회 확률이 더욱 높아짐
- 이용자는 수노에 업로드한 후에는 수노 스튜디오에서 음원을 원래 속도로 복원하고 화이트 노이즈를 제거할 수 있음
- 가사의 경우, 공식 가사를 그대로 붙여넣으면 저작권 필터가 이를 차단하지만, 첫 번째 절과 코러스 일부의 철자를 변경하면 이후 가사는 원본 그대로 입력해도 수노의 저작권 필터를 통과함
- 위와 같은 방식으로 비욘세(Beyoncé)의 '프리덤(Freedom)', 블랙 사바스(Black Sabbath)의 '파라노이드(Paranoid)' 등 유명곡의 AI 커버를 수노에서 생성할 수 있는 것으로 나타남

* 오다시티(Audacity): 무료 오픈소스 오디오 편집 소프트웨어로, 속도 변경, 노이즈 추가·제거 등 기본적인 음원 조작 기능을 제공함

• 인디 아티스트 곡에 대한 저작권 필터의 보호 공백

- 유명 아티스트의 곡은 저작권 필터를 우회하기 위해 조작이 필요한 반면, 인디 아티스트의 곡은 아무런 조작 없이도 저작권 필터를 통과하는 것으로 확인됨
- 소규모 레이블 소속 아티스트나 밴드캠프(Bandcamp)*, 디스트로키드(DistroKid)** 등을 통해 자체 배포하는 아티스트의 곡이 특히 취약하며, 저작권 필터가 이들 곡을 보호 대상으로 인식하지 못하는 것으로 보임
- 싱어송라이터 맷 윌슨(Matt Wilson)·찰스 비셀(Charles Bissell), 실험 음악가 클레어 루세이(Claire Rousay) 등의 곡은 의도적인 조작 없이도 저작권 필터를 통과함

* 밴드캠프(Bandcamp): 인디 음악가들이 중간 유통사 없이 팬에게 직접 음원을 판매·배포할 수 있는 온라인 플랫폼

** 디스트로키드(DistroKid): 독립 음악가들이 스포티파이, 애플 뮤직 등 주요 스트리밍 플랫폼에 음원을 배포할 수 있도록 지원하는 디지털 음악 배포 서비스

• 수노에서 생성된 AI 커버, 스트리밍 플랫폼에 업로드 및 무단 수익화 가능

- 수노는 음원 업로드 시점에만 저작권 필터를 적용하며, 생성된 음원은 재검사하지 않으며, 외부 배포 전 추가 검사도 하지 않는 것으로 보임
- 따라서 수노의 저작권 필터를 우회해 생성한 AI 커버를 디스트로키드 등의 음원 배포 서비스를 통해 스트리밍 플랫폼에 업로드할 수 있음. 이를 통해 무단으로 수익화할 수 있는 경로가 열려 있음
- 허가받은 커버곡의 경우 원곡자에게 저작권료가 지급되지만, 무허가로 업로드된 AI 커버의 경우 원곡자에게 저작권료가 지급되지 않을 가능성이 있음

AI 커버에 대한 스트리밍 플랫폼의 대응 현황과 협력 체계 구축의 필요성

• 주요 스트리밍 플랫폼의 AI 커버 대응 현황과 한계

- 디저(Deezer), 코부즈(Qobuz), 스포티파이 등 주요 스트리밍 플랫폼은 AI 커버에 대응하는 조치를 시행 중임
- 스포티파이 대변인 크리스 마코스키(Chris Macowski)는 더 버지에 "아티스트 권리 보호를 중요하게 여기며 다각도로 접근하고 있다"고 밝힘[2]. 구체적으로 무단으로 업로드되는 콘텐츠를 방지하는 장치를 운영하며, 중복·유사 음원 식별 시스템을 활용하고, 이러한 기술적 대응을 사람이 직접 검토하는 과정으로 보완하고 있다고 설명함
- 다만 마코스키는 "이러한 무단 콘텐츠를 탐지하는 데 기술적 어려움이 있으며, 새로운 기술이 등장함에 따라 계속 투자하고 발전시켜야 할 영역"이라며 현재 대응의 한계를 인정함
- 수노의 저작권 필터가 쉽게 우회되는 현 상황에서, 스트리밍 플랫폼의 사후 대응만으로는 원곡자에게 미치는 피해를 충분히 방지하기 어렵다는 한계가 있음
- 더 버지는 수노 뿐 아니라 배포·스트리밍 플랫폼을 포함한 전체 시스템에 문제가 있다고 지적함. 이 문제를 해결하려면, AI 음악 생성 플랫폼과 배포·스트리밍 플랫폼 간 협력 체계 구축이 필요할 것으로 전망됨

참고문헌

• News Room, "Suno is a music copyright nightmare capable of pumping out AI cover slop", Daily Guardian, 2026.04.05., https://dailyguardian.ca/suno-is-a-music-copyright-nightmare-capable-of-pumping-out-ai-cover-slop/
• Terrence O'Brien, "Suno is a music copyright nightmare", The Verge, 2026.04.06., https://www.theverge.com/ai-artificial-intelligence/906896/sunos-copyright-ai-music-covers

2) News Room, "Suno is a music copyright nightmare capable of pumping out AI cover slop", Daily Guardian, 2026.04.05., https://dailyguardian.ca/suno-is-a-music-copyright-nightmare-capable-of-pumping-out-ai-cover-slop/

이르데토, 대규모 OTT 환경을 위한 포렌식 워터마킹 기술 공개

OTT 환경에서의 요구사항과 이르데토의 기술적 차별점

- **이르데토의 포렌식 워터마킹 기술, 트레이스마크**

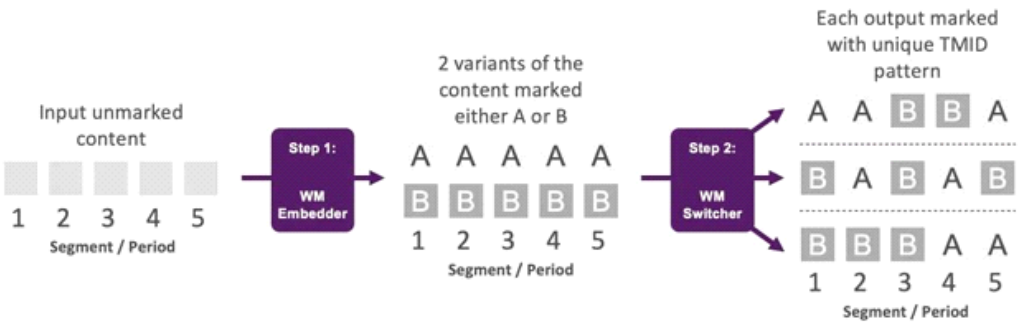
- 네덜란드의 디지털 콘텐츠 보안 및 불법복제 방지 솔루션 기업인 이르데토(Irdeto)는 대규모 OTT 스트리밍 환경을 위한 포렌식 워터마킹 기술인 트레이스마크(TraceMark™)를 공개함
- 포렌식 워터마킹은 영상에 비가시적 식별 정보를 삽입하여 유출 경로를 추적하는 기술로 재인코딩, 화면 캡처, 압축 등 불법복제 과정의 변형에도 워터마크가 유지되어야 함
- 실제 불법복제 환경에서는 영상 일부만 캡처되거나 여러 유출본을 결합하는 공격이 발생하므로, 워터마크가 부분적으로만 존재하거나 훼손된 상황에서도 신뢰성 있게 유출자를 식별할 수 있어야 함

- **균일 전환 식별자(U-SWIDs) 방식의 기술적 우위**

- 대규모 OTT 환경에서는 각 영상 구간을 A, B 두 버전으로 준비한 뒤 사용자별로 다른 조합을 제공하는 A/B 워터마킹 방식이 널리 사용되고 있으나, 일부 비트 손실 시 식별 정확도가 낮아지는 한계가 있음
- 이르데토는 이러한 한계를 극복하기 위해 각 워터마크 비트가 동등한 통계적 중요도를 갖도록 설계한 균일 전환 식별자(Uniform Switching Identities, 이하 U-SWIDs) 방식을 독자 개발하여, 일부 비트가 손실되더라도 나머지 비트만으로도 신뢰도가 균등하게 유지되도록 함
- U-SWIDs는 필요 시점에 즉시 생성 가능한 구조로 설계되어 CDN 및 엣지 환경에서도 실용적으로 배치할 수 있으며, 영상 일부만 확보된 상황이나 콜루전 공격(Collusion Attack)* 공격 환경에서도 안정적인 식별 성능을 제공하여 대규모 OTT 스트리밍 환경에서 확장성과 내구성을 동시에 확보함

* 콜루전 공격(Collusion Attack): 디지털 워터마킹 기술에서 여러 명의 사용자가 공모하여(Collusion), 같은 영상이나 이미지의 서로 다른 복사본을 비교 분석함으로써 워터마크를 제거하거나 변형시키는 공격 방식

[그림 1] A/B 워터마킹 방식 프로세스



출처: Lau Zuydervelt, "Beyond the lab: What modern forensic watermarking must solve at scale", Irdeto, 2026.04.01., https://irdeto.com/blog/modern-forensic-watermarking-at-scale

이르데토 트레이스마크의 통합 방어 체계와 산업 적용

• 모니터링·탐지·식별·집행을 연계하는 통합 워크플로우

- 이르데토 트레이스마크는 워터마크 삽입만으로 끝나지 않고, 유출 콘텐츠를 실시간으로 모니터링하고 워터마크를 추출하여 유출 경로를 식별한 뒤 법적 조치까지 연계하는 통합 체계를 구축함
- 모니터링 단계에서는 소셜 미디어, 불법 스트리밍 사이트 등을 자동으로 크롤링하여 유출 콘텐츠를 탐지하고, 탐지된 영상에서 워터마크를 추출하여 유출 경로를 역추적함
- 식별된 유출자 정보는 콘텐츠 권리자 및 법 집행 기관과 공유되어 법적 조치로 연결되며, 이를 통해 단순 기술 제공을 넘어 실질적인 불법복제 억제 효과를 창출함
- 이 시스템은 불완전하거나 훼손된 영상에서도 워터마크를 신뢰성 있게 복원할 수 있도록 설계되어, 실전 불법복제 환경에서 유출자 추적 가능성을 극대화함
- 단순 기술 제공을 넘어 모니터링·탐지·식별·집행을 하나의 워크플로우로 통합함으로써, 불법복제에 대한 실질적 억제 효과를 창출함

• 하모닉의 라이브 스포츠 스트리밍 솔루션 통합 사례

- 미국의 방송 인프라 솔루션 기업 하모닉(Harmonic)은 자사의 라이브 스포츠 스트리밍 솔루션에 이르데토 트레이스마크 워터마킹 기술을 통합하여, 프리미엄 스포츠 콘텐츠의 불법 스트리밍을 차단하고 유출자를 신속하게 식별할 수 있는 체계를 구축함
- 하모닉은 송출 시점에 삽입되는 A/B 워터마크과, CDN이나 파트너사 등 배포 경로별 워터마킹을 모두 지원함. 또한 지연시간 최소화를 강조하여 실시간 스포츠 중계에 최적화된 솔루션을 제공하고 있으며, 이는 이르데토의 기술이 업계 주요 솔루션으로 자리잡았음을 보여줌

시사점: 상용화 가능한 워터마킹 기술의 산업 확산 전망

• 프리미엄 콘텐츠 보호를 위한 기술·운영 통합의 필요성

- 이르데토 트레이스마크 사례는 워터마킹 기술의 실효성이 기술적 완성도뿐 아니라 실제 운영 환경 적합성과 법적 집행 연계 여부에 의해 결정됨을 보여주며, 불법복제 수법이 고도화되면서 모니터링, 식별, 집행까지 통합한 체계 구축이 필수 조건으로 부상함
- 하모닉 등 주요 스트리밍 인프라 기업들이 이르데토 기술을 도입하는 사례가 확산되면서, 상용화 가능한 워터마킹 기술이 OTT 산업 표준으로 자리잡을 가능성이 커지고 있음
- 향후 실시간 스포츠 중계, 영화 동시 개봉 등 고가치 콘텐츠의 온라인 유통이 확대될수록, 실전 환경에서 검증된 워터마킹 기술의 도입이 OTT 플랫폼 경쟁력 확보의 핵심 요소로 자리잡을 전망임

참고문헌

- Lau Zuydervelt, "Beyond the lab: What modern forensic watermarking must solve at scale", Irdeto, 2026.04.01., https://irdeto.com/blog/modern-forensic-watermarking-at-scale
- PR Newswire, “Harmonic Delivers Winning Live Sports Streaming Innovations”, 2026.03.26., https://www.prnewswire.com/news-releases/harmonic-delivers-winning-live-sports-streaming-innovations-302725609.html

주간 기술 동향

딥페이크 사전 방어를 위한 출처 추적 워터마킹 SAiW 기술

• AI 생성 콘텐츠 범람 시대, 디지털 정체성 보호를 위한 워터마킹 기술의 부상

생성형 AI 기술의 급속한 발전은 디지털 콘텐츠 창작의 새로운 가능성을 열었지만, 동시에 개인의 디지털 정체성을 무단으로 복제하고 악용하는 딥페이크 문제를 심화시키고 있다. 특히 AI가 생성한 이미지, 영상, 음성이 실제와 구분하기 어려울 정도로 정교해지면서, 유명인은 물론 일반인까지 자신의 얼굴과 목소리가 동의 없이 사용되는 피해를 겪고 있다. 이러한 상황에서 개인이 자신의 얼굴과 목소리가 무단으로 사용되는 것을 막을 수 있어야 한다는 인식이 확산되고 있으며, 이는 기술 발전뿐 아니라 법적 보호가 필요한 문제로 인식되고 있다.

이에 대응하기 위해 주요 플랫폼들은 AI 생성 콘텐츠를 탐지하고 관리하는 자체 시스템을 구축하고 있다. 유튜브는 개인의 얼굴, 목소리 등을 무단으로 사용한 AI 생성 콘텐츠를 식별할 수 있는 유사성 관리 도구를 도입하여, 사용자가 자신의 디지털 복제물을 발견하고 삭제를 요청할 수 있는 경로를 마련했다. 또한 저작권 보호에 사용되던 자동화 기술을 딥페이크 탐지에도 적용하여, 합성 콘텐츠가 광범위하게 확산되기 전에 선제적으로 식별하려는 시도를 진행 중이다. 그러나 이러한 플랫폼 주도 방식만으로는 급증하는 딥페이크 콘텐츠에 충분히 대응하기 어렵다는 한계가 지적되고 있다.

플랫폼 중심의 사후 탐지 방식은 근본적인 구조적 한계를 지니고 있다. 개인이 수동으로 딥페이크를 찾아 신고해야 하는 부담은 AI 생성 콘텐츠의 폭발적 증가 속도를 따라잡기 어렵고, 한 플랫폼에서 보호받더라도 다른 플랫폼에서는 여전히 무방비 상태로 남는 문제가 존재한다. 더욱이 악의적 행위자들은 지속적으로 탐지 시스템을 우회하는 새로운 기법을 개발하기 때문에, 방어 기술은 항상 한발 뒤처질 수밖에 없는 취약점을 안고 있다. 이는 사후 대응이 아닌 사전 방어 중심의 근본적인 기술적 전환이 필요함을 시사한다.

이러한 배경에서 콘텐츠 생성 단계부터 출처 정보를 삽입하는 사전 방어 기술이 주목받고 있다. 특히 출처 조건부 비가시 워터마킹 기술인 SAiW(Source-Aware Invisible Watermarking)는 AI가 생성한 이미지에 육안으로 식별할 수 없는 워터마크를 삽입하여, 해당 콘텐츠가 어떤 생성 모델에서 만들어졌는지 추적할 수 있게 한다. 이 기술은 단순히 딥페이크 여부를 판별하는 것을 넘어 정확한 출처를 식별함으로써, 책임 소재를 명확히 하고 법적 대응의 투명한 근거를 제공할 수 있는 가능성을 제시하고 있다.

[사례] 출처 조건부 비가시 워터마킹 기술 SAiW

• 현재 딥페이크 방어 기술의 특징과 한계

- 기존 워터마킹 기술은 AI가 만든 이미지에 눈에 보이지 않는 고유 표시를 넣어 출처를 추적하지만, 대부분 한 가지 AI 도구에만 적용할 수 있기 때문에 여러 생성 프로그램이 동시에 사용되는 실제 산업 환경에서는 제대로 작동하기 어려움

- 여러 AI 생성 도구를 동시에 추적하려면 각각에 맞는 별도 시스템을 만들어야 하는데, 이는 기술적으로 매우 복잡하고 비용도 많이 들어 실제로 사용하기 어렵게 만듦

- 또한, 워터마크를 넣었더라도 이미지를 압축하거나, 밝기를 조정하거나, 노이즈를 추가하는 등 일상적인 편집 과정에서 쉽게 지워지거나 손상되어 법적 증거로 활용하기 어려운 문제가 있음

- 가장 근본적인 한계는 단순히 'AI 생성 이미지인가'만 판별할 뿐 구체적으로 어떤 프로그램이 언제 어떤 목적으로 사용되었는지 알 수 없어, 딥페이크 제작자의 책임 소재를 명확히 밝히기 어렵다는 점임

• 출처 조건부 워터마크 임베딩 구조

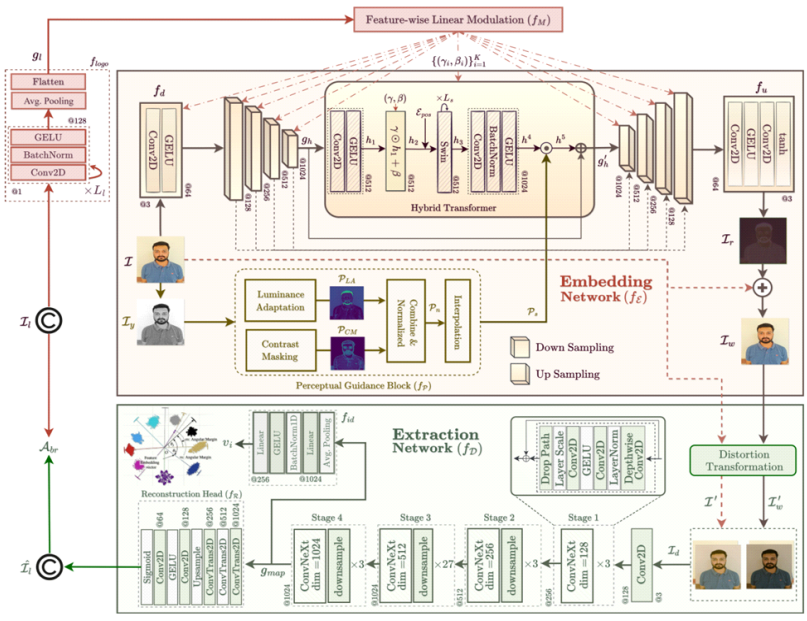
- 출처 조건부 비가시 워터마킹(Source-Attributable Invisible Watermarking, 이하 SAiW)는 기존 기술이 'AI 생성 여부'만 표시하던 것과 달리, 어떤 AI 도구로, 언제 만들었는지 구체적인 정보까지 추적 가능한 워터마크를 삽입하는 구조로, 각 AI 생성 도구마다 고유한 식별 코드를 이미지에 기록함

- 워터마크 인코더는 원본 이미지와 출처 정보를 동시에 입력받아, 출처별로 구별 가능한 패턴을 이미지 픽셀값에 미세하게 변조하여 삽입하는 방식으로 작동함. 이 과정에서 인간의 시각 시스템이 감지하기 어려운 고주파 영역에 워터마크를 배치하여, 이미지 품질의 저하 없이 정보를 은닉할 수 있도록 설계됨

- SAiW는 단일 디코더 네트워크만으로 여러 생성 모델의 출처를 동시에 식별할 수 있어, 기존 방식 대비 시스템 복잡도를 크게 낮추면서도 확장성을 확보함

- 출처 정보는 비트 문자열 형태로 인코딩되어 워터마크에 삽입되며, 디코더는 이를 해독하여 생성 모델의 종류, 버전, 생성 시점 등의 메타데이터를 복원할 수 있음

[그림 1] SAiW 작동 프로세스



출처: Bibek Das 외 4인, "SAiW: Source-Attributable Invisible Watermarking for Proactive Deepfake Defense", arXiv, 2026.03.24., https://arxiv.org/pdf/2603.23178

• **지각적 최적화 및 비가시성 확보**

- SAiW는 인간 시각 시스템의 특성을 모델링하여, 사람이 변화를 인지하기 어려운 이미지 영역을 골라 우선적으로 워터마크를 삽입함
- 워터마크가 삽입된 이미지는 원본과 비교했을 때 매우 높은 품질 지표를 기록하여, 전문가도 육안으로 구별하는 것이 사실상 불가능한 수준의 이미지 품질을 유지함
- 딥러닝 기반 손실 함수를 통해 워터마크 강도와 이미지 품질 간의 최적 균형점을 자동으로 학습하여, 수동 조정 없이도 안정적인 성능을 보장함
- 이러한 비가시성은 워터마크가 콘텐츠의 미적 가치나 상업적 활용도를 저해하지 않으면서도, 필요 시 법적 증거로 활용될 수 있는 기술적 기반을 제공함

• **다중 공격에 대한 강건성**

- SAiW는 강건성 확보를 위해 워터마크 신호를 이미지 전체에 분산시키는 방식을 채택하여, 일부 영역이 손상되더라도 전체 정보를 복원할 수 있는 구조를 갖춤
- 밝기 조정, 대비 변경, JPEG 압축, 가우시안 노이즈*, 가우시안 블러** 등 5가지 주요 이미지 변형 공격에 대해 평균 98% 이상의 워터마크 추출 정확도를 유지함
- 특히 JPEG 압축 품질 75%까지 워터마크가 손실 없이 보존되는데, 이는 인스타그램, 페이스북, 트위터 등 소셜 미디어 플랫폼에서 자동으로 적용되는 압축 수준을 충분히 견딜 수 있는 수준임
- 또한 인스타그램 필터 3종(Aden, Brooklyn, Clarendon)을 단독 또는 조합하여 적용한 경우에도 99% 이상의 추출 정확도를 보여, 실제 사용자들이 자주 사용하는 이미지 보정에도 안정적으로 유지됨
- 이미지가 여러 플랫폼을 거치며 반복적으로 재압축되거나 편집되는 실제 유통 환경에서도 워터마크가 유지되어, 출처 추적의 실효성을 보장함

* 가우시안 노이즈(Gaussian Noise): 이미지에 무작위로 밝기 변화를 주는 잡음으로, 디지털 카메라 촬영이나 이미지 압축 과정에서 자연스럽게 발생하는 화질 저하 현상

** 가우시안 블러(Gaussian Blur): 이미지를 흐릿하게 만드는 효과로, 배경을 부드럽게 처리하거나 모자이크 효과를 줄 때 사용되는 기법

• **이중 검증 메커니즘 및 성능 비교**

- SAiW는 워터마크 추출과 동시에 이미지가 AI로 생성되었는지 여부를 판별하는 이진 분류 기능을 통합하여, 두 가지 독립적인 검증 경로를 제공함
- 워터마크 디코더는 출처 정보를 추출하는 동시에 신뢰도 점수를 산출하여, 워터마크가 의도적으로 제거되거나 손상된 경우를 감지할 수 있음
- 이중 검증 구조는 한 가지 방법이 우회되더라도 다른 방법으로 AI 생성 콘텐츠를 식별할 수 있어, 악의적 공격에 대한 방어력을 높임
- SAiW는 기존 워터마킹 기술과의 비교 실험에서 비가시성과 강건성 모두에서 기존 기술 대비 우수한 성능을 기록하며, 특히 다중 출처 추적이 가능한 기술로 평가됨

결론 및 시사점

• 기술적 의의 및 산업적 활용 가능성

- SAiW는 콘텐츠 생성 단계부터 출처 정보를 삽입하는 접근법을 통해, 딥페이크가 확산된 후 추적하는 기존 방식의 한계를 극복하고 AI 생성 콘텐츠에 대한 책임 소재를 명확히 할 수 있게 함
- 다중 출처 추적 능력과 높은 강건성은 플랫폼 간 협력을 통한 통합 워터마킹 생태계 구축의 가능성을 보여주며, 유튜브, 인스타그램, 페이스북 등 주요 플랫폼이 공통 표준을 채택할 경우 개인의 디지털 정체성 보호를 위한 실질적인 방어막이 될 수 있음

• 현재 한계와 향후 발전 과제

- 워터마크 제거를 목적으로 정교하게 설계된 적대적 공격에 대한 방어 능력은 여전히 제한적이며, 실시간 처리 속도 개선과 동영상·오디오 등 멀티모달 콘텐츠로의 확장이 기술적 과제로 남아 있음
- 현재는 이미지에만 적용 가능하기 때문에 딥페이크 영상이나 음성 복제에 대응하기 위한 멀티모달 워터마킹 체계 구축이 필요함
- 기술적 발전과 함께 AI 생성 콘텐츠에 대한 워터마킹 의무화, 출처 정보 공개 기준, 법적 증거로서의 워터마크 인정 범위 등 정책적 프레임워크가 마련되어야 기술이 실효성 있는 방어 수단으로 자리잡을 수 있음

참고문헌

• Bibek Das 외 4인, "SAiW: Source-Attributable Invisible Watermarking for Proactive Deepfake Defense", arXiv, 2026.03.24., https://arxiv.org/pdf/2603.23178