문화체육관광부 Ministry of Culture, Sports and Tourism
한국저작권위원회 KOREA COPYRIGHT COMMISSION
한국저작권보호원 KOREA COPYRIGHT PROTECTION AGENCY

ICOTEC 2024

2024 국제저작권기술 콘퍼런스 International Copyright Technology Conference 2024 발표자료집 I PROGRAM BOOK

문화체육관광부 Ministry of Culture, Sports and Tourism
한국저작권위원회 KOREA COPYRIGHT COMMISSION
한국저작권보호원 KOREA COPYRIGHT PROTECTION AGENCY

ICOTEC 2024

저작권 보호 기술, 초연결 디지털 전환 시대의 창의성 수호

COPYRIGHT PROTECTION TECHNOLOGY
: Safeguarding Creativity in the
Hyperconnected Era of Digital Transformation

발표자료집
Program Book

2024 국제저작권기술 콘퍼런스

International Copyright Technology Conference 2024

2024. 11. 6. (수) 10:00~17:00 I JW메리어트 동대문 스퀘어 서울

문화체육관광부 Ministry of Culture, Sports and Tourism
한국저작권위원회 KOREA COPYRIGHT COMMISSION
한국저작권보호원 KOREA COPYRIGHT PROTECTION AGENCY

ICOTEC 2024

저작권 보호 기술,
초연결 디지털 전환 시대의
창의성 수호

COPYRIGHT PROTECTION TECHNOLOGY
: Safeguarding Creativity in the
Hyperconnected Era of Digital Transformation

발표자료집
Program Book

2024 International Copyright Technology Conference 2024
국제저작권기술 콘퍼런스

2024. 11. 6. (수) 10:00~17:00 I JW메리어트 동대문 스퀘어 서울

목차

개요

행 사 명 | 2024 국제저작권기술 콘퍼런스

일　　시 | 2024년 11월 6일 (수) 10:00 ~ 17:00

장　　소 | JW 메리어트 동대문 스퀘어 서울 (YouTube 생중계)

주　　제 | 저작권 보호 기술, 초연결 디지털 전환 시대의 창의성 수호

주　　최 | 문화체육관광부 Ministry of Culture, Sports and Tourism

주　　관 | 한국저작권위원회 KOREA COPYRIGHT COMMISSION　한국저작권보호원 KOREA COPYRIGHT PROTECTION AGENCY

프로그램

시간	프로그램		
09:30 ~ 10:00	행사 등록		
10:00 ~ 10:05	ICOTEC 2024 오프닝 영상 시청		
10:05 ~ 10:20	개회식	축 사 | **문화체육관광부 저작권국장**	
		개회사 | **한국저작권위원회 위원장**	
		환영사 | **한국저작권보호원 원장**	
10:20 ~ 10:40	2024 저작권기술 어워드 시상식		
10:40 ~ 10:50	브레이크 타임		
10:50 ~ 11:30	기조연설	사이버보안 관점에서 바라본 초연결 시대의 저작권 기술 **원유재** 교수 | 충남대학교	
11:30 ~ 12:00	초청연설	디지털 시대의 EU 저작권 인프라 강화: EUIPO의 새로운 저작권 계획 **해리 테밍크 서비스 총괄** | 유럽연합 지식재산청(EUIPO)	
12:00 ~ 13:10	오찬 및 커피브레이크		
Session 1 디지털 혁신 속 저작권 보호 기술			
13:10 ~ 13:40	I	물리적 복제 불가능 기술과 저작권 보호 기술 **박욱** 교수 | 경희대학교	튜토리얼 최신 인공지능기반 기술을 활용한 저작권 침해와 보호사례 및 AI 보안과 보호 이슈 (13:30~14:00) **우사이먼 성일** 교수 | 성균관대학교
13:40 ~ 14:00	II	생성형 AI 시대의 콘텐츠 진위성 **일케 데미르** 선임연구원 | 인텔	
14:00 ~ 14:30	III	30년간 비가시성 워터마크를 연구한 기업이 말해주는 "다양한 비가시성 워터마크 활용" 및 "생성형 AI 콘텐츠 위한 고속 비가시성 워터마킹 기술" 이야기 **최고** 대표 | 마크애니	
14:30 ~ 14:50	IV	저작물 방송 송출을 위한 콘텐츠 보안 적용 **로날드 휠러** 전무이사 | A3SA	
14:50 ~ 15:10	커피브레이크		
Session 2 콘텐츠 창작의 토대, 저작권 보호 기술			
15:10 ~ 15:40	I	웹툰 불법유출에 대한 기술적 대응의 중요성 **서충현** 실장 | 네이버웹툰	
15:40 ~ 16:00	II	콘텐츠 보호: 트렌드와 과제 **에릭 딜** 보안 및 미디어 기술 부사장 | 소니 픽처스 엔터테인먼트	
16:00 ~ 16:30	III	OTT 콘텐츠 불법 유출 현황과 이에 대응하는 콘텐츠 보안 기술 소개 **김준호** 프로덕트 매니저 | 잉카엔트웍스	
16:30 ~ 16:50	IV	콘텐츠 분석 및 워터마킹을 통한 이미지 복제 감지 **마테이스 두즈** 연구원 | 메타	
16:50~17:00	폐회 선언		

연사 소개

기조연설

사이버보안 관점에서 바라본 초연결 시대의 저작권 기술

원유재

충남대학교 교수

초청연설

디지털 시대의 EU 저작권 인프라 강화: EUIPO의 새로운 저작권 계획

해리 테밍크

유럽연합 지식재산청(EUIPO) 서비스 총괄

세션 1 : 디지털 혁신 속 저작권 보호 기술

1-1. 물리적 복제 불가능 기술과 저작권 보호 기술

박욱

경희대학교 교수

1-2. 생성형 AI 시대의 콘텐츠 진위성

일케 데미르

인텔 선임 연구원

1-3. 30년간 비가시성 워터마크를 연구한 기업이 말해주는 "다양한 비가시성 워터마크 활용" 및 "생성형 AI 콘텐츠 위한 고속 비가시성 워터마킹 기술" 이야기

최고

마크애니 대표

1-4. 저작물 방송 송출을 위한 콘텐츠 보안 적용

로날드 휠러

A3SA 전무이사

튜토리얼 세션

최신 인공지능기반 기술을 활용한 저작권 침해와 보호사례 및 AI 보안과 보호 이슈

우사이먼 성일

성균관대학교 교수

세션 2 : 콘텐츠 창작의 토대, 저작권 보호 기술

2-1. 웹툰 불법유출에 대한 기술적 대응의 중요성

서충현

네이버웹툰 실장

2-2. 콘텐츠 보호: 트렌드와 과제

에릭 딜

소니 픽처스 엔터테인먼트 보안 및 미디어 기술 부사장

2-3. OTT 콘텐츠 불법 유출 현황과 이에 대응하는 콘텐츠 보안 기술 소개

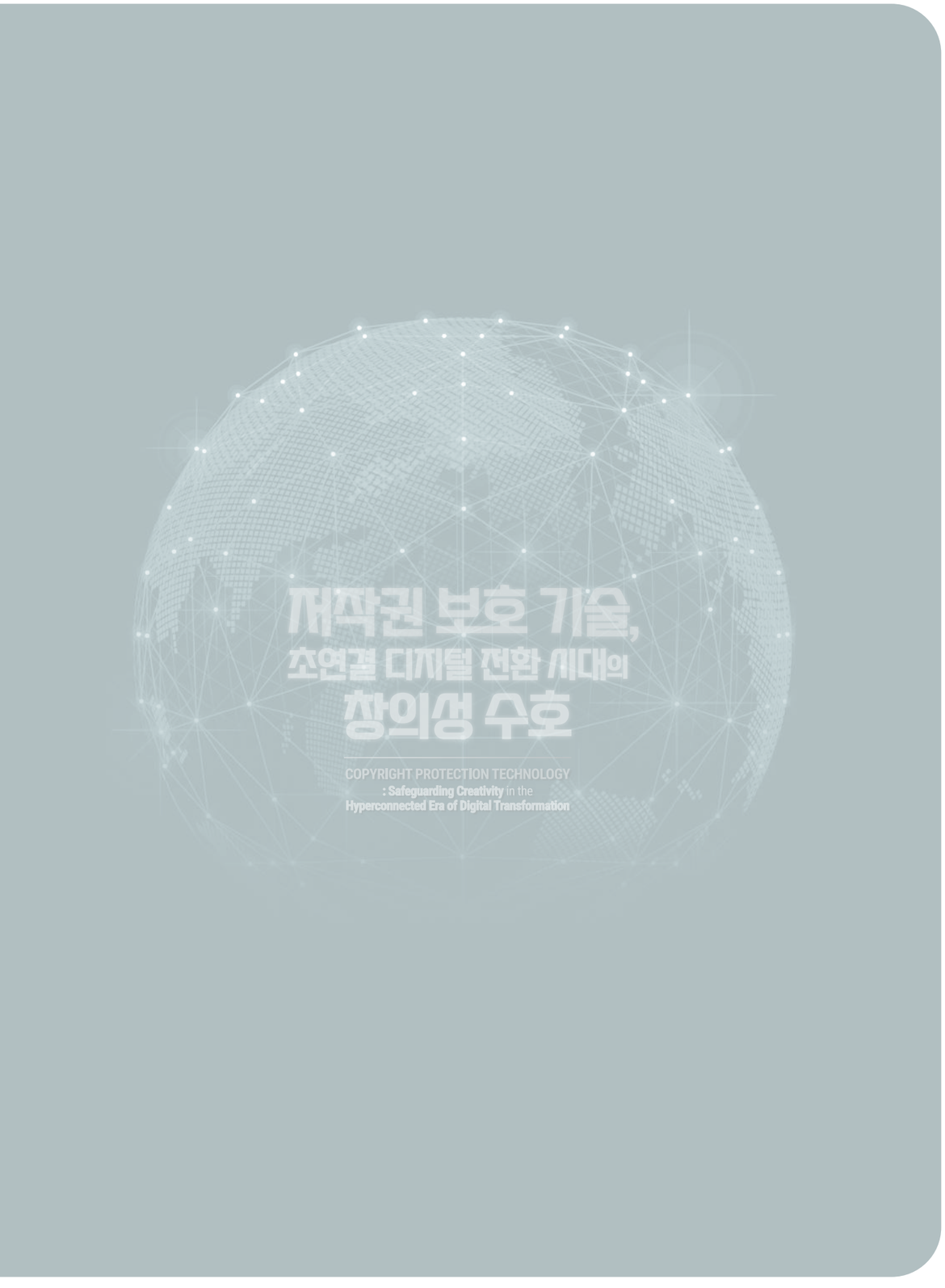
김준호

잉카엔트웍스 프로덕트 매니저

2-4. 콘텐츠 분석 및 워터마킹을 통한 이미지 복제 감지

마테이스 두즈

메타 연구원



기조연설

사이버보안 관점에서 바라본 초연결 시대의 저작권 기술

원유재 | 충남대학교 교수

기조연설

사이버보안 관점에서 바라본 초연결 시대의 저작권 기술

원유재
충남대학교 교수

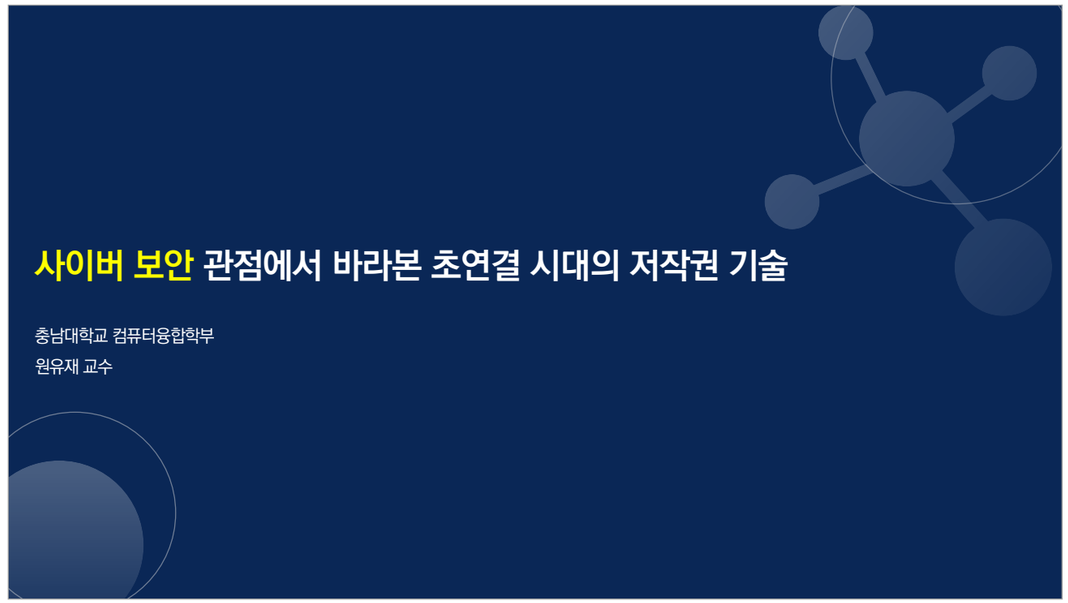
연사 이력

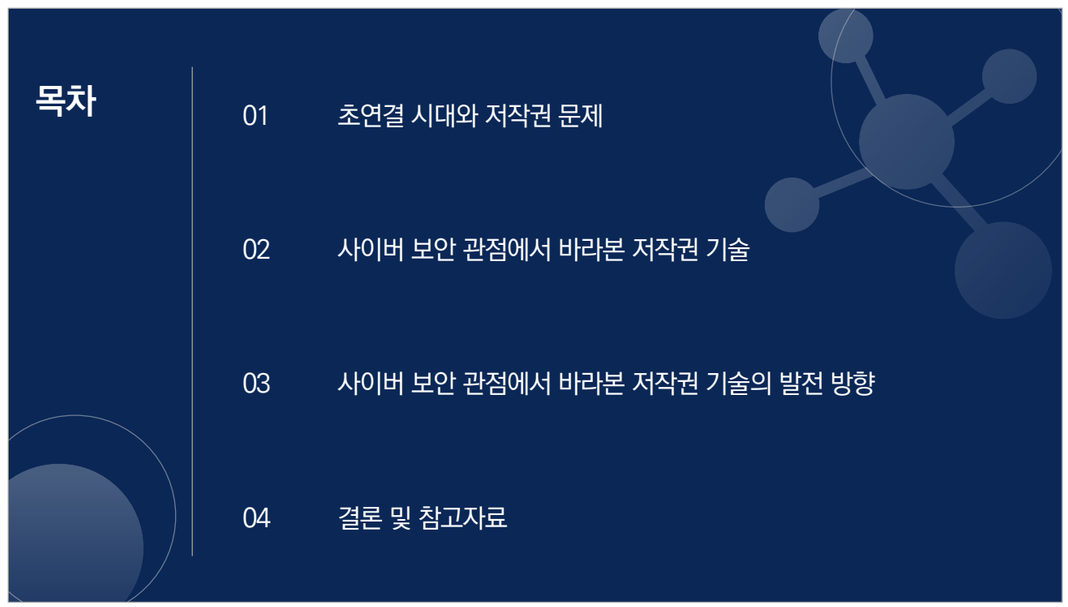
- 한국전자통신연구원(ETRI) 팀장 (1987~2001)
- 안랩유비웨어, 안철수연구소 CTO (2001~2004)
- 한국인터넷진흥원(KISA) 인터넷침해대응센터 본부장 (2004~2014)
- 충남대학교 컴퓨터융합학부 교수 (2014~현재)
- 충남대학교 융합보안연구센터 센터장 (2020~현재)

발표 내용

초연결 시대의 진전으로 시공간의 제약을 넘어 정보의 빠른 유통이 가능해진 시대이지만, 이로 인한 저작권 문제 또한 심화하고 있습니다. 사이버보안 관점에서 워터마크, DRM의 한계 점, 보안 취약성 및 생성형 AI의 저작권 침해 사례를 알아보고 최신 사이버 보안 기술을 활용한 대응에 대해 탐구합니다. 특히 AI/ML, 제로트러스트, 블록체인, 클라우드, 암호화 기술 동향을 기반으로 저작권 보호를 강화하기 위한 실질적인 방안을 제시하고 향후 발전 방향에 대해 알아봅니다.

The progression of the hyperconnected era has transcended spatial and temporal barriers, facilitating rapid dissemination of information, yet simultaneously exacerbating copyright issues. This presentation explores the limitations and security vulnerabilities of watermarking and DRM, as well as copyright infringement cases involving generative AI, from a cybersecurity perspective. It delves into the use of cutting- edge cybersecurity technologies as countermeasures. Specifically, it discusses trends in AI/ML, zero trust, blockchain, cloud, and encr yption technologies, proposing practical measures to enhance copyright protection and exploring future directions for development.





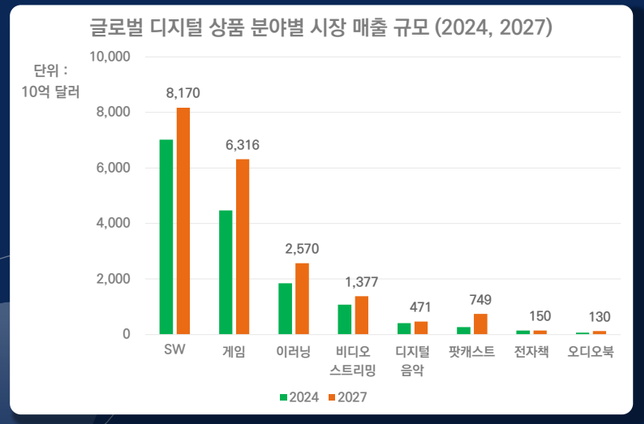
01 초연결 시대와 저작권 문제

초연결 시대란?

사람과 사물, 공간 등 모든 것(Things)들이 인터넷을 통해 서로 연결되어 모든 정보가 생성 및 수집, 공유 및 활용되는 시대

01 초연결 시대와 저작권 문제

콘텐츠 생산/소비 시장의 변화



Statista

글로벌 디지털 상품 시장 분야별 매출
연평균 1~30% 상승 (~2027)

[분야별 연평균 성장률]

- SW 5%
- 게임 10%
- 이러닝 8%
- 비디오 스트리밍 8%
- 디지털 음악 7%
- 팟캐스트 27%
- 오디오북 10%
- 전자책 2%

01 초연결 시대와 저작권 문제

콘텐츠 생산/소비 트렌드 변화

세계 경제 포럼(WEF)

모바일 기기 사용 증가

콘텐츠 생성 / 배포 단순화로 인한 급증

SNS 사용 증가

멀티 채널 / 파편화

마스터카드

디지털 상품 거래량 2년 간 70% 증가

맥킨지

[디지털 상품 지출 증가]

2021년 – 2024년

$ 1,100억 → $ 1,350억

21년 → 24년(예상)

01 초연결 시대와 저작권 문제

콘텐츠 생산/소비 트렌드 변화의 배경

트렌드 변화 배경

인터넷 사용률 증가

전 세계 인구 66%(53억 5천만) 인구 인터넷 사용

소셜 미디어 사용률 증가

활성 소셜 미디어 ID 50억 돌파
최신 사용자 수 인구 대비 62.3%(50억)

모바일 기기 보급 증가

전 세계 인구 69%(56억 1천만) 인구 모바일 기기 사용

네트워크 성능 향상

2018년 대비 브로드 밴드, 모바일 속도 3배 이상 증가

01 초연결 시대와 저작권 문제

트렌드 변화에 따른 저작권 문제

인터넷 사용률 증가
모바일 기기 보급 증가
→ 지리적 한계 없이 콘텐츠 불법 복제, 유통
모바일 기반 불법 사이트 접근성 상승

소셜 미디어 사용률 증가
네트워크 성능 향상
→ 개인 간 불법 콘텐츠 공유 용이
불법 콘텐츠 빠른 접근 가능

01 초연결 시대와 저작권 문제

트렌드 변화에 따른 저작권 문제

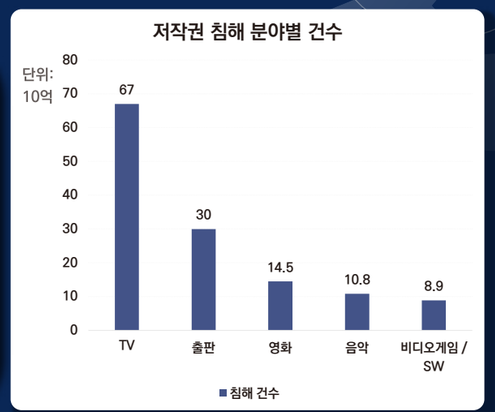
아카마이(Akamai)

22년 9월 저작권 침해 현황 보고서 발표

코로나 이후, 수많은 스트리밍 채널을 이용한
콘텐츠 생산/소비 증가

21년 1~9월 간 브라우저/모바일 App 이용한
불법 콘텐츠 사이트 접속 수는 총 1,320억 건

아카마이 연구원은 저작권 침해에
대처할 특효약은 없고 관련 산업에 악영향이 크다고 평가



01 초연결 시대와 저작권 문제

생성형 AI 기반 콘텐츠 저작권 문제

생성형 AI는 대규모 학습을 통해 텍스트, 이미지, 음악, 영상 등을 빠르게 생성하여 저작권 분쟁 다수 발생
→ 23년 미국에서만 관련 소송 13건 발생

23년 1월,
예술가 집단의
미드저니 소송
유사 이미지 생성

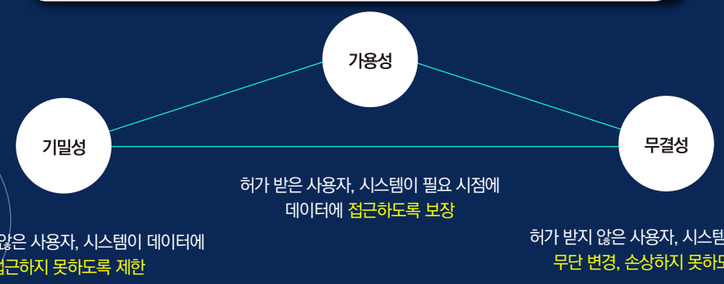
24년 10월,
뉴욕타임즈의
퍼플렉시티 소송
뉴스기사 무단 도용

24년 10월,
알콘 엔터테인먼트
의 일론머스크 소송
블레이드 러너 2049
스틸컷 무단 사용

02 사이버 보안 관점에서 바라본 저작권 기술

사이버 보안의 개요

사이버 보안이란?
컴퓨터, 전자 통신 시스템 및 서비스, 데이터, SW와, 관련 인프라의 손상 방지, 보호 및 복구를 위해 가용성/무결성/기밀성을 보장하는 것



02 사이버 보안 관점에서 바라본 저작권 기술

사이버 보안과 저작권 기술의 연관성

워터마크 / 포렌식마크

콘텐츠에 사람이 인지할 수 없는 저작권 정보를 삽입하고 검출기 통해 삽입 정보 식별하는 기술

저작권자 정보를 삽입하여 지적 재산권 분쟁의 정당성 증명

포렌식마크는 워터마크 기술의 확장으로 유포자, 배포 경로 추적하는 기술

DRM / CAS

허가된 사용자만 이용 가능하도록 라이선스 발급하여 콘텐츠 사용 권한 지속 통제하는 기술

암호화 기술 기반 불법 접근 및 복제 방지

CAS는 DRM과 연동하여 콘텐츠를 인증된 사용자에게 전송

인증 기반 저작권 콘텐츠 접근 제어, 복제/수정 차단 등 기밀성/무결성/가용성 확보 기술

02 사이버 보안 관점에서 바라본 저작권 기술

사이버 보안과 저작권 기술의 연관성

사이버보안과 저작권을 위한 기술은 공동의 목표가 있으며
이에 대응하기 위한 기술적 연관성 높음

	디지털 자산 보호	위협 방어, 대응	지속적 발전
사이버보안	방화벽, 제로 트러스트 등을 통한 접근 제어	IDS 등 공격 탐지 및 감사 네트워크 패킷, 데이터 암호화	AI 기반 탐지 및 방어 기반시스템 개선
저작권 보호 기술	DRM 기술을 통한 접근 제어, 복제 방지	포렌식 워터마크를 이용한 불법 유통 추적 DRM 기술을 통한 암호화	AI 워터마크, 멀티 DRM 등 기술 개선

02 사이버 보안 관점에서 바라본 저작권 기술

기존 저작권 기술의 한계점 (워터마크)

1990~
디지털 워터마크 기술 발전

2000~
인공신경망, 딥러닝 기반 워터마크 기술 등장

구글 논문 발표 (2017)
"AI 기반 워터마크 삭제 알고리즘 학습 및 비가시성 워터마크 제거 취약점 공개"

2020~
AI 생성 콘텐츠 식별 기능 필요성 제기

메릴랜드대 논문 발표 (2023)
"신뢰할 수 있는 워터마킹 없음 주장, 워터마크 세척(Washing out) 취약성 공개"

강인한(Robust) 워터마크 필요성 지속 제기, 제거, 변조 방지 필요 → 무결성 보장

02 사이버 보안 관점에서 바라본 저작권 기술

기존 저작권 기술의 한계점 (DRM)

2000~

2007.1 Windows Vista HD-DVD 인증 우회 및 불법 콘텐츠 재생 방법 공개

2007.7 MS Zune DRM 해킹

2009 아마존 킨들 DRM 무력화

2010 구글 안드로이드 마켓 LVL 해킹

2023 알라딘 DRM 키 탈취 및 전자책 무단 공개

2024 MS SW 기반 PlayReady XOR 키 공격에 무력화

DRM 기술의 지속적 해킹 공격 노출, 보안 기술 개선 통한 위협 방지 필요
→ 컨텐츠 기밀성, 무결성 보장

03 사이버 보안 관점에서 바라본 저작권 기술의 발전 방향

주요 발전 방향

디지털 워터마크 제거, 변조 대응	DRM 해킹, 정보 유출 방지	생성형 AI 콘텐츠 대응
비가시성 워터마크의 견고성(Robustness) 향상을 위한 기술 적용 필요	DRM 해킹 및 내부 사용자에 의한 민감 정보 유출 방지를 위한 기술 적용 필요	생성형 AI에 의한 허위 정보 확인 및 콘텐츠 저작권자 권리 보호를 위한 기술 적용 필요

03 사이버 보안 관점에서 바라본 저작권 기술의 발전 방향

워터마크 기술

디지털 워터마크 제거, 변조 대응

- AI/ML → AI/ML 기반 견고성(Robustness) 높은 워터마크 기술 개발
- 양자 컴퓨팅 → 양자 워터마크 기술 활용

[관련 연구]

- Google, Watermarking AI-generated text and video with SynthID
- Zhengyuan Jiang et al, 제거 및 위조 공격에 견고성 인증된 이미지 워터마크 제안
- Swapnaneel Dhar et al., 양자 워터마크 기술 Survey

03 사이버 보안 관점에서 바라본 저작권 기술의 발전 방향

워터마크 기술

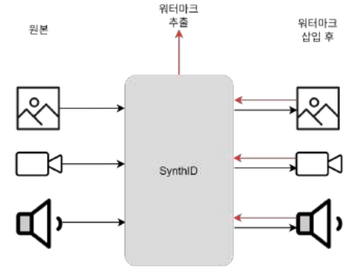
AI/ML – Google SynthID를 활용한 워터마크 삽입 및 추출

※ SynthID : 딥러닝 모델과 알고리즘을 사용해 비가시적 워터마크 삽입 및 추출하는 솔루션

GAN(생성적 적대 네트워크) 기술로 색상 분포, 물체 질감, 조명 등 고유 특징 식별을 훈련하여 기존 콘텐츠 품질 손상 없이 고유한 워터마크 삽입

이미지, 비디오의 개별 프레임 스캔하여 삽입된 워터마크 감지, 수정되어도 감지 가능

음악의 경우 오디오 파형을 학습하여 품질 하락 없이 SynthID 워터마크 삽입



- Zhengyuan Jiang et al, 제거 및 위조 공격에 견고성 인증된 이미지 워터마크 제안
- Swapnaneel Dhar et al., 양자 워터마크 기술 Survey

03 사이버 보안 관점에서 바라본 저작권 기술의 발전 방향

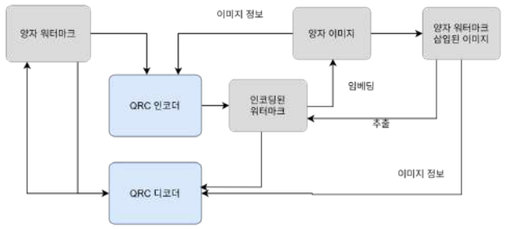
워터마크 기술

양자 컴퓨팅 – 양자 이미지에 양자 워터마크 삽입 및 추출 예시

※ 양자 이미지 : 양자 컴퓨팅 또는 양자 정보 처리를 사용해 생성된 이미지

QRC : Quantum Robust Coding, 워터마크 이미지의 밝기 및 어둠 특성에 관한 워터마크 큐비트 변경

양자 워터마크 임베딩 전, QRC를 통해 워터마크 인코딩 후 이미지에 워터마크 삽입 하여 견고성 확보



- Zhengyuan Jiang et al, 제거 및 위조 공격에 견고성 인증된 이미지 워터마크 제안
- Swapnaneel Dhar et al., 양자 워터마크 기술 Survey

03 사이버 보안 관점에서 바라본 저작권 기술의 발전 방향

DRM 기술 및 보안 강화

DRM 해킹, 정보 유출 방지

- 암호화 → 차세대 암호 기술을 통한 DRM 해킹 방지
- 제로 트러스트 → 제로 트러스트 아키텍처 기반 정보 유출 최소화
- 클라우드 → 클라우드 기반 IAM, MFA, Key 관리로 보안 강화

[관련 연구]

- Nist, Migration to Post-Quantum Cryptography
- Google, BeyondCorp: A New Approach to Enterprise Security (제로트러스트)
- Abhijit Mali, Cloud Security with AWS

03 사이버 보안 관점에서 바라본 저작권 기술의 발전 방향

DRM 기술 및 보안 강화

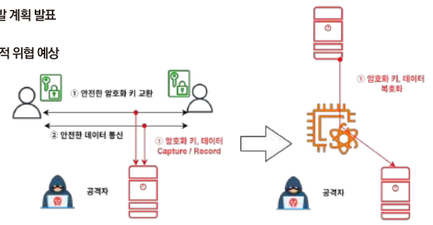
암호화 – PQC를 활용한 안전한 통신

※ PQC : Post-Quantum Cryptography로 양자 컴퓨터로도 풀어내기 어려운 복잡한 알고리즘 사용하는 암호화 방식

양자 컴퓨터는 IBM의 경우 25년 까지 4,000 큐비트, 26년 까지 수만 큐비트 처리 프로세서 개발 계획 발표
2030년까지 현재의 비대칭 암호화 알고리즘인 RSA-2048 공격 가능할 것으로 판단
캐나다 Global Risk Institute는 전문가 대상 설문하여 15년내 50% 이상이 양자 컴퓨터의 현실적 위협 예상

미국 국립표준기술연구소(NIST)는 PQC 알고리즘 표준화 진행하여 23년 8월 초안 발표
IBM은 24년 8월 NIST 포스트 양자 암호 표준에 적합한 알고리즘 2종 채택
현재 NIST는 PQC Migration을 용이하도록 하는 프로젝트 시작

우측 그림과 같은 양자 위협을 방지하기 위한 PQC Migration 준비 필요



- Google, BeyondCorp: A New Approach to Enterprise Security (제로트러스트)
- Abhijit Mali, Cloud Security with AWS

03 사이버 보안 관점에서 바라본 저작권 기술의 발전 방향

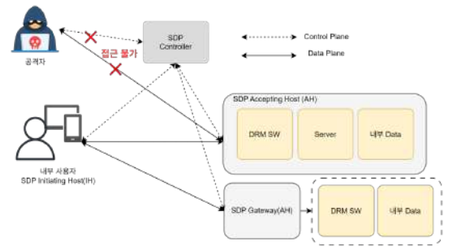
DRM 기술 및 보안 강화

제로 트러스트 – SDP 모델을 적용한 DRM 시스템 및 내부 정보 접근 제어 예시

※ SDP : Software Defined Perimeter, 네트워크 보안 경계 기능을 동적으로 유연하게 배포 가능한 제로 트러스트 모델

SDP Initiating Host : 접속 사용자 장치 등
SDP Controller : 사용자 인증, 권한 부여 등 정책 결정 관리
SDP Accepting Host / Gateway: SDP로 보호되는 자원에 대한 정책 시행

모든 내부 접근에 대해 Controller를 통한 인증 진행 후 접근
공격자는 인증 수행이 불가하여 접근 차단



- Google, BeyondCorp: A New Approach to Enterprise Security (제로트러스트)
- Abhijit Mali, Cloud Security with AWS

03 사이버 보안 관점에서 바라본 저작권 기술의 발전 방향

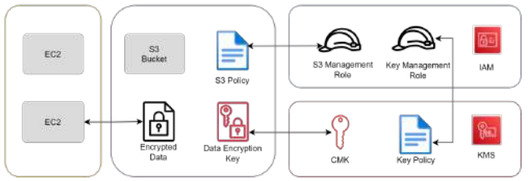
DRM 기술 및 보안 강화

클라우드 – AWS 클라우드 IAM과 KMS를 활용한 보안 강화 예시

※ IAM : Identity and Access Management : 내부 데이터에 대한 세부적인 권한 및 접근 제어 설정
KMS : Key Management Service : 데이터 보호용 암호화 키 제어 서비스

내부 민감 데이터를 AWS Cloud의 S3 등 저장소에 저장하며, 정책을 통해 접근 제어

데이터 암호화를 위한 Key를 KMS 를 통해 관리하고 해당 Key의 수명, 권한 등 정책 설정 및 Key 접근 권한을 Identity 별로 설정



- Google, BeyondCorp: A New Approach to Enterprise Security (제로트러스트)
- Abhijit Mali, Cloud Security with AWS

03 사이버 보안 관점에서 바라본 저작권 기술의 발전 방향

생성형 AI 대응

생성형 AI 콘텐츠 대응

- 정책 → 해외 입법 사례 참고 및 포괄적, 체계적 법제화
- 블록체인 → 블록체인 기술을 활용한 콘텐츠 소유권, 진위성 확보

[관련 연구]

- MIT, Using AI to protect against AI image manipulation (포토가드 활용한 악의적 AI 편집 방지)
- Xiaowan Wang et al., Blockchain-based fake news traceability and verification mechanism
- Xiangli Xiao et al., Blockchain-based reliable image copyright protection

03 사이버 보안 관점에서 바라본 저작권 기술의 발전 방향

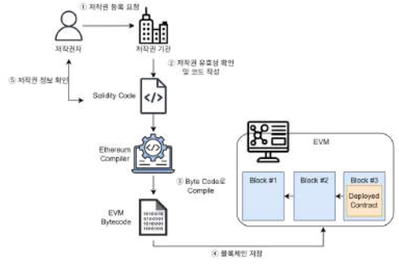
생성형 AI 대응

블록체인 – 이더리움 기반 Smart Contract를 활용한 콘텐츠 저작권 확인 예시

※ Smart Contract : 정해진 조건이 충족되면 자동으로 실행되는 코드

Solidity : Smart Contract를 구현하기 위한 객체 지향 고수준 언어
Ethereum Compiler : Solidity 언어로 작성된 코드를 Bytecode로 변환하는 객체
EVM : Ethereum Virtual Machine의 약어, 이더리움 Smart Contract를 위한 런타임 환경
EVM Bytecode : Compiler에 의해 저수준으로 변환된 EVM이 실행 가능한 코드

저작권 정보를 Code를 이용하여 이더리움 기반 Smart Contract 환경에 저장하여
저작권자의 권리 보호



- Xiaowan Wang et al., Blockchain-based fake news traceability and verification mechanism
- Xiangli Xiao et al., Blockchain-based reliable image copyright protection

03 사이버 보안 관점에서 바라본 저작권 기술의 발전 방향

생성형 AI 대응

정책 – 미국, EU의 AI 입법 현황 및 국내 동향

미국 : AI 분야 주도권 유지 위한 적극적 대응
- 2020년 국가 인공지능 계획법(National Artificial intelligence Initiative Act of 2020) 제정하여 AI 기술 발전 위한 국가 전략 수립 기구, 위원회 설립, 인공 지능 정의
- 2022년 미국 인공지능 진흥법(Advancing American AI Act) 제정하여 인공지능 관련 산업, 연구 활성화 법적 장치 마련, AI 사용 가이드라인 규정
- 2023년 안전하고 신뢰할 수 있는 인공지능 개발 및 사용(Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence) 행정 명령 발표하여 AI 위험 평가 및 관리 강화

유럽 : AI 포괄적 규제 법률 세계 최초 제정
- 2024년 인공지능법(Artificial Intelligence Act) 세계 최초 제정하여 위험 기반 규율로 위험 수준에 따라 순차적으로 인공지능 시스템 규제, 인간 존엄성/자유 평등/기본권 등 중대 침해 우려되는 부분에서 인공지능 관행 금지(Prohibited AI Practices)
- 위 법 기반 생체인식, 중요기반시설, 교육 및 직업훈련 등 고위험 AI 시스템은 출시 전 영향 평가, 출시 후 모니터링, 보고, 위험관리 의무 등 엄격한 기준이 적용

국내
- 2019년 인공지능 국가전략 수립하여 AI 기술 발전 및 윤리적 사용 위한 기반 마련
- 이후 여러 법안 발의 되었으나 아직 포괄적, 체계적 법제화는 이루어지지 않았으며, 22대 국회에서 인공지능 발전 및 신뢰 확보 관련 법률안이 발의되어 진행 중

- Xiaowan Wang et al., Blockchain-based fake news traceability and verification mechanism
- Xiangli Xiao et al., Blockchain-based reliable image copyright protection

04 결론 및 참고자료

결론

사이버 보안은 디지털 자산을 보호하기 위한 기술들로 구성되어 저작권 보호 기술과의 연관성이 높음

디지털 워터마크 견고성(Robustness) 확보, DRM 해킹 방지, 생성형 AI 기반 콘텐츠 식별 등 과제 해결 방향 제시

AI/ML, 양자컴퓨팅, 제로 트러스트, 클라우드, 암호화 기술, 블록체인 등 사이버 보안 관련 기술 기반 해결책 방향성 제시

04 결론 및 참고자료

참고자료

Emily Dean, (March 13, 2024), 100+ Digital Products Statistics for 2024
WEF, (January 2016), Digital Media and Society Implications in a Hyperconnected Era
MasterCard, Out of Sight, Out of Mind: The Problem with Digital Goods and Services
Mckinsey, (May, 2022), The State of Fashion Technology
Simon Kemp, (January 31, 2024), Digital 2024: Global Overwiew Report
Cisco, (March 9, 2020), Cisco Annual Internet Report(2018–2023) White Paper
Rachel Kim, (January 4, 2024), AI and Copyright in 2023: In the Courts
박찬, (August 14, 2024), 이미지 생성 AI 소송에서 '저작권 침해 혐의' 최초 인정… 본격 조사 돌입
박찬, (October 22, 2024), 뉴스 코프, WSJ, 뉴욕포스트 저작권 침해로 퍼플렉시티 고소
이정현, (October 22, 2024), 생성형 AI로 이미지 만들어 쓰다가… 피소당한 일론 머스크
김영명, (April 19, 2023), 故 이우영 작가 사태로 불거진 저작권 이슈, 저작권 보호 기술 어떤 게 있나
장민기, (March 18, 2022), [미국] 아카마이(Akamai), 세계 저작권 침해 현황 조사 보고서 발표
김영희, (July 5, 2024), 디지털 워터마크(Digital Watermark) 산업 현황 보고서
임유정, (August 21, 2017), 구글, AI로 워터마크 제거… '감쪽같네'
임대준, (October 5, 2023), '눈에 안 보이는 워터마크도 무용지물'… 실험결과 모두 '세척' 가능
오성흔, (July 6, 2012), [기술] 주요 저작권 보호 기술의 해킹 사례

04 결론 및 참고자료

참고자료

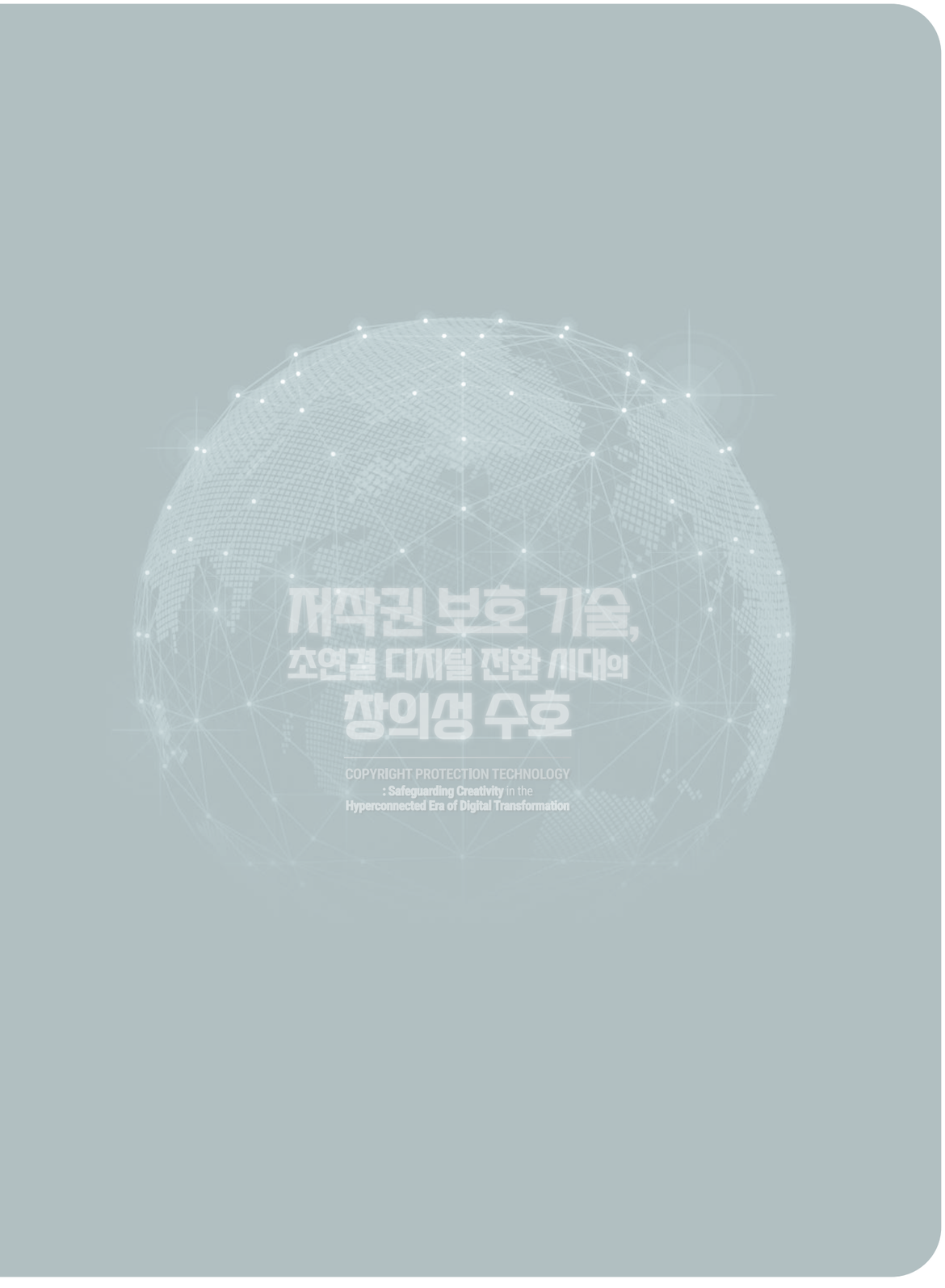
Hiren Dhaduk, (July 11, 2023), AWS HIPAA Compliance: Ensuring Data Security in Healthcare
MIT, (July 31, 2023), Using AI to protect against AI image manipulation
Xiaowan Wang et al., (July, 2023), Blockchain-based fake news traceability and verification mechanism
Xiangli Xiao et al., (March 02, 2023), Blockchain-based reliable image copyright protection
Sumit S Shevtekar, (November, 2022), Fundraising Tracking System Using Blockchain
법제처 미래법제혁신기획단, (July, 2024), 인공지능(AI) 관련 국내외 법제 동향

04 결론 및 참고자료

참고자료

김영명, (June 26, 2023), 알라딘 e북 해킹 유출 사고 한 달… 전자책 업계 빅 5, 어떻게 대응하고 있나
장하민, (September 21, 2023), 알라딘, 입시학원등 해킹한 범인 검거돼… '암호화 키' 관리 실패로 데이터 유출
Dennis Schirrmacher, (April 24, 2024), Researcher extracts DRM key from Microsoft and downloads Netflix movies
Ashraful Islam, https://www.flaticon.com/kr/free-icons/title="암호 아이콘"
Ch.designer, https://www.flaticon.com/kr/free-icons/title="보안 아이콘"
Risman Muhammad, https://www.flaticon.com/kr/free-icons/title="진화 아이콘"
Google, (May 14, 2024), Watermarking AI-generated text and video with SynthID
EMRE CITAK, (September 4, 2023), Google Declares war on deepfakes
Zhengyuan Jiang et al., (July 4, 2024), Certifiably Robust Image Watermark
Swapnaneel Dhar et al., (November, 2024), Digital to quantum watermarking: A journey from past to present and into the future
Zheng Xing, (August 5, 2024), Quantum Robust Coding for Quantum Image Watermarking
Nist, (April 24, 2023), Migration to Post-Quantum Cryptography: NIST SP 1800-38A Preliminary Draft Available for Comment
Rory Ward et al., (2014), BeyondCorp: A New Approach to Enterprise Security
Abhijit Mali, (November, 2022), Cloud Security with AWS
조지훈 et al., (March 15, 2023), 양자컴퓨팅으로 더 커지는 보안 위협, 지금이 바로 PQC 전환을 시작할 때!
Jason Garbis et al., (March 10, 2022), Software-Defined Perimeter (SDP) Specification v2.0

감사합니다.



초청연설

디지털 시대의 EU 저작권 인프라 강화: EUIPO의 새로운 저작권 계획

해리 테밍크 | 유럽연합 지식재산청(EUIPO) 서비스 총괄

초청연설

디지털 시대의 EU 저작권 인프라 강화: EUIPO의 새로운 저작권 계획

해리 테밍크

유럽연합 지식재산청(EUIPO) 서비스 총괄

연사 이력

- 2023년 2월부터 유럽연합 지식재산청(EUIPO)에서 "디지털 세계의 지식재산(IP)" 총괄로 근무 중
- 온라인 중개자, 저작권, 법률 문제, 지식 구축, EU 외부의 지식재산권(IPR) 집행을 감독
- 유럽연합 집행위원회(GROW) 내무시장, 산업, 기업가정신 및 중소기업 총국(DG GROW) 전 부국장
- 유해 화학물질(REACH), 화학법 집행, 지식재산권, 위조 방지, 전자상거래, 중개자 책임, 우편 및 소포 서비스를 관리
- 유럽 소비자보호위원 메글레나 쿠네바(Meglena Kuneva) 내각의 전 구성원
- 유럽연합 사법재판소 법률서기관으로 근무
- 전 위트레흐트 대학교 공공경제법 강사
- 네덜란드 경쟁 당국의 고문으로 근무
- 위트레흐트 대학교에서 네덜란드 법학 학위 및 스페인어 및 문학 학위 소지

발표 내용

EUIPO(유럽연합 지식재산청)의 2030 전략 계획 초안에서는 저작권을 우선 과제로 삼고 있으며, EU 저작권 정책의 이행을 촉진하려고 합니다. 이 계획에는 저작권 관련 기존 이니셔티브(상업적 이용이 불가능한 작품 포털 포함)를 통합하고, 점진적으로 새로운 프로젝트를 개발할 EUIPO 저작권 지식 센터를 설립할 예정입니다. 이 센터는 저작권에 관한 기존 및 미래의 정보 자원을 하나의 랜딩 페이지에 결합한 "저작권 지식 허브"를 포함할 것으로 예상됩니다.
또한, EUIPO는 국가 및 EU 저작권 데이터베이스의 정보를 통합하여 작품의 저작권 상태, 저자, 소유권 등을 확인할 수 있는 고급 검색 기능을 제공하는 "CopyrightView" 데이터베이스를 만들 계획입니다.
아울러, EUIPO는 저작권자가 텍스트 및 데이터 마이닝에 대한 옵트아웃(거부권)을 표시할 수 있도록 하는 서비스를 개발하는 가능성도 모색하고 있으며, 이는 해당 옵트아웃을 준수해야 하는 AI 회사들에도 혜택을 줄 것입니다.

The draft EUIPO Strategic Plan 2030 considers copyright a priority and promote the implementation of EU copyright policies. It is envisaged that the Office will establish the EUIPO Copyright Knowledge Center which will consolidate ongoing copyright initiatives (including the out of commerce works portal) and gradually develop new projects. This should include a “Copyright Knowledge Hub” that will combine, in a single landing page, existing and future information resources on copyright. Furthermore, the EUIPO plans to create a “CopyrightView” database consolidating information from national and EU copyright databases, offering advanced search functionalities to determine the copyright status, authorship and ownership of works. The EUIPO is also exploring the possibility of developing services facilitating opt-out for text and data mining expressed by copyright holders that should benefit also AI companies that must respect the opt-out,

1

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

Overview

- **Introduction** EUIPO Observatory
- Copyright in the digital age – **current** initiatives of the EUIPO Observatory
  - Online Copyright Infringement in the European Union (study 2024)
  - “Good practices” online intermediaries to combat infringement of copyright
  - Combat live event piracy
  - Out of commerce Works Portal
- Creation **EUIPO Copyright Knowledge Centre** (envisaged for 2025)
- **Future** initiatives
  - AI and copyright
  - “CopyrightView”
- **Final comments**

2

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

Introduction EUIPO Observatory

1

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

개요

- EUIPO 관측소 **소개**
- 디지털 시대의 저작권 – EUIPO 관측소의 현재 이니셔티브
  - EU 온라인 저작권 침해(2024년 연구)
  - 저작권 침해 방지를 위한 온라인 중개자의 "모범 사례”
  - 라이브 이벤트 불법 복제 타파
  - 비상용 저작물 포털
- **EUIPO 저작권 지식 센터** 설립(2025 예정)

- **향후 과제**
  - AI와 저작권
  - “CopyrightView”
- **최종 의견**

2

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

EUIPO 관측소 소개

3

EUIPO
EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

European Observatory on Infringements of IP rights – General objectives

Build trust and respect for IP → Support EU policies

Goal 1	Goal 2	Goal 3
Facts & evidence	Tools & resources for enforcement	Awareness

Strengthening the network → International cooperation

3

EUIPO
EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

IP 권리 침해에 대한 유럽 관측소 - 일반 목표

IP에 대한 신뢰와 존중 구축 → EU 정책 지원

목표 1	목표 2	목표 3
사실과 증거	시행을 위한 도구 및 리소스	인식

네트워크 강화 → 국제 협력

EUIPO
EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

Contribution EUIPO Observatory to intellectual property

EUIPO
EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

지적 재산권에 대한 EUIPO 관측소 의 기여

5

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

EUIPO Observatory supports EU policies

In particular, the implementation of EU legal instruments on copyright such as:

- ✓ Recommendation on **combatting online piracy** of sports and other live events, 2023
- ✓ **Digital Services Act** (DSA), 2022
- ✓ Directive on copyright and related rights in the Digital Single Market (**Out of Commerce Works Portal**), 2019
- ✓ Directive on certain permitted uses of **Orphan works**, 2012
- ✓ Directive on the **Enforcement** of intellectual property rights, 2004

6

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

Online Copyright Infringement in the EU

5

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

EUIPO 관측소는 EU 정책 지원

특히 저작권 관련 EU 법률 도구 실행 :

- ✓ 스포츠 및 기타 라이브 이벤트의 **온라인 불법 복제 방지**에 대한 권고안, 2023
- ✓ 디지털 서비스법(DSA) 2022.
- ✓ 디지털 단일 시장(**비상용 저작권 포털**)의 저작권 및 관련 권리에 대한 지침2019
- ✓ **고아저작물**의 특정 허용 사용에 대한 지침 2012
- ✓ 지적 재산권 **집행**에 대한 지침, 2004

6

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

EU 온라인 저작권 침해

7

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

Trends in digital copyright infringement in the European Union

- EUIPO economic study originally from 2019 with regular updates, next version "Online Copyright Infringement in the European Union" scheduled to be published in November 2024
- Films, Music, Publications, Software and TV
- Analysis per EU Member State, per type of content, access method, type of device…
- New for 2024: illegal streaming of **IPTV piracy** (over Internet Protocol networks)

- Key findings 2024 (tentative):
  - **Overall piracy** stabilizing at about **10.5 accesses** per internet user per month
    - **TV piracy** in the EU stabilized in 2023 at 5.2 accesses per user/month but with variations
    - **Film piracy** decreased by 38% in 2023
    - **Music** piracy slightly increased at 0.64 accesses per internet user/month
    - **Publications** piracy remains flat with an average of 2.7 access per internet user/month
    - **Software piracy** increased 6% in 2023
    - **Sports live event piracy** increased from 2021 to 2023, peaking in the summer months
  - Certain **substitution between pirated and legal content**

8

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

Good practices (online) intermediaries to combat copyright infringements

7

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

EU 디지털 저작권 침해 동향

- EUIPO의 경제 연구는 2019년에 시작되어 정기적으로 업데이트 중 다음 버전인 "유럽연합의 온라인 저작권 침해"가 2024년 11월에 발행될 예정
- 영화, 음악, 출판물, 소프트웨어, TV
- EU 회원국별, 콘텐츠 유형별, 접근 방법별, 장치 유형별 분석...
- 2024년의 새로운 내용: **IPTV 불법 스트리밍**(인터넷 프로토콜 네트워크를 통한)

- 2024년 주요 발견 (잠정적):
  - **전반적인 저작권 침해**는 월 평균 인터넷 사용자당 약 **10.5회**로 안정세를 보임.
    - EU 내 **TV 저작권 침해**는 2023년에 사용자 1인당 월 5.2회로 안정화되었으나 변동 있음
    - **영화 저작권 침해**는 2023년에 38% 감소
    - **음악 저작권 침해**는 인터넷 사용자 1인당 월 0.64회로 약간 증가
    - **출판물 저작권 침해**는 평균 1인당 월 2.7회로 유지
    - **소프트웨어 저작권 침해**는 2023년에 6% 증가
    - **스포츠 라이브 이벤트 저작권 침해**는 2021년부터 2023년까지 증가하며 여름철에 정점을 찍음
  - **불법 콘텐츠와 합법 콘텐츠 간의 대체 관계 발견**

8

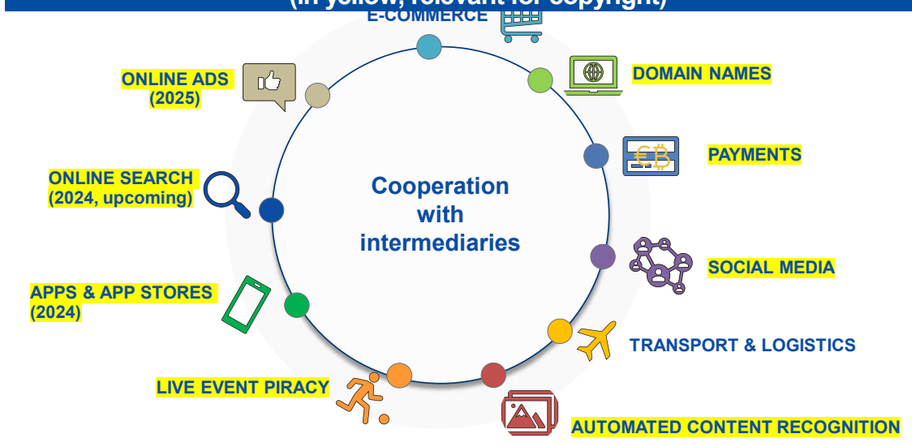
EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

저작권 침해에 맞서는 (온라인) 중개자 우수 사례

9

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

"Good practices" papers online intermediaries to combat misuse of IP (in yellow, relevant for copyright)



9

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

"IP 남용을 막기 위한 온라인 중개업체의 우수 사례" (노란색은 저작권 관련)

10

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

Live event piracy

10

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

라이브 이벤트 불법 복제

EUIPO
EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

CHALLENGES AND SPECIFICITIES OF LIVE EVENT PIRACY

Live nature

- Main commercial value during live transmissions

Unauthorized retransmissions of live sport and other cultural event

- Significant loss in revenues
- Undermines the viability of the services

New methods of piracy and resilience strategy

- Increasingly sophisticated means (illegal IPTV, apps, website, 'Piracy-as-a-Service') / mirrors

Misuse of intermediaries services (upstream, downstream)

- Crucial role intermediaries in assisting holders of rights
- Need for effective legal tools tailored to respective functions of intermediaries

11

EUIPO
EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

Tackling live event piracy The EU Journey

MAY 2021 — European Parliament resolution **on the challenges of sports events**

MARCH 2023 — Publication of the EUIPO **Live Piracy Discussion Paper**

MAY 2023 — Adoption of the European **Commission Recommendation**

JULY 2023 — **Publication of Key Performance Indicators** after stakeholders' consultations

OCT 2023 — **High level Conference** Implementation of the Recommendation and first meeting of the public sector network

12

EUIPO
EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

라이브 이벤트 불법 복제 문제점 및 특징

라이브 환경

- 라이브 전송 중 주요 상업적 가치

라이브 스포츠 및 기타 문화 행사의 무단 재전송

- 상당한 매출 손실
- 서비스 실행 가능성 훼손

새로운 해적 행위 방법 및 회복력 전략

- 점점 정교해지는 수단(불법 IPTV, 앱, 웹사이트, 'PaaS')/미러

아웃리치 인식 및 교육

- 권리 보유자를 지원하는 중요한 중개자 역할
- 중개자 각 기능에 맞게 조정된 효과적 법적 도구 필요성

11

EUIPO
EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

라이브 이벤트 불법 복제 대처 EU 여정

2021년 5월 — 스포츠 이벤트 과제에 대한 유럽 의회 결의안

2023년 3월 — EUIPO 라이브 해적 행위 논의 문서 발행

2023년 5월 — 유럽 위원회 권고안 채택

2023년 7월 — 이해 관계자 협의 후 주요 성과 지표 발행

2023년 10월 — 고위급 회의 권고안 이행 및 공공 부문 네트워크 첫 회의

12

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

European Commission Recommendation on combating online piracy of sports and other live events

Adopted on 4 may 2023
Evaluation by 17 Nov 2025

NOTICES on live events: Prompt treatment & Cooperation

INJUNCTIONS: Dynamic injunctions, Safeguards & Voluntary Cooperation

RAISING AWARENESS and VOLUNTARY COOPERATION between public authorities

FOLLOW UP AND MONITORING

13

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

https://www.euipo.europa.eu/en/observatory/enforcement/combating-piracy

Combating piracy

The EUIPO Observatory is supporting the fight against online piracy, including online piracy of live events.

In line with the European Commission's Recommendation on combating online piracy of live events (the Recommendation'), the EUIPO has established a specialised network of national administrative authorities to facilitate regular information exchange and will assist the European Commission in monitoring the implementation and effects of the Recommendation. The EUIPO is also encouraged to develop and organise knowledge-building activities for national judges and authorities in this area (see the Observatory's knowledge-building events page).

The Observatory has already explored several topics on the fight against IP infringement online. The discussion paper on Live Event Piracy is one such example. Further information can also be found in Observatory publications.

Dedicated network of administrative authorities

In line with the European Commission's Recommendation, the Observatory has established a dedicated network of administrative authorities (Dedicated Network) to regularly exchange information on the measures and good practices used to address the issues covered in the Recommendation and the challenges encountered.

The Dedicated Network is composed of

- representatives from administrative authorities with specific competence in the enforcement of intellectual property rights, particularly online piracy, and/or
- representatives from other administrative authorities that are competent in policy or regulatory developments relating to the enforcement of intellectual property rights, particularly online piracy.

Further information

Mission statement of the Dedicated Network

Monitoring live event piracy

The Observatory is supporting the European Commission in **monitoring the effects** of the Recommendation by gathering data from national authorities, rights holders and a number of intermediary services.

There will be 3 milestones for data collection during the period January 2024 to June 2025

- April 2024: gathering of test data
- January 2025: gathering of data for the period January to December 2024
- July 2025: gathering of data for the period January to June 2025

Monitoring timeline

2024 - 2025

April 2024	January 2025	July 2025
Test data gathering for Q1 2024	data gathering for 2024	data gathering for 2025
Testing batch	Batch for 2024	Batch for Q1-Q2 2025
Period covered 1 January 2024 to 31 March 2024	Period covered 1 January 2024 to 31 December 2024	Period covered 1 January 2025 to 30 June 2025

The data, based on identified key performance indicators (KPIs), can be submitted using standardised templates to the EUIPO Observatory at the following address: observatory.piracy@euipo.europa.eu.

From this data analysis, the European Commission will assess the effects of the Recommendation on unauthorised retransmissions of live sports and other live events, no later than 17 November 2025.

Templates

Participants are kindly invited to indicate the 'year' and 'name of organisation' when naming the file. It is possible to provide additional clarifications by email for further sub-transtation. For encrypted transmissions you may use the PGP Public Key

- Advertising and payment services
- Content delivery networks (CDN)
- Holders of rights
- Hosting providers/Online platforms/Dedicated service providers
- Internet access providers
- Public and national authorities

14

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

스포츠 및 기타 라이브 이벤트 온라인 불법 복제 방지 관련 유럽 위원회 권고안

2023년 5월 4일 채택
2025년 11월 17일까지 평가

라이브 이벤트 공지: 신속한 처리 및 협조

가처분 명령: 동적 가처분 명령, 보호 조치 및 자발적 협력

공공 기관 간 인식 제고 및 자발적 협력

후속 조치 및 모니터링

13

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

https://www.euipo.europa.eu/en/observatory/enforcement/combating-piracy

Combating piracy

The EUIPO Observatory is supporting the fight against online piracy, including online piracy of live events.

In line with the European Commission's Recommendation on combating online piracy of live events (the Recommendation'), the EUIPO has established a specialised network of national administrative authorities to facilitate regular information exchange and will assist the European Commission in monitoring the implementation and effects of the Recommendation. The EUIPO is also encouraged to develop and organise knowledge-building activities for national judges and authorities in this area (see the Observatory's knowledge-building events page).

The Observatory has already explored several topics on the fight against IP infringement online. The discussion paper on Live Event Piracy is one such example. Further information can also be found in Observatory publications.

Dedicated network of administrative authorities

In line with the European Commission's Recommendation, the Observatory has established a dedicated network of administrative authorities (Dedicated Network) to regularly exchange information on the measures and good practices used to address the issues covered in the Recommendation and the challenges encountered.

The Dedicated Network is composed of

- representatives from administrative authorities with specific competence in the enforcement of intellectual property rights, particularly online piracy, and/or
- representatives from other administrative authorities that are competent in policy or regulatory developments relating to the enforcement of intellectual property rights, particularly online piracy.

Further information

Mission statement of the Dedicated Network

Monitoring live event piracy

The Observatory is supporting the European Commission in **monitoring the effects** of the Recommendation by gathering data from national authorities, rights holders and a number of intermediary services.

There will be 3 milestones for data collection during the period January 2024 to June 2025

- April 2024: gathering of test data
- January 2025: gathering of data for the period January to December 2024
- July 2025: gathering of data for the period January to June 2025

Monitoring timeline

2024 - 2025

April 2024	January 2025	July 2025
Test data gathering for Q1 2024	data gathering for 2024	data gathering for 2025
Testing batch	Batch for 2024	Batch for Q1-Q2 2025
Period covered 1 January 2024 to 31 March 2024	Period covered 1 January 2024 to 31 December 2024	Period covered 1 January 2025 to 30 June 2025

The data, based on identified key performance indicators (KPIs), can be submitted using standardised templates to the EUIPO Observatory at the following address: observatory.piracy@euipo.europa.eu.

From this data analysis, the European Commission will assess the effects of the Recommendation on unauthorised retransmissions of live sports and other live events, no later than 17 November 2025.

Templates

Participants are kindly invited to indicate the 'year' and 'name of organisation' when naming the file. It is possible to provide additional clarifications by email for further sub-transtation. For encrypted transmissions you may use the PGP Public Key

- Advertising and payment services
- Content delivery networks (CDN)
- Holders of rights
- Hosting providers/Online platforms/Dedicated service providers
- Internet access providers
- Public and national authorities

14

15

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

Out of Commerce Works Portal

15

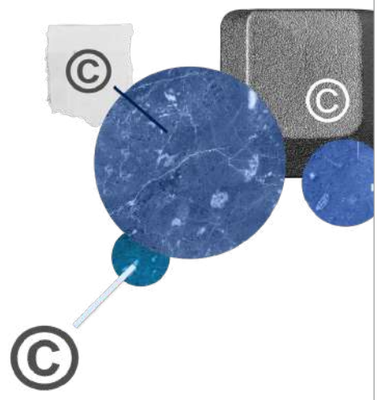
EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

비상업적 작품

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

Out of commerce works

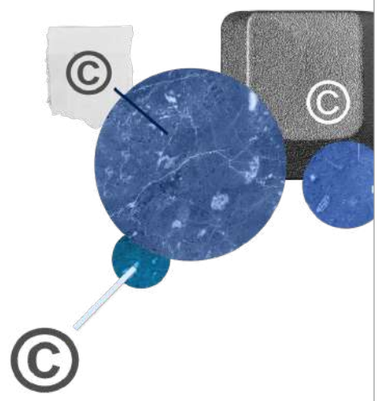
Out-of-commerce works are works that are still protected by copyright but are **no longer or have never been commercially available** (a reasonable effort to determine their unavailability)



EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

비상업적 작품

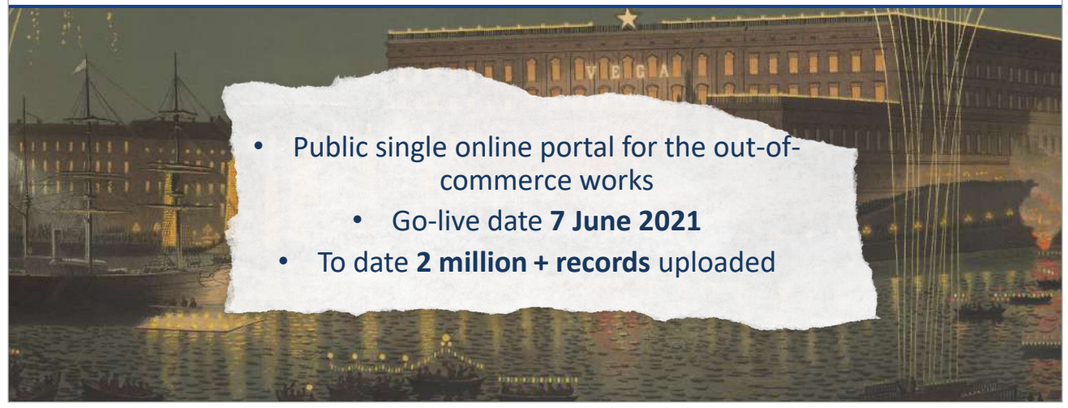
비상업적 작품은 저작권으로 보호받고 있지만 더 이상 상업적으로 이용 가능하지 않거나 과거에 상업적으로 이용 가능하지 않았던 작품을 뜻함(불가용성 판단을 위한 합리적 노력 필요)



Background

- New out-of-commerce works regime under the **Directive (EU) 2019/790**, to enable mass digitization projects from cultural heritage institutions
- New mandatory **exception/limitation** to copyright (to allow cultural heritage institutions to digitalize) with an **"opt out"** mechanism (in case right holders such as authors and publishers do not wish their works to be used this way)
- Role of EUIPO Observatory in establishing and managing **a single online portal** for out-of-commerce works

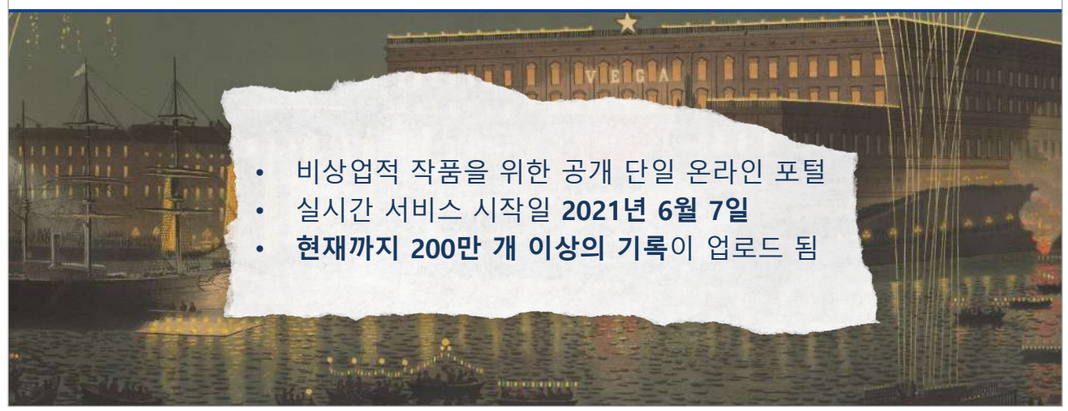
The Out of Commerce Works Portal



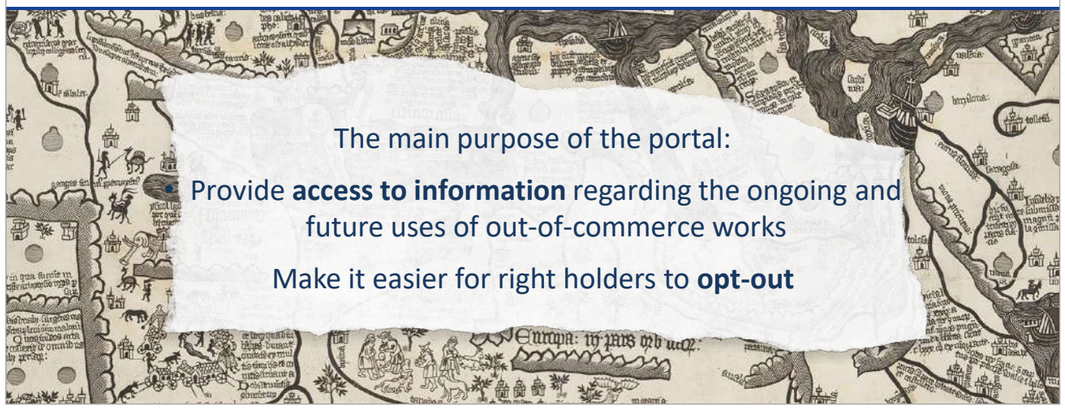
배경

- **EU 지침 (EU) 2019/790에 따른 새로운 상용되지 않는 저작물 시스템**
  문화유산 기관의 대규모 디지털화 프로젝트 가능함
- **새로운 의무적 저작권 예외/제한**
  문화유산 기관이 디지털화할 수 있도록 하며, 권리자가 자신의 저작물이 사용되는 것을 원하지 않을 경우 **"옵트 아웃"** 메커니즘 포함
- **EUIPO 관측소의 기능**
  상용되지 않는 저작물에 대한 단일 온라인 포털을 설계하고 관리

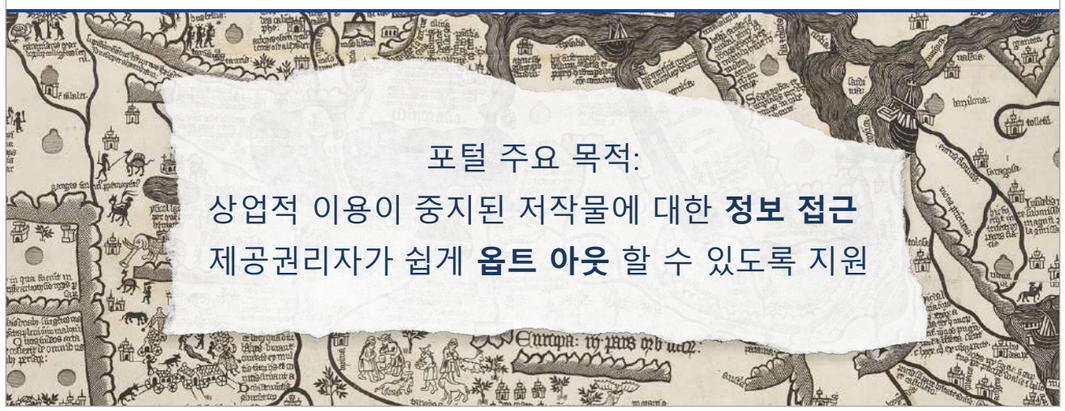
비상업적 작품 포털



Purpose of the Portal



포털의 주요 목적

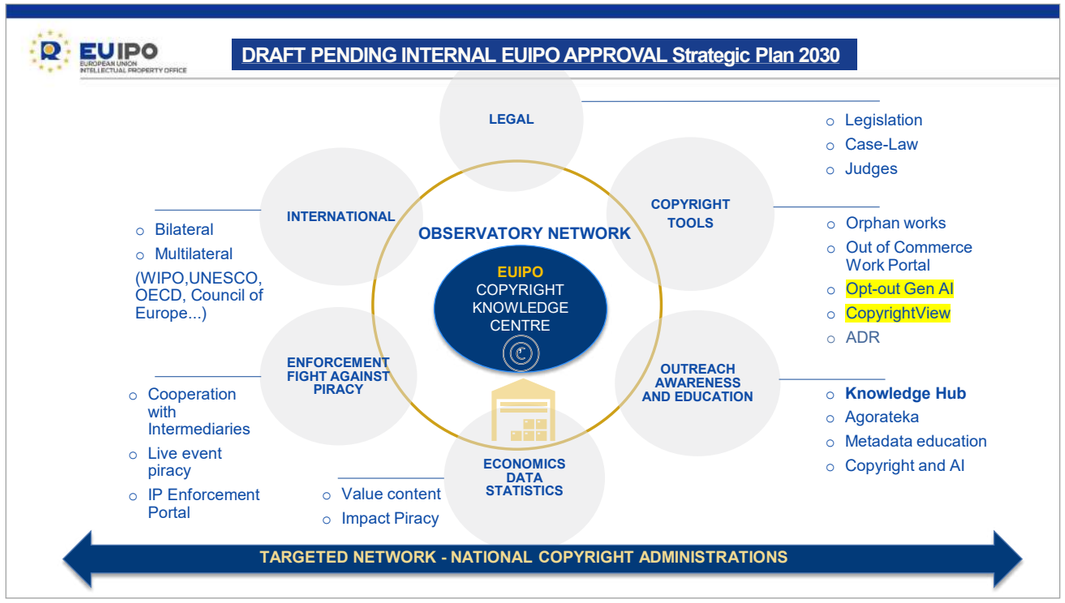


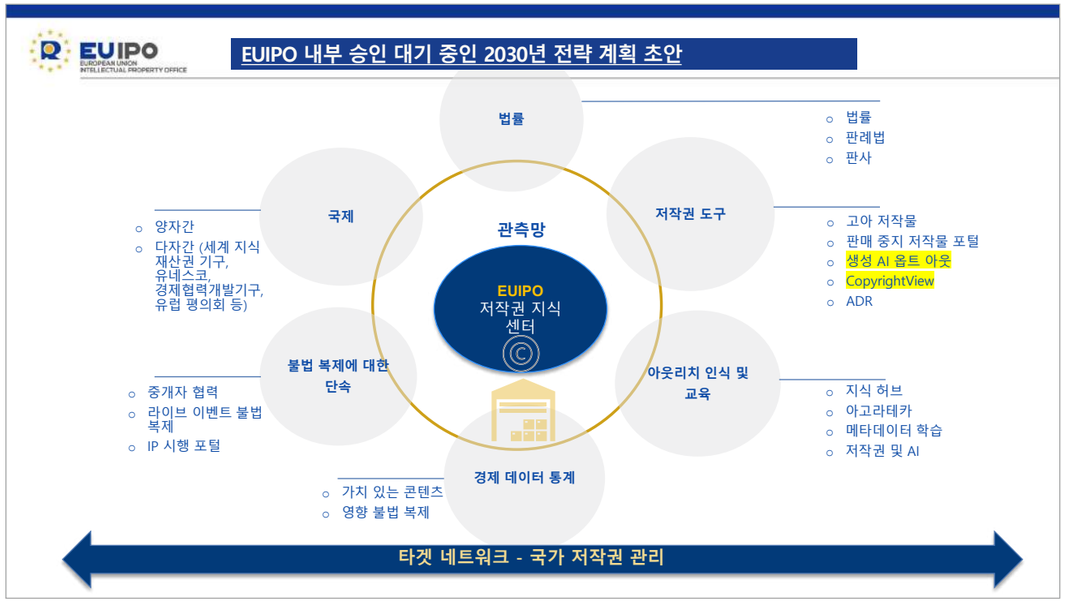
20

Creation “EUIPO Copyright Knowledge Centre” (2025)

20

“EUIPO 저작권 지식 센터” 설립(2025)





22

EUIPO
EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

AI and copyright – technical study

22

EUIPO
EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

AI와 저작권 - 기술 연구

23

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

EU Artificial Intelligence Act (2024) and copyright

- The new EU law on AI includes two requirements for providers of general-purpose AI models:
  - To put in place a policy to respect Union copyright law in particular to identify and respect **the reservation of rights ("opt out") under the Text and Data Mining exception** of the Copyright in the DSM Directive (2019)
  - To draw up and make publicly available a sufficiently detailed summary about the content used for training, according to a template
- Aim: to support the enforcement of EU copyright rules in the context of the development of generative AI

23

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

EU 인공지능법(2024) 및 저작권

- 새로운 EU AI 법률, 일반 목적 AI 모델 제공자에게 두 가지 조건 요구:
  - 유럽 연합 저작권법을 준수하기 위한 정책 수립, 저작권 DSM 지침(2019)의 **텍스트 및 데이터 마이닝 예외에 따라 권리 예약("옵트 아웃")**을 식별 및 존중
  - 교육에 사용된 콘텐츠에 대한 충분히 상세한 요약 작성 및 공개
- 목적: 생성적 AI 개발에 대한 EU 저작권 규칙 집행 지원

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

GenAI

EUIPO study on GenAI and Copyright - INPUT

KEY ELEMENTS OF THE ANALYSIS

USE OF CONTENT

Through different methods & at the different phases of GenAI system training

SOLUTIONS & PRACTICES

To limit, reserve, support access or license the use of copyright protected works as training materials

TDM·AI
robots.txt
IPTC
TDM Reservation Protocol (TDMRep)
Final Community Group Report 02 February 2024
RIGHTSANDAI.COM
Fairly Trained Certified

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

GenAI

EUIPO의 생성형 AI와 저작권 연구 결과- 입력

분석의 핵심 요소

콘텐츠 사용

생성형AI 시스템 훈련의 다양한 단계와 방법을 통해

솔루션 및 실행

저작권 보호 작품의 사용을 훈련 자료로 제한, 예약, 지원하거나 라이센스 부여

TDM·AI
robots.txt
IPTC
TDM Reservation Protocol (TDMRep)
Final Community Group Report 02 February 2024
RIGHTSANDAI.COM
Fairly Trained Certified

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

GenAI

EUIPO study on GenAI and Copyright OUTPUT

KEY ELEMENTS OF THE ANALYSIS

GENERATION OF CONTENT

Methods & phases including defining and analysing prompts, refining output, content moderation and validation

SOLUTIONS & PRACTICES

To identify and/or mark AI generated works

C2PA

Meta
Labeling AI-Generated Images on Facebook, Instagram and Threads

STEAM

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

CopyrightView

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

GenAI

EUIPO의 생성형 AI와 저작권 연구 결과

분석의 핵심 요소

콘텐츠 생성

프롬프트 정의 및 분석, 출력 정제, 콘텐츠 조정 및 검증을 포함한 방법 및 단계

솔루션 및 실행

AI로 제작된 저작물을 확인하거나 표시하기 위해

C2PA

Meta
Labeling AI-Generated Images on Facebook, Instagram and Threads

STEAM

26

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

CopyrightView

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

Copyright View

A **central EU copyright search service**, linking the **existing databases providing information on copyright** (incl. national deposit systems, registries, orphan and out of commerce databases, private copyright registers, 'CMO' databases) with **advanced search functionalities** for participating databases.

27

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

Objectives

- First step in contributing to the development of an EU copyright infrastructure
- To the benefit of creators in the management of their rights and users to identify works
- Improving quality of and access to information in existing copyright databases

Gradual approach in 4 successives steps

- Map existing databases (public/private, national/EU, compulsory/voluntary)
- Engage with relevant stakeholders to discuss feasibility of a search functionality
- Engage in standardisation initiatives on content identifiers
- Set up and operate search tools.

Flexible approach

- Journey with a clear destination but multiple paths
- Interaction with other initiatives EUIPO Copyright Knowledge Centre

28

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

CopyrightView

중앙 EU 저작권 검색 서비스: **저작권 관련 정보를 제공하는 기존 데이터베이스**(국가 예치 시스템, 등록부, 고아 저작물 및 상용되지 않는 저작물 데이터베이스, 민간 저작권 등록부, 'CMO' 데이터베이스 포함) 연결 및 참여하는 데이터베이스를 위한 **고급 검색 기능** 제공

27

EUIPO EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

목표

- EU 저작권 인프라 개발에 기여하는 첫 단계창작자들이 권리를 관리하는 데 도움
- 사용자들이 저작물을 식별할 수 있도록 지원
- 기존 저작권 데이터베이스에서 정보의 질과 접근성 개선

점진적인 접근 방식: 4단계

- 기존 데이터베이스 매핑 (공공/민간, 국가/EU, 의무/자발적)
- 검색 기능 실현 가능성 논의를 위해 관련 이해관계자 참여
- 콘텐츠 식별자에 대한 표준화 이니셔티브에 참여
- 검색 도구를 설정하고 운영

유연한 접근 방식

- 명확한 목적지를 가진 여정, 하지만 다양한 경로가 존재
- EUIPO 저작권 지식 센터와의 다른 이니셔티브와의 상호작용

28

29

EUIPO
EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

FINAL COMMENTS

29

EUIPO
EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

최종 의견

EUIPO
EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

www.euipo.europa.eu

@EU_IPO
EUIPO
EUIPO.EU
@EUIPO
EUIPO

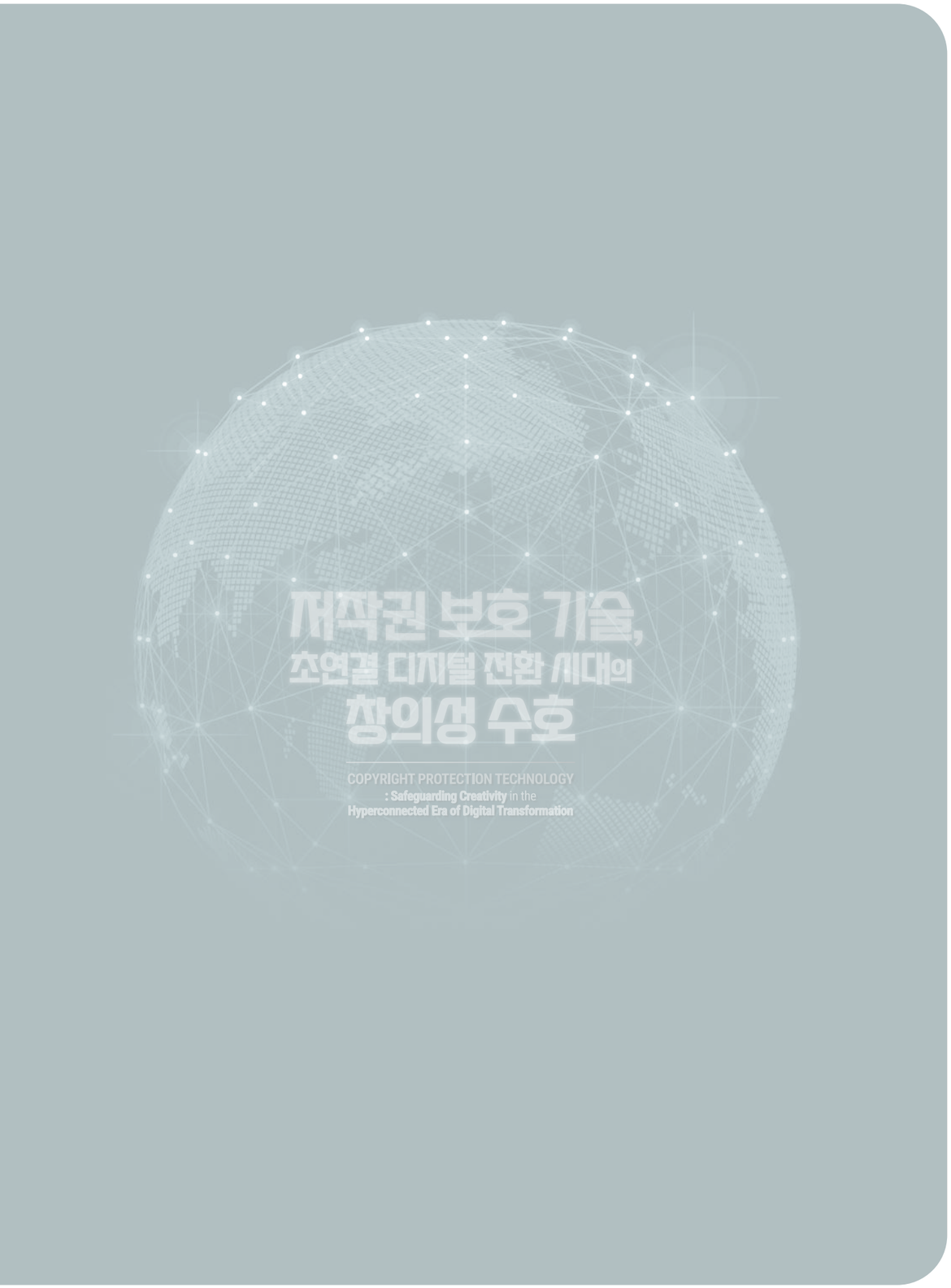
THANK YOU

EUIPO
EUROPEAN UNION INTELLECTUAL PROPERTY OFFICE

www.euipo.europa.eu

@EU_IPO
EUIPO
EUIPO.EU
@EUIPO
EUIPO

감사합니다



Session 1

디지털 혁신 속 저작권 보호 기술

Ⅰ 물리적 복제 불가능 기술과 저작권 보호 기술

박욱 | 경희대학교 교수

Ⅱ 생성형 AI 시대의 콘텐츠 진위성

일케 데미르 | 인텔 선임 연구원

Ⅲ 30년간 비가시성 워터마크를 연구한 기업이 말해주는 "다양한 비가시성 워터마크 활용" 및 "생성형 AI 콘텐츠 위한 고속 비가시성 워터마킹 기술" 이야기

최고 | 마크애니 대표

Ⅳ 저작물 방송 송출을 위한 콘텐츠 보안 적용

로날드 휠러 | A3SA 전무이사

Session 1 디지털 혁신 속 저작권 보호 기술

I 물리적 복제 불가능 기술과 저작권 보호 기술

박욱
경희대학교 교수

연사 이력

- 경희대학교 전자공학과 교수(2012~)
- 경희대학교 산학협력단 부단장(2022~)
- 경희대학교 기술지주회사 이사(2022~)
- 서울대학교 반도체공동연구소 연구원(2011~2012)
- 대통령포스닥펠로우십(2011~2015)

발표 내용

Physical Unclonable Functions (PUFs)는 물리적인 구조의 미세한 불규칙성을 이용해 복제가 불가능한 고유의 암호학적 키를 생성하는 기술로, 각 PUF는 동일한 환경에서조차 완벽히 동일한 출력을 만들 수 없는 특성을 지닙니다. 이러한 특성 덕분에 PUF는 사물인터넷(IoT) 기기의 인증과 암호화 과정에서 중요한 보안 요소로 부각되고 있습니다.

본 발표에서는 **4차원 PUFs(4D PUFs)**를 통해 시간에 따라 변화하는 혼돈적인 발광 패턴을 구현한 새로운 방식의 보안 기술을 소개합니다. MoS2 원자 씨앗을 기반으로 형성된 4D PUF는 서로 다른 수명을 지닌 발광체들이 결합된 상태를 형성하며, 그 결과로 나타나는 비정규적인 수명 분포는 복제할 수 없는 무작위성을 제공합니다. 이를 활용한 비트스트림 생성 및 암호화/인증 과정이 시연되며, 기존의 광학 PUF와 비교하여 훨씬 강력한 위조 방지 및 보안 성능을 자랑합니다.

이 기술은 저작권 보호에도 중요한 응용 가능성을 지니고 있습니다. 제품에 부착되는 위조 방지 태그는 저작권 보호의 핵심 도구로, 경제적이며 대량 생산이 가능하고 빠른 인증 절차를 필요로 합니다. 본 발표에서는 레이저 어블레이션을 활용하여 무작위로 분포된 크레이터 패턴을 형성하는 새로운 태그 생성 기법을 소개합니다. 이 태그는 레이저를 이용해 짧은 시간 내에 생성되며, 고유한 무작위 패턴을 통해 제품의 진위를 확인할 수 있습니다. 이러한 태그는 NIST 통계 테스트를 통해 높은 무작위성과 매우 낮은 오차율을 입증했습니다.

결론적으로, 4D PUFs와 레이저 기반 위조 방지 태그 생성 기술은 저작권 보호와 위조 방지에서 획기적인 보안 솔루션을 제공하며, 향후 IoT 기기와 디지털 콘텐츠 보호에 중요한 역할을 할 것으로 기대됩니다.

Physical Unclonable Functions (PUFs) are a technology that utilizes the subtle physical irregularities in material structures to generate unique cryptographic keys that are impossible to replicate. Even in identical environments, each PUF produces outputs that are distinct, making it a crucial security element for authentication and cryptographic processes in Internet of Things (IoT) devices.

In this presentation, we introduce a novel security technology using **four-dimensional PUFs (4D PUFs)** that implement time-varying chaotic emission patterns. These 4D PUFs are based on MoS2 atomic seeds, forming hybrid states with emitters of varying lifetimes, resulting in irregular lifetime distributions that cannot be duplicated. The process of generating bitstreams from these patterns is demonstrated, highlighting their use in encryption and authentication. Compared to traditional optical PUFs, 4D PUFs exhibit significantly enhanced counterfeit deterrence and security performance.

This technology also has important implications for copyright protection. Anti-counterfeiting tags affixed to products are a key tool in copyright enforcement, and these tags must be cost-effective, mass-producible, and capable of rapid authentication. In this presentation, we introduce a **laser ablation** technique that creates randomly distributed crater patterns on laser-sensitive materials. These patterns, generated in a short amount of time, provide unique randomness that can be digitized to verify product authenticity. The efficacy of these tags is demonstrated through statistical NIST tests, which show high randomness and extremely low error rates.

In conclusion, 4D PUFs and laser-based anti-counterfeiting tag generation techniques offer groundbreaking security solutions for copyright protection and counterfeiting deterrence. These technologies are expected to play a critical role in the future protection of IoT devices and digital content.

물리적 복제 불가능 기술과 저작권 보호 기술

경희대학교 박욱

Physically unclonable technology and copyright protection technology

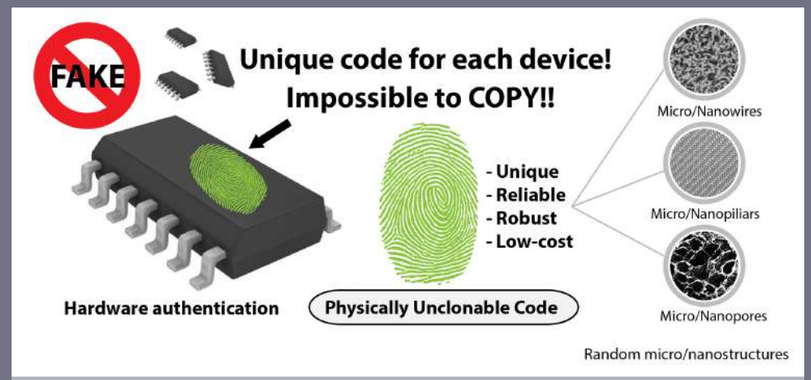
Wook Park

물리적 복제 불가능 함수 (PUF)

- A function that is embodied in a physical structures and is easy to evaluate but hard to clone

Physically unclonable function (PUF)

odied in a physical structures and is easy to evaluate but hard to clone



물리적 복제 불가능 함수(PUF)

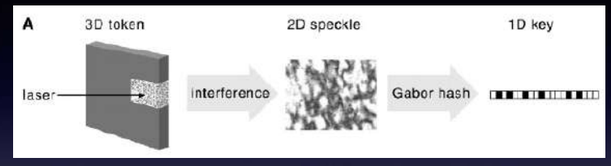
- 위조 방지 시스템을 위한 효과적인 도구
- 챌린지-응답 쌍 (CRP)
- 챌린지: 구조에 가해진 물리적 자극
- 응답: PUF의 반응
- PUF의 출력: 비트 문자열
- PUF 식별: 거리 측정 (해밍 거리) 인터-거리 특정 챌린지를 두 개의 다른 PUF에 적용했을 때 발생하는 두 응답 간의 거리
- μinter: 고유성 (비트 길이의 50% = 가장 잘 구별됨) 인트라-거리

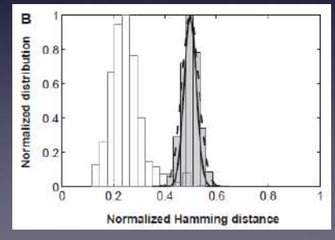
Physically unclonable function (PUF)

- An effective tool for anti-counterfeiting system
- Challenge-response pair (CRP)
  - Challenge: physical stimulus applied to the structure
  - Response: reaction of the PUF
- Extracted output of PUF: bit string
- Identification of PUF: distance measure (Hamming distance)
- Inter-distance
  - Distance between the two responses resulting from applying a particular challenge to two different PUFs
  - μ_{inter}: uniqueness (50% of bit length = best distinguishable)
- Intra-distance
  - Distance between the two responses resulting from applying a particular challenge twice to one PUF
  - μ_{intra}: robustness (small = reliable PUF response)

광학 PUF

- Coherent multiple scattering from inhomogeneous structures



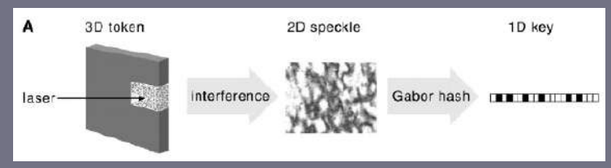


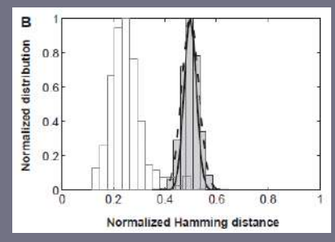
- CRP
  - challenge: beam angle
  - response: speckle pattern
- Data set (576 keys)
  - 4 tokens
  - 144 angles

R. Pappu et al., *Science*, **297** (2002)

Optical PUF

- Coherent multiple scattering from inhomogeneous structures



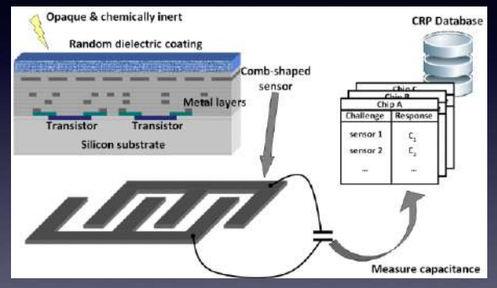


- CRP
  - challenge: beam angle
  - response: speckle pattern
- Data set (576 keys)
  - 4 tokens
  - 144 angles

R. Pappu et al., *Science*, **297** (2002)

코팅 PUF

- Randomness of capacitance in comb-shaped sensors in IC

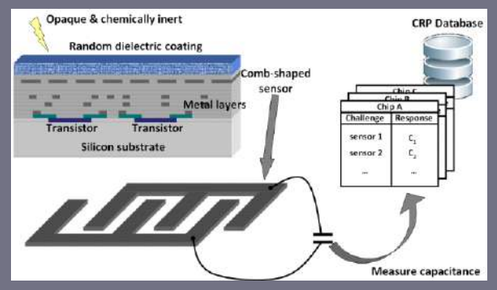


- Coating layer
  - matrix (aluminophosphate)
  - random dielectric particles (TiO_2, TiN)
- CRP
  - challenge: sensor
  - Response: capacitance
- Data set
  - 36 chips
  - 31 sensors

R. Maes and I. Verbauwhede, *Towards Hardware-Intrinsic Security Information Security and Cryptography* (2010)
P. Tuyls et al., *Cryptogrpahic Hardware and Embedded Systems Workshop* (2006)

Coating PUF

- Randomness of capacitance in comb-shaped sensors in IC

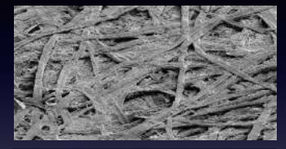


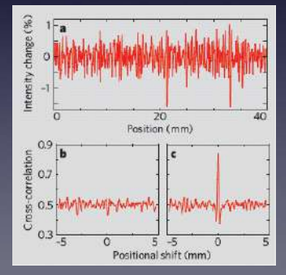
- Coating layer
  - matrix (aluminophosphate)
  - random dielectric particles (TiO_2, TiN)
- CRP
  - challenge: sensor
  - Response: capacitance
- Data set
  - 36 chips
  - 31 sensors

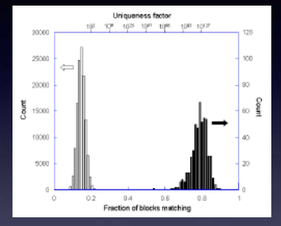
R. Maes and I. Verbauwhede, *Towards Hardware-Intrinsic Security Information Security and Cryptography* (2010)
P. Tuyls et al., *Cryptogrpahic Hardware and Embedded Systems Workshop* (2006)

종이 PUF

- Unique surface imperfections in documents





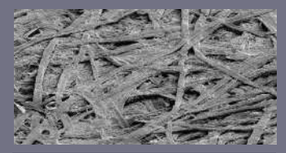


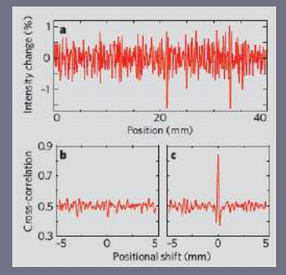
- CRP
  - challenge: laser beam
  - response: speckle pattern
- Data set
  - 500 sheets

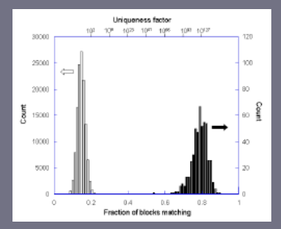
J. D. R. Buchanan et al., *Nature*, **436** (2005)

Paper PUF

- Unique surface imperfections in documents





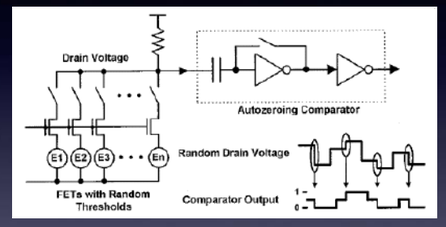


- CRP
  - challenge: laser beam
  - response: speckle pattern
- Data set
  - 500 sheets

J. D. R. Buchanan et al., *Nature*, **436** (2005)

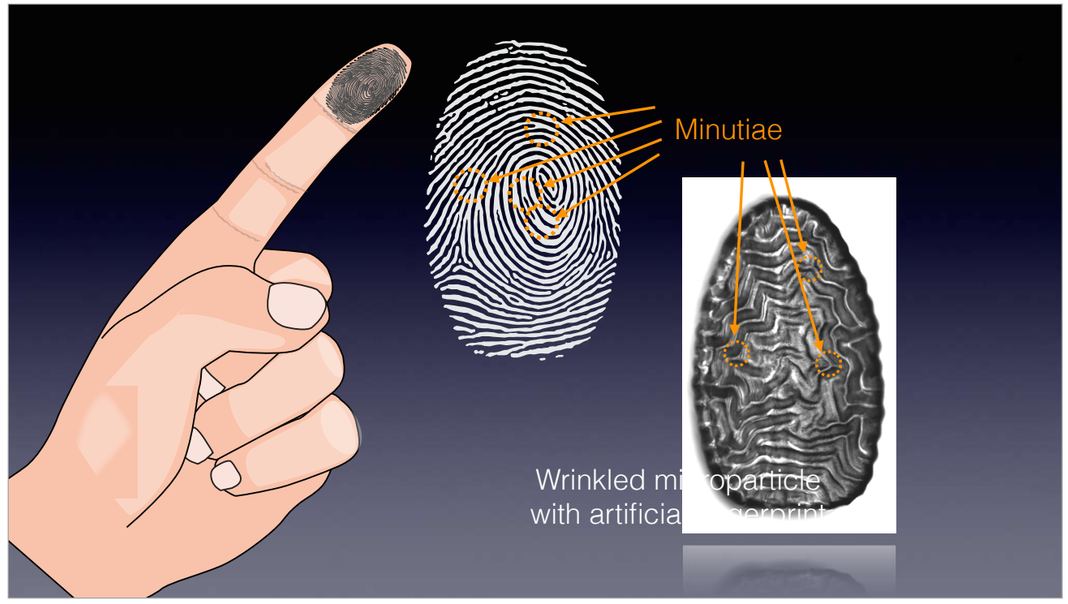
실리콘 PUF

- Manufacturing process variations in integrated circuit (IC) design



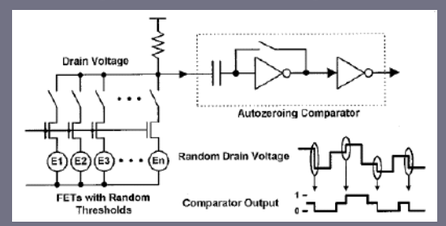
- CRP
  - challenge: MOSFET block
  - response: load voltages
- Data set
  - 55 chips
  - 132 blocks

R. Maes and I. Verbauwhede, *Towards Hardware-Intrinsic Security Information Security and Cryptography* (2010)
K. Losfstrom et al., *In proceeding of ISSCC 2000* (2000)



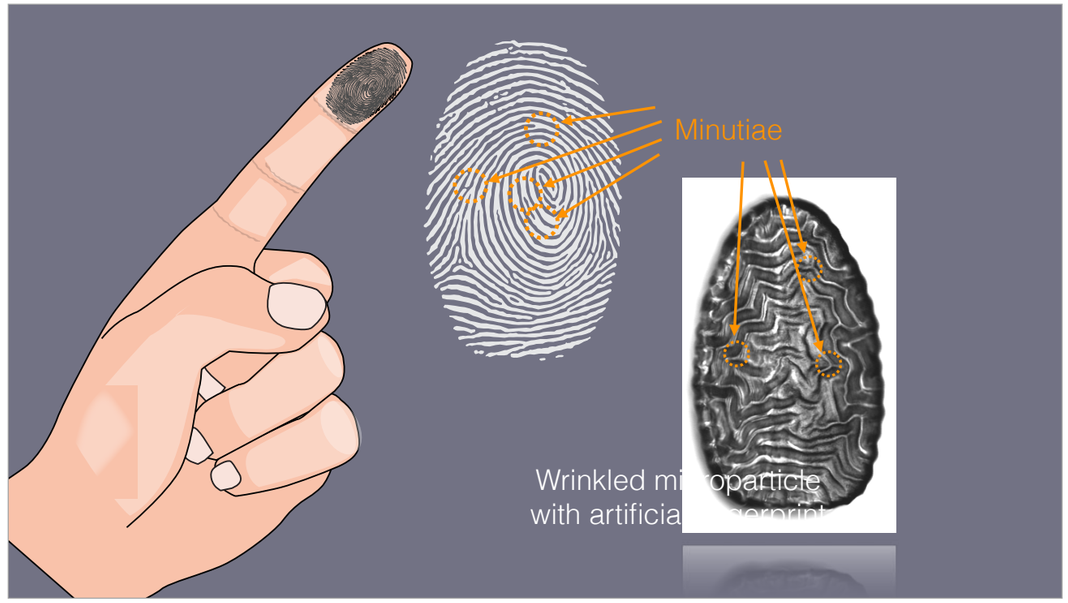
Silicon PUF

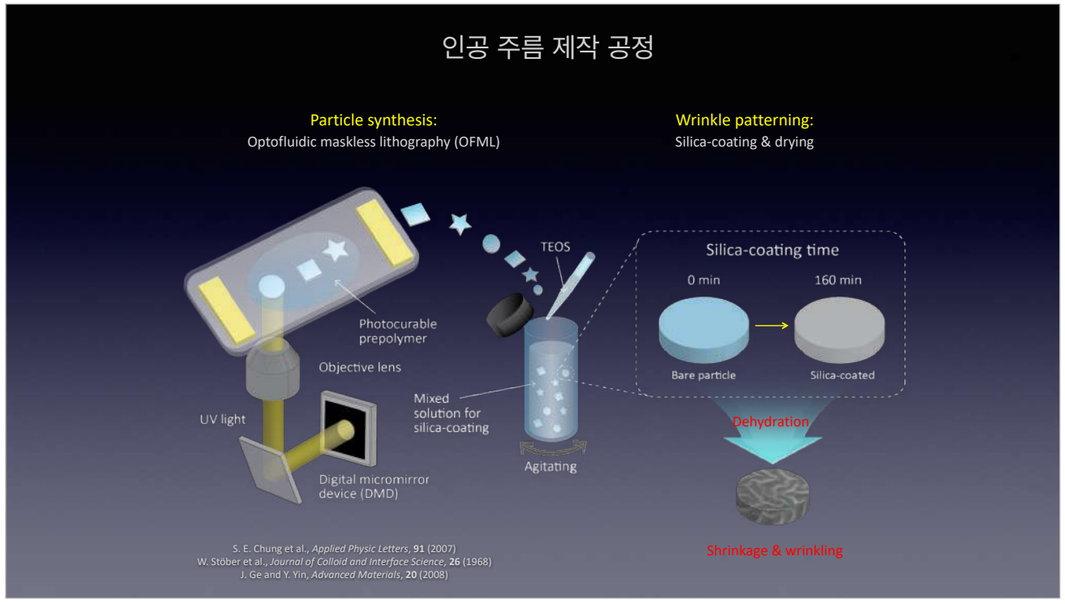
- Manufacturing process variations in integrated circuit (IC) design

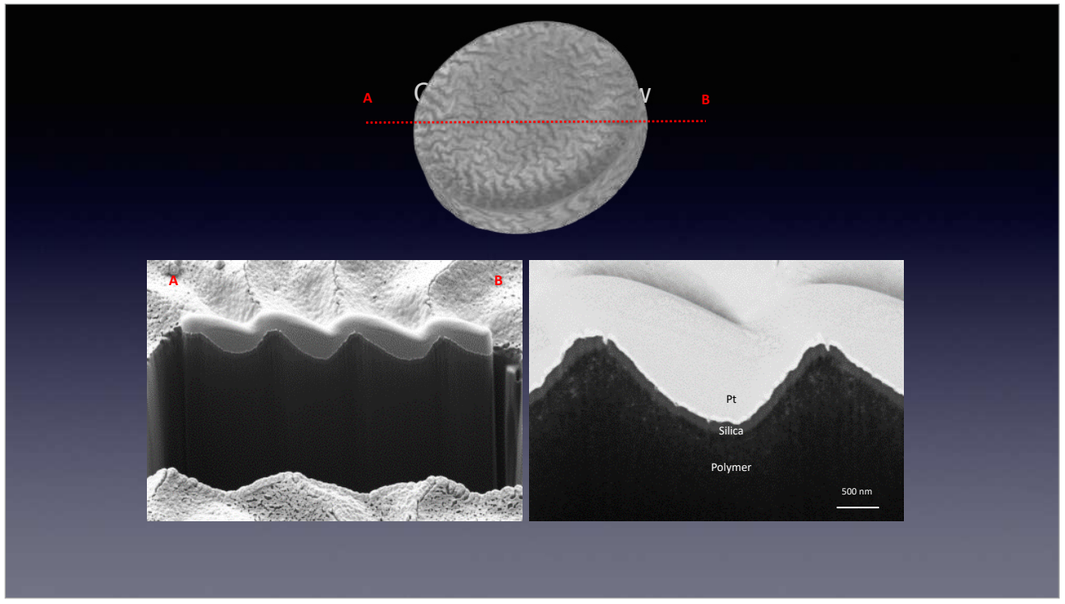


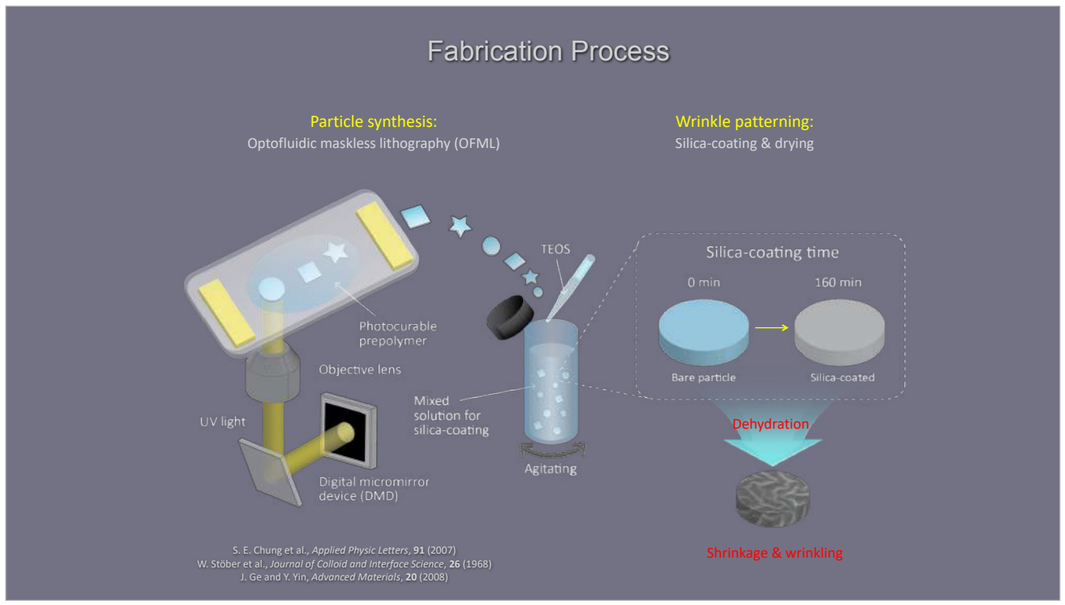
- CRP
  - challenge: MOSFET block
  - response: load voltages
- Data set
  - 55 chips
  - 132 blocks

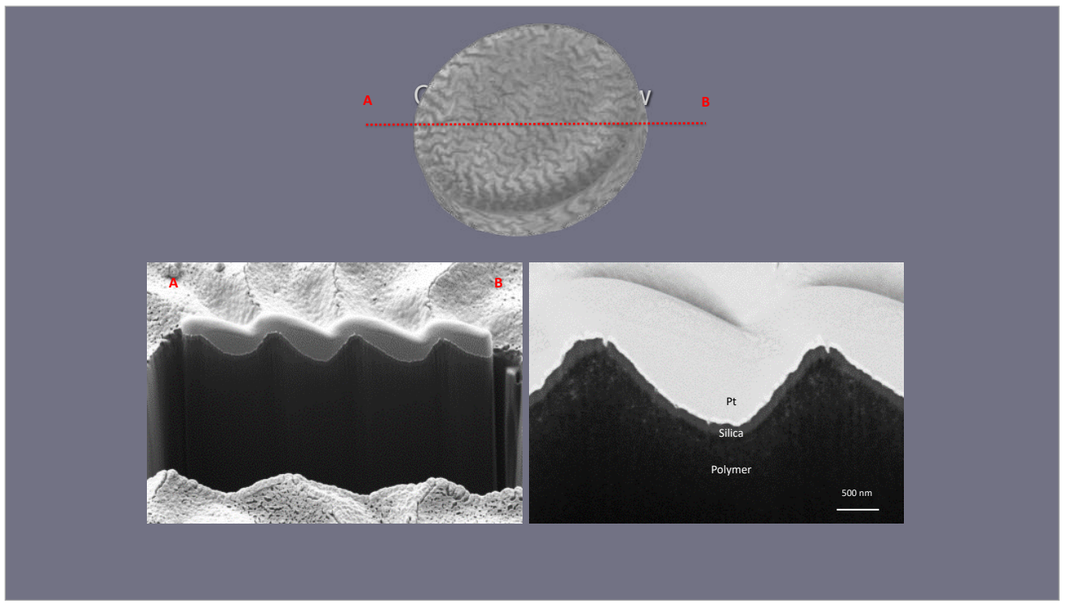
R. Maes and I. Verbauwhede, *Towards Hardware-Intrinsic Security Information Security and Cryptography* (2010)
K. Losfstrom et al., *In proceeding of ISSCC 2000* (2000)



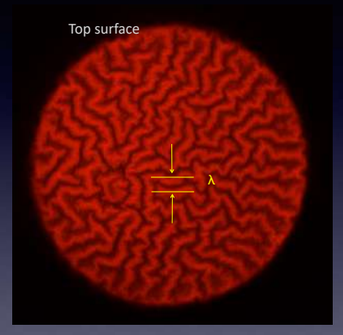

인공 주름 제작 공정
Particle synthesis:
Optofluidic maskless lithography (OFML)
Wrinkle patterning:
Silica-coating & drying
TEOS
Photocurable prepolymer
Objective lens
UV light
Mixed solution for silica-coating
Digital micromirror device (DMD)
Agitating
Silica-coating time
0 min
160 min
Bare particle
Silica-coated
Dehydration
Shrinkage & wrinkling
S. E. Chung et al., Applied Physic Letters, 91 (2007)
W. Stöber et al., Journal of Colloid and Interface Science, 26 (1968)
J. Ge and Y. Yin, Advanced Materials, 20 (2008)


A
B
A
B
Pt
Silica
Polymer
500 nm


Fabrication Process
Particle synthesis:
Optofluidic maskless lithography (OFML)
Wrinkle patterning:
Silica-coating & drying
TEOS
Photocurable prepolymer
Objective lens
UV light
Mixed solution for silica-coating
Digital micromirror device (DMD)
Agitating
Silica-coating time
0 min
160 min
Bare particle
Silica-coated
Dehydration
Shrinkage & wrinkling
S. E. Chung et al., Applied Physic Letters, 91 (2007)
W. Stöber et al., Journal of Colloid and Interface Science, 26 (1968)
J. Ge and Y. Yin, Advanced Materials, 20 (2008)


A
B
A
B
Pt
Silica
Polymer
500 nm

Wavelength Control



E_s: Elastic modulus of substrate (=polymer)
E_f: Elastic modulus of film (=silica layer)
v_s: Poisson ratio of substrate
v_f: Poisson ratio of film
t :Thickness of film

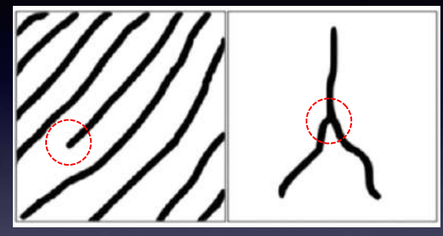
$$\lambda_c = 2\pi t \left[\frac{(1 - v_s^2) E_f}{3(1 - v_f^2) E_s} \right]^{1/3}$$

1. Coating time↑ → t↑ → λ↑
2. UV light dose↑ → E_s↑ → λ↓

J. Genzer and J. Groenewold, *Soft Matter*, **2** (2006)

특이점

- Major features of a fingerprint



Ridge ending Ridge bifurcation

No two fingers have identical minutiae patterns!

Wavelength Control

Top surface
λ

E_s: Elastic modulus of substrate (=polymer)
E_f: Elastic modulus of film (=silica layer)
v_s: Poisson ratio of substrate
v_f: Poisson ratio of film
t :Thickness of film

$$\lambda_c = 2\pi t \left[\frac{(1 - v_s^2) E_f}{3(1 - v_f^2) E_s} \right]^{1/3}$$

1. Coating time↑ → t↑ → λ↑
2. UV light dose↑ → E_s↑ → λ↓

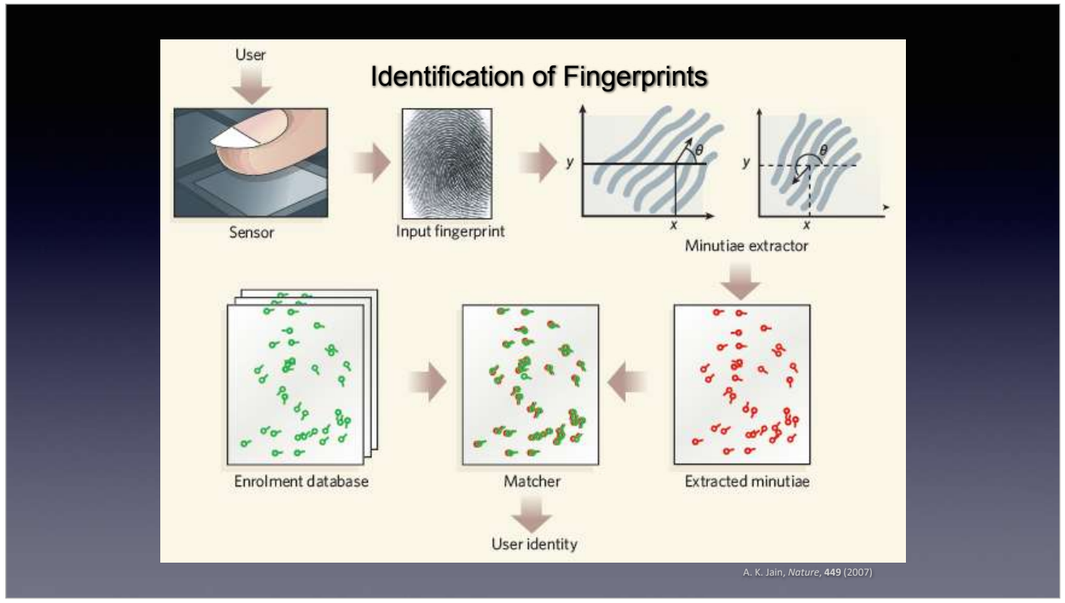
J. Genzer and J. Groenewold, *Soft Matter*, **2** (2006)

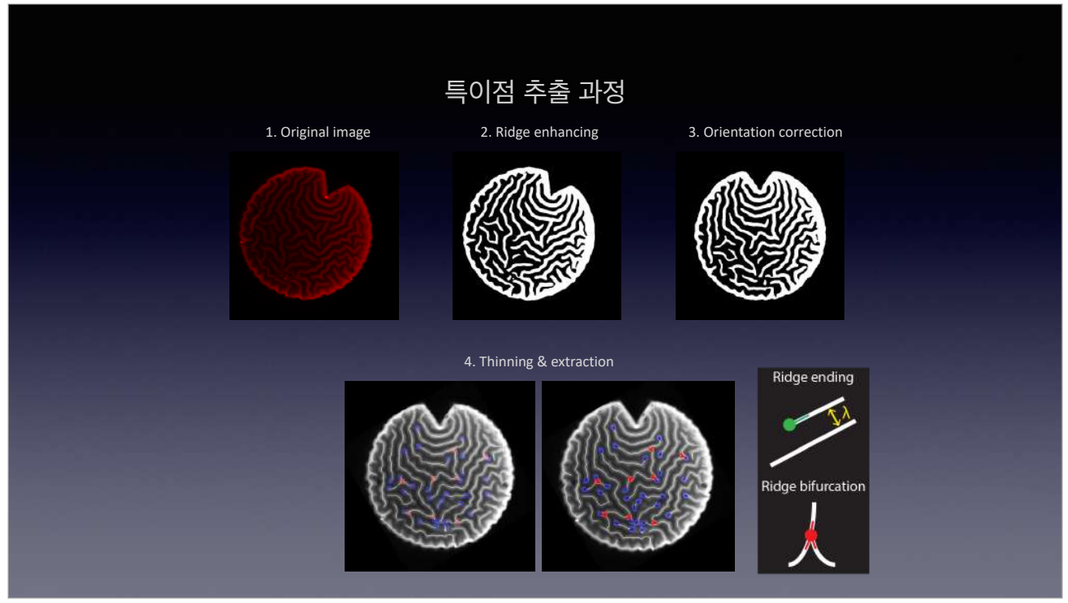
Minutiae

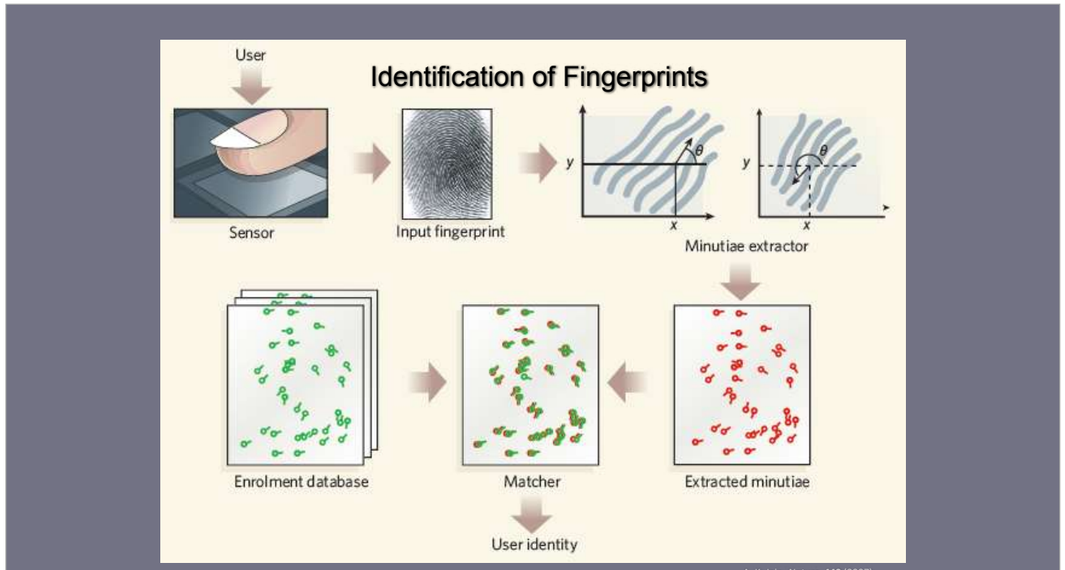
- Major features of a fingerprint

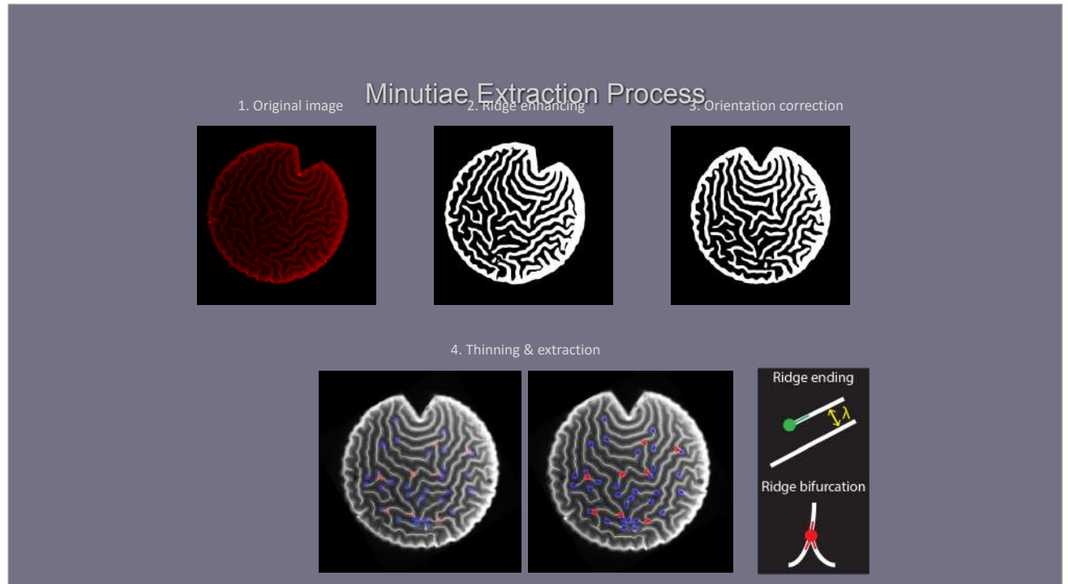
Ridge ending Ridge bifurcation

No two fingers have identical minutiae patterns!


User
Identification of Fingerprints
Sensor
Input fingerprint
Minutiae extractor
Enrolment database
Matcher
Extracted minutiae
User identity
A. K. Jain, Nature, 449 (2007)

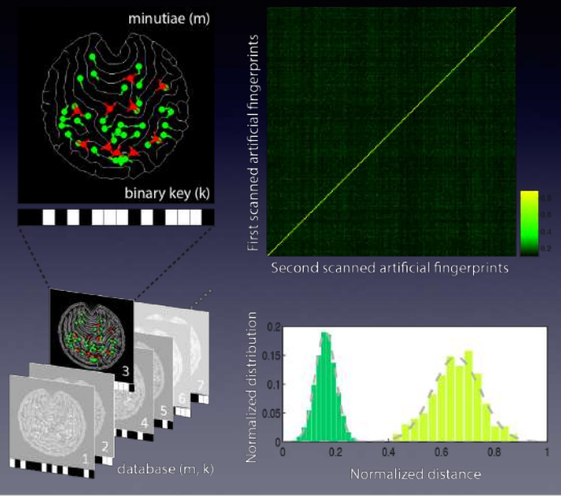

특이점 추출 과정
1. Original image
2. Ridge enhancing
3. Orientation correction
4. Thinning & extraction
Ridge ending
Ridge bifurcation


User
Identification of Fingerprints
Sensor
Input fingerprint
Minutiae extractor
Enrolment database
Matcher
Extracted minutiae
User identity
A. K. Jain, Nature, 449 (2007)

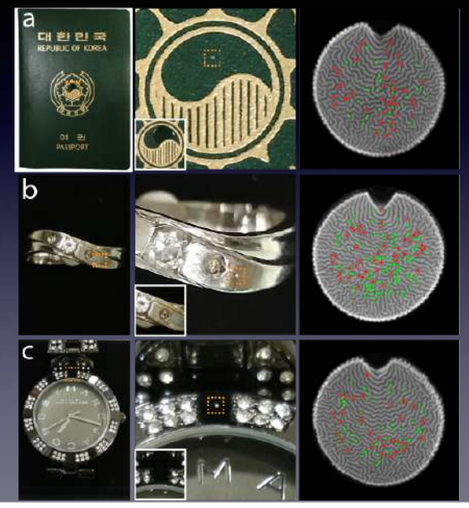

Minutiae Extraction Process
1. Original image
2. Ridge enhancing
3. Orientation correction
4. Thinning & extraction
Ridge ending
Ridge bifurcation

특이점의 유일성 검증 과정

- Cross-correlation of 350 particle images

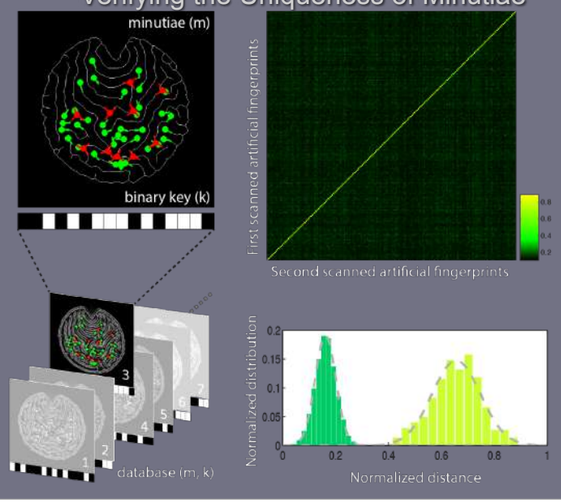


인증 응용

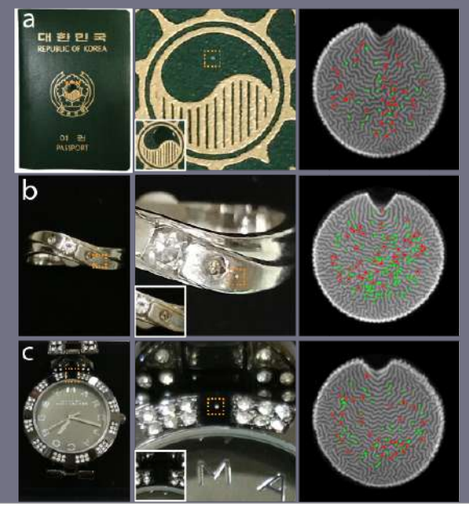


Verifying the Uniqueness of Minutiae

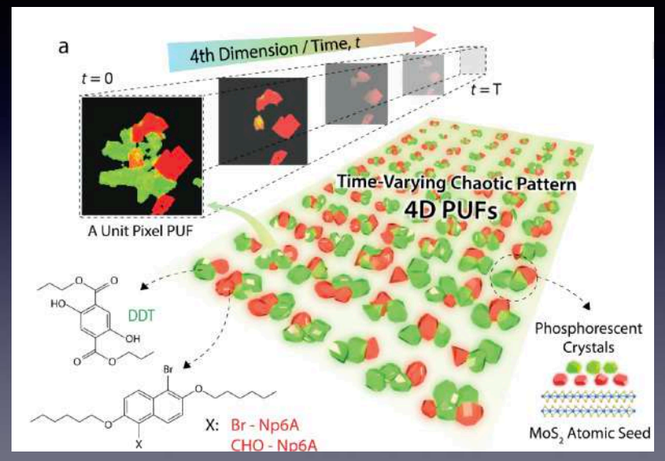
- Cross-correlation of 350 particle images



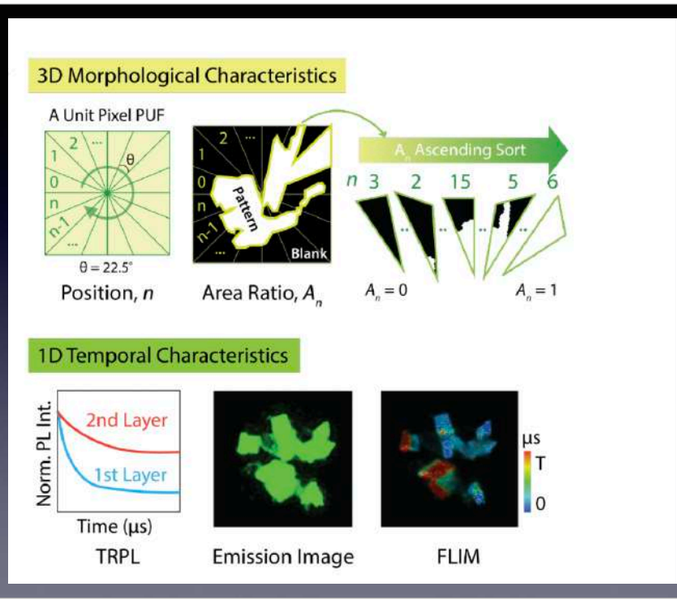
Authentication examples



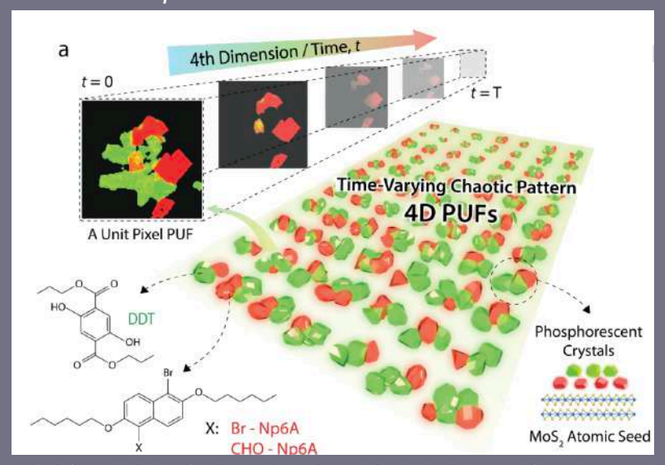
4차원 물리적 복제 불가능 함수와 시간에 따라 변하는 혼돈의 인광 패턴에 기반한 암호화 응용 및 인증 예제



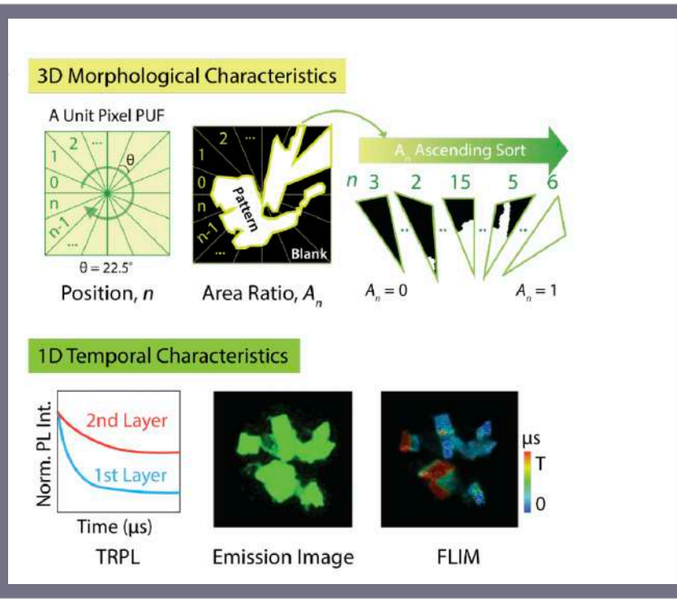
Im, H., Yoon, J., So, B., Choi, J., Park, D. H., Kim, S., & Park, W. (2024). Four-Dimensional Physical Unclonable Functions and Cryptographic Applications Based on Time-Varying Chaotic Phosphorescent Patterns. *ACS Nano*, *18*(18), 11703–11716.

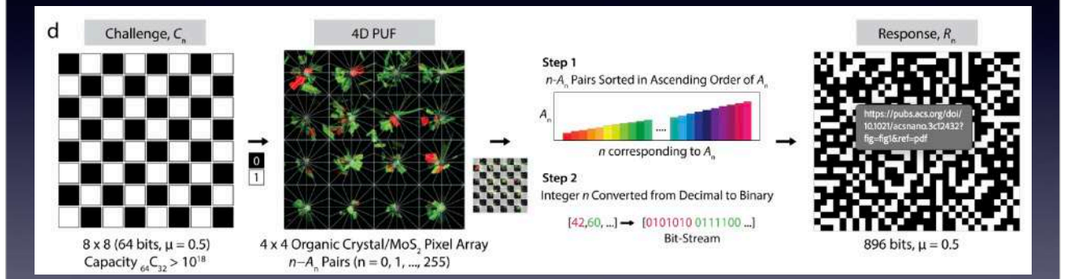


Four-Dimensional Physical Unclonable Functions and Cryptographic Applications Based on Time-Varying Chaotic Phosphorescent Patterns Authentication examples

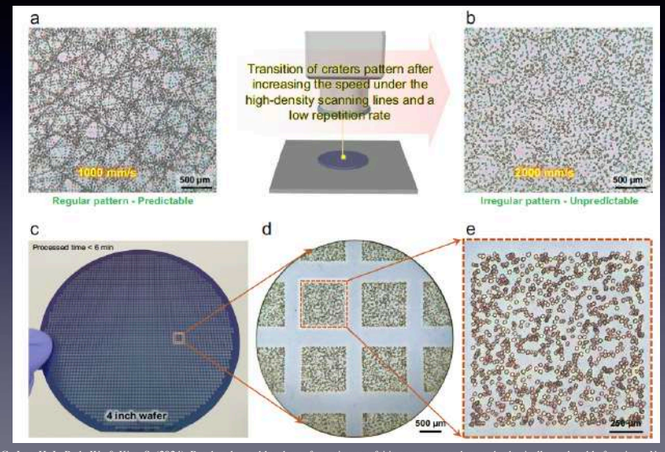


Im, H., Yoon, J., So, B., Choi, J., Park, D. H., Kim, S., & Park, W. (2024). Four-Dimensional Physical Unclonable Functions and Cryptographic Applications Based on Time-Varying Chaotic Phosphorescent Patterns. *ACS Nano*, *18*(18), 11703–11716.

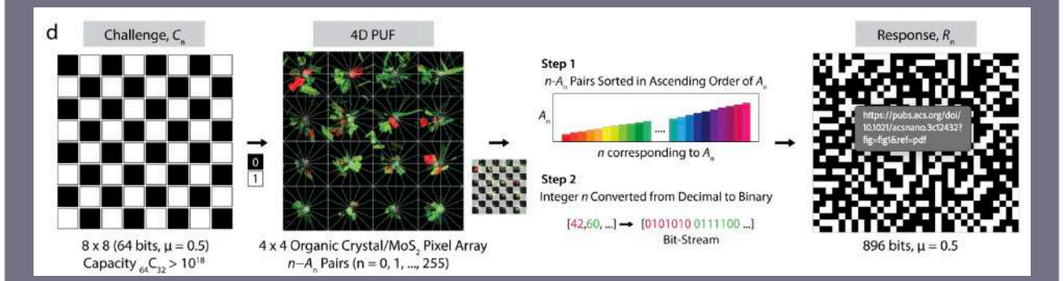




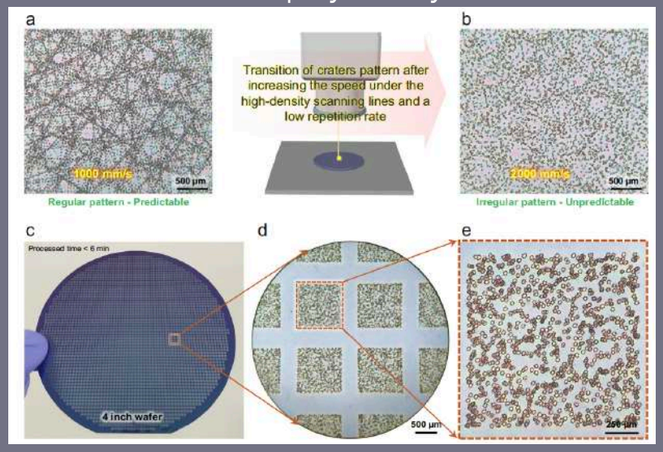
위조 방지 목적과 물리적 복제 불가능 함수로의 응용을 위한 랜덤 레이저 어블레이션 태그



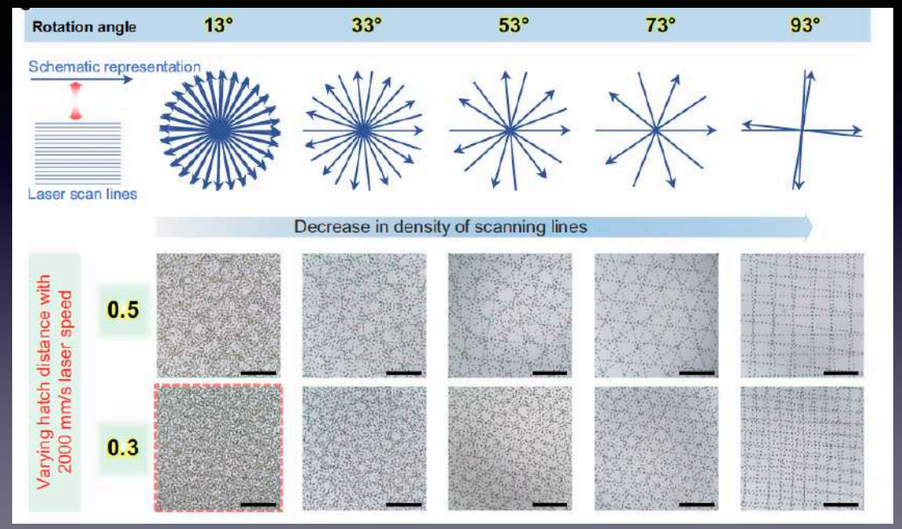
Gandla, S., Yoon, J., Yang, C., Lee, H. J., Park, W., & Kim, S. (2024). Random laser ablated tags for anticounterfeiting purposes and towards physically unclonable functions. *Nature Communications* , *15*(1).

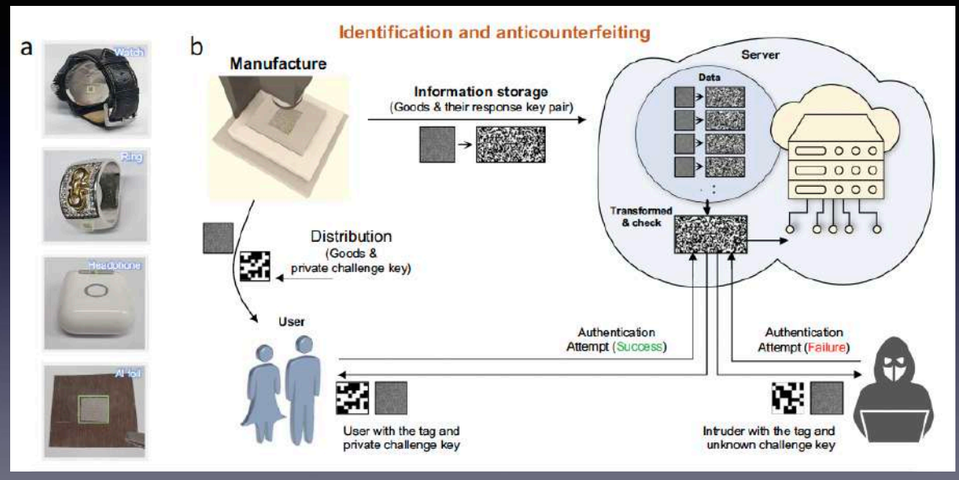


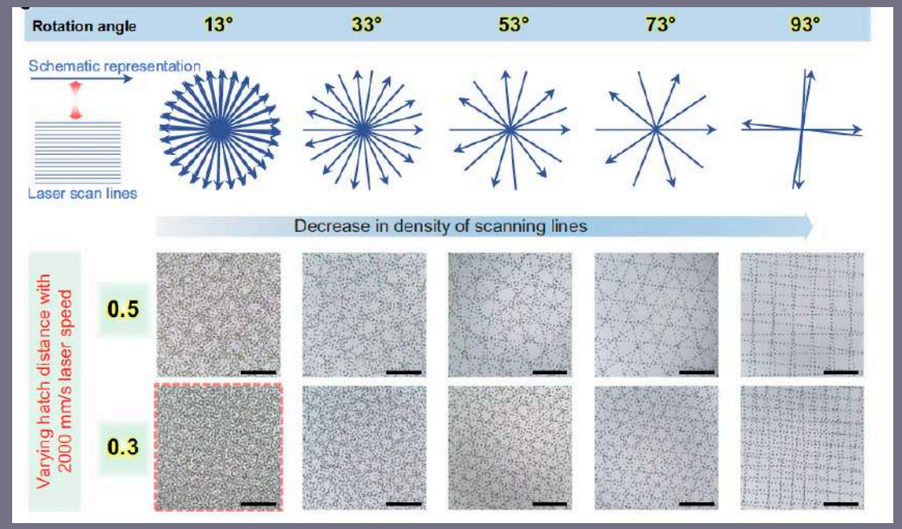
Random laser ablated tags for anticounterfeiting purposes and towards physically unclonable functions

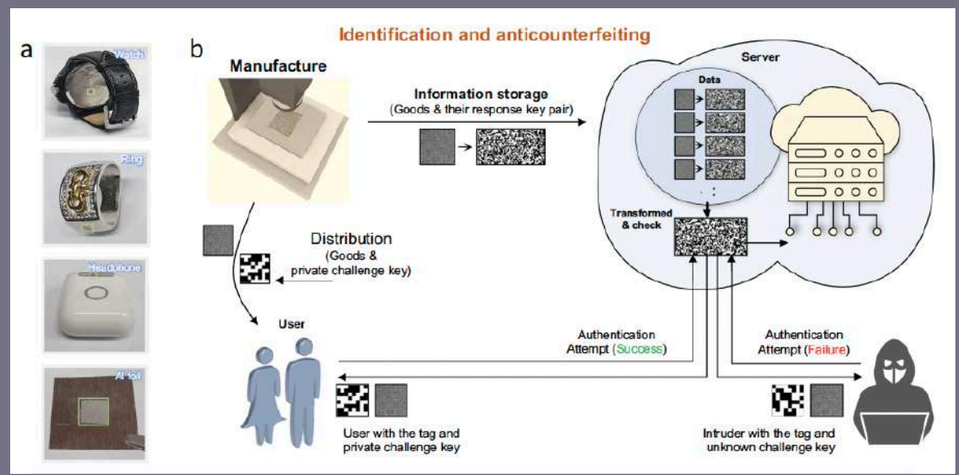


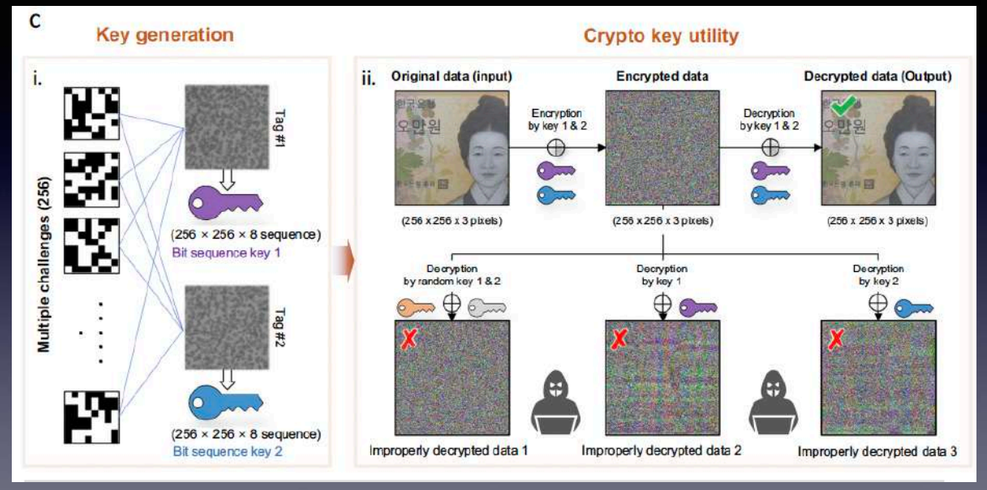
Gandla, S., Yoon, J., Yang, C., Lee, H. J., Park, W., & Kim, S. (2024). Random laser ablated tags for anticounterfeiting purposes and towards physically unclonable functions. *Nature Communications* , *15*(1).


Rotation angle
13°
33°
53°
73°
93°
Schematic representation
Laser scan lines
Decrease in density of scanning lines
Varying hatch distance with 2000 mm/s laser speed
0.5
0.3


a
b
Identification and anticounterfeiting
Manufacture
Information storage
(Goods & their response key pair)
Server
Data
Transformed & check
Distribution
(Goods & private challenge key)
User
Authentication Attempt (Success)
Authentication Attempt (Failure)
User with the tag and private challenge key
Intruder with the tag and unknown challenge key


Rotation angle
13°
33°
53°
73°
93°
Schematic representation
Laser scan lines
Decrease in density of scanning lines
Varying hatch distance with 2000 mm/s laser speed
0.5
0.3


a
b
Identification and anticounterfeiting
Manufacture
Information storage
(Goods & their response key pair)
Server
Data
Transformed & check
Distribution
(Goods & private challenge key)
User
Authentication Attempt (Success)
Authentication Attempt (Failure)
User with the tag and private challenge key
Intruder with the tag and unknown challenge key

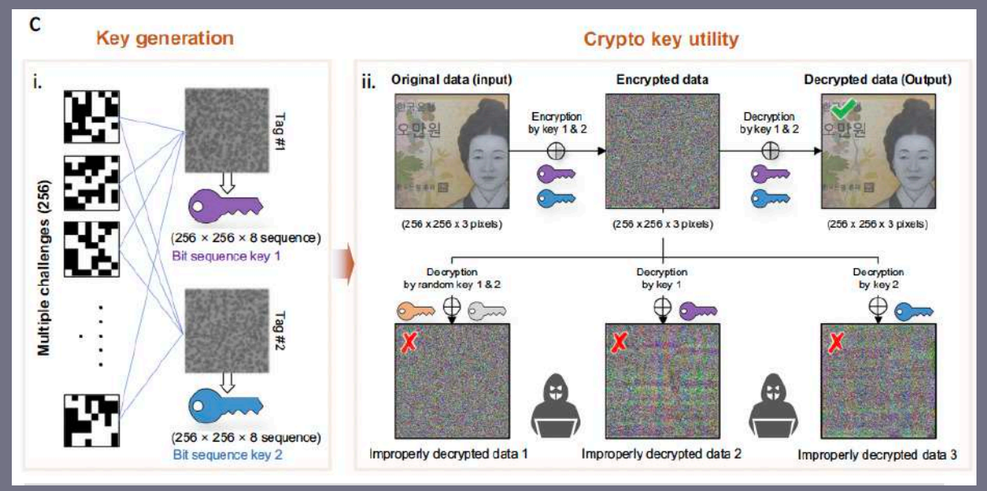


결론

4D PUF는 시간에 따라 변화하는 혼돈적 발산 패턴을 활용하여 IoT 장치 인증과 암호화 보안을 강화

레이저 어블레이션 기법은 비용 효율적이며 쉽게 검증 가능한 고유하고 무작위적인 위조 방지 태그를 생성

저작권 보호와 위조 방지를 위한 혁신적인 해결책을 제공



Conclusion

4D PUFs utilize time-varying, chaotic emission patterns for enhanced IoT device authentication and cryptographic security.

A laser ablation technique generates unique, random anti-counterfeiting tags that are cost-effective and easily verifiable.

• These technologies offer innovative solutions for copyright protection and counterfeiting deterrence.

Session 1 디지털 혁신 속 저작권 보호 기술

II 생성형 AI 시대의 콘텐츠 진위성

일케 데미르
인텔 선임 연구원

연사 이력

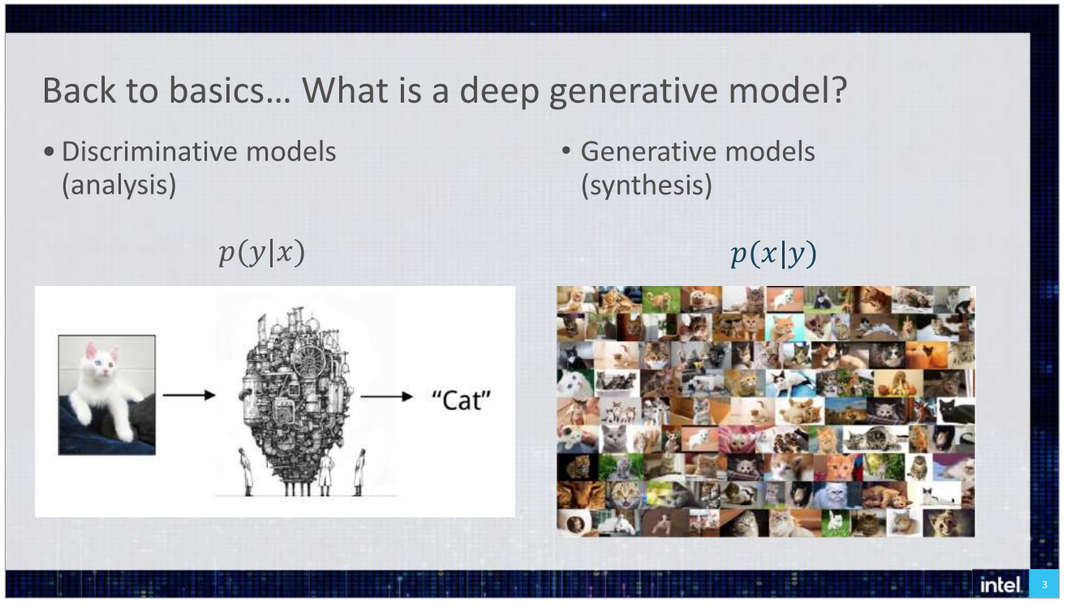
- Purdue University에서 컴퓨터 과학 박사 및 석사 학위 취득
- Middle East Technical University에서 컴퓨터 공학 학사 및 전기 공학 부전공 수료
- 현재 Intel Corporation에서 Senior Staff Research Scientist로 재직, 신뢰할 수 있는 미디어(Trusted Media) 이니셔티브를 이끌고 있음
- Facebook에서 박사후 연구원으로 근무하며 생성적 거리 주소 개발에 기여
- Pixar Animation Studios와 Intel Studios에서 애니메이션 및 VR/AR 영화에 기여
- Intel에서 세계 최대 볼륨 캡처 스튜디오의 연구 기반 구축
- 스타트업 경험(성공적인 인수)과 UCLA에서 방문 학자 경력
- National Academy of Engineering FOE 참가자, Jack Dangermond Award, Bilsland Dissertation Fellowship 등 다수의 수상 경력
- ACM Distinguished Speaker, IEEE Industry Distinguished Lecturer로 활동

발표 내용

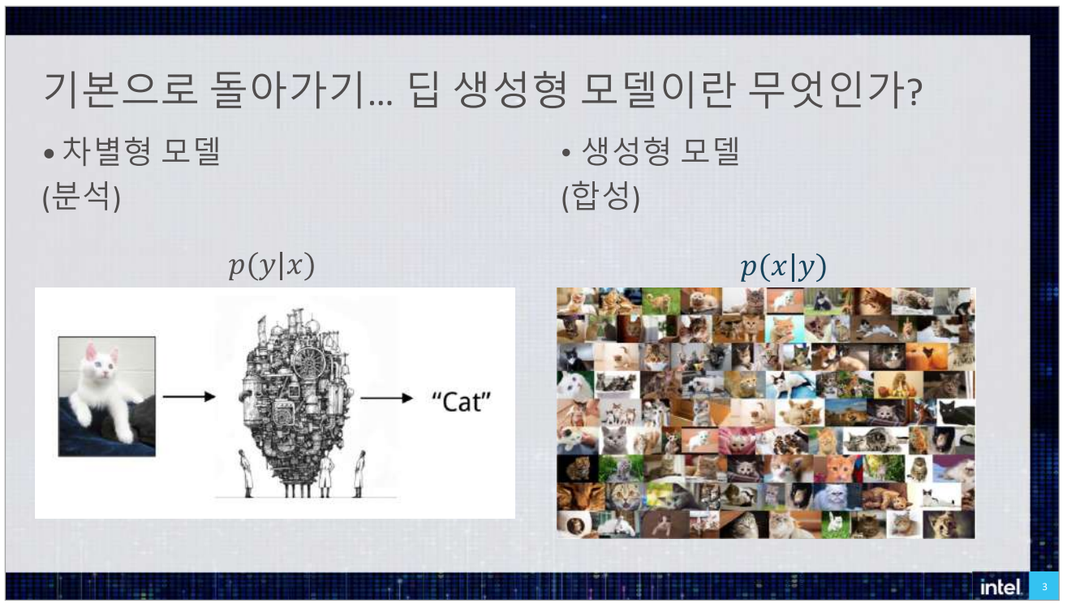
생성형 AI 접근법이 급격히 발전하면서, 우리가 온라인에서 보는 것과 듣는 것을 더 이상 신뢰할 수 없게 되어 디지털 콘텐츠에 대한 사회적 신뢰가 점차 무너지고 있습니다. 이들의 포토리얼리즘(사진 같은 사실성)은 주목을 받고 있지만, 동시에 사기, 윤리 문제, 신분 도용, 허위 정보와 관련된 심각한 우려도 제기되고 있습니다. 더 나아가, 이러한 모델들이 예술, 스타일, 유사성 등을 복제할 때, 진본과 인공 콘텐츠의 소유권 경계가 모호해지고 있습니다. 이번 강연에서 합성 콘텐츠 탐지, 책임 있는 생성형 AI, 미디어 출처 추적에 관한 혁신적인 연구를 통해 이 분야를 어떻게 변화시키고 있는지 공유할 것입니다. 이 연구는 모두를 위한 신뢰할 수 있는 온라인 미래를 만드는 세계적인 솔루션으로 꽃피울 것입니다.

Generative AI approaches are dramatically improving, silently causing social erosion of trust in digital content as we can no longer trust what we see or hear online. Their photorealistic prowess is drawing attention – but also raising grave concerns in frauds, ethics, impersonation, and misinformation. Furthermore, ownership of both authentic and synthetic content is getting blurry when these models replicate art, style, likeness, etc. In this talk, Dr. Demir will share how her team changes this landscape by their innovative research on synthetic content detection, responsible generative AI, and media provenance; blooming the research into world-changing solutions to create a trusted online future for everyone.

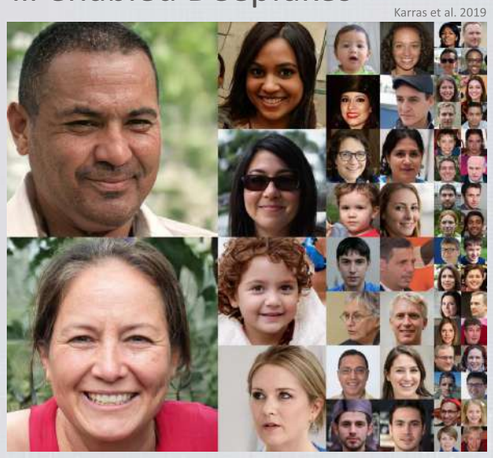

Content Authenticity
in the Age of Generative AI
Dr. Ilke Demir
Sr. Staff Research Scientist
Intel Labs
ICOTEC 2024
2


Back to basics… What is a deep generative model?
• Discriminative models (analysis)
• Generative models (synthesis)
$p(y|x)$
$p(x|y)$
"Cat"
intel
3


생성형 AI 시대의
콘텐츠 진위성
Dr. Ilke Demir
Sr. Staff Research Scientist
Intel Labs
ICOTEC 2024
2

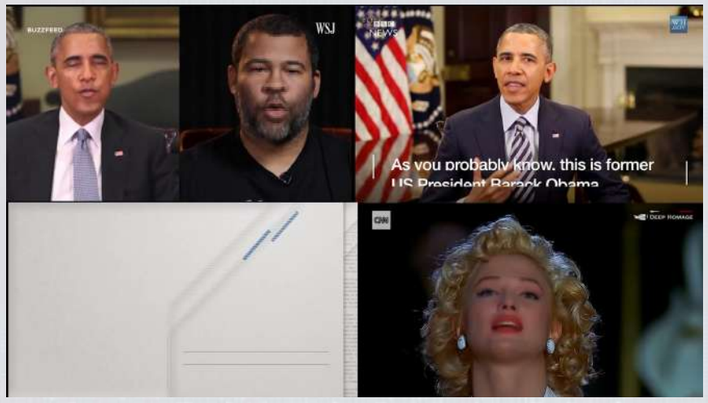

기본으로 돌아가기… 딥 생성형 모델이란 무엇인가?
• 차별형 모델 (분석)
• 생성형 모델 (합성)
$p(y|x)$
$p(x|y)$
"Cat"
intel
3

… enabled Deepfakes


Karras et al. 2019


Karras et al. 2020

Deepfake Dystopia

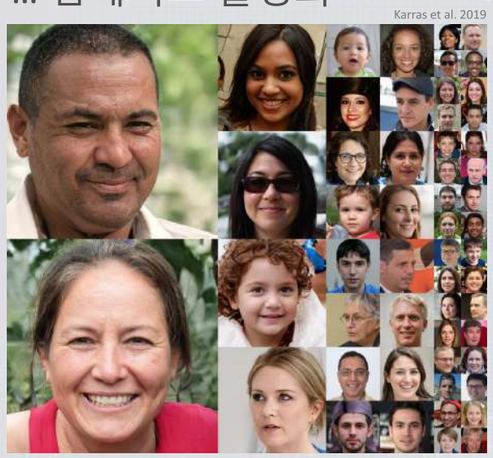


- **Political misinformation**
  - False coup attempt
- **Adult content**
  - Bot nudifies 680K women
  - Taylor Swift case
- **Impersonation & Forgery**
  - $243K by synthetic voice
  - $25M video call with CFO
- **Fake court evidences**
  - Fake cheerleaders -> real
  - Real threats -> fake
  - Fake audio in custody hearing

Social erosion of trust!

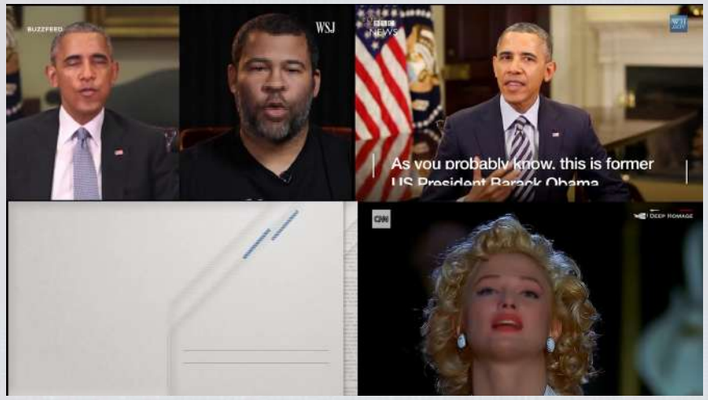
[*] D. Chu, **I. Demir**, K. Eichensehr, J. G. Foster, M. L. Green, K. Lerman, F. Menczer et al. "White Paper: DEEP FAKERY—An Action Plan.", IPAM – UCLA, Technical Report.

… 딥페이크 활성화


Karras et al. 2019


Karras et al. 2020

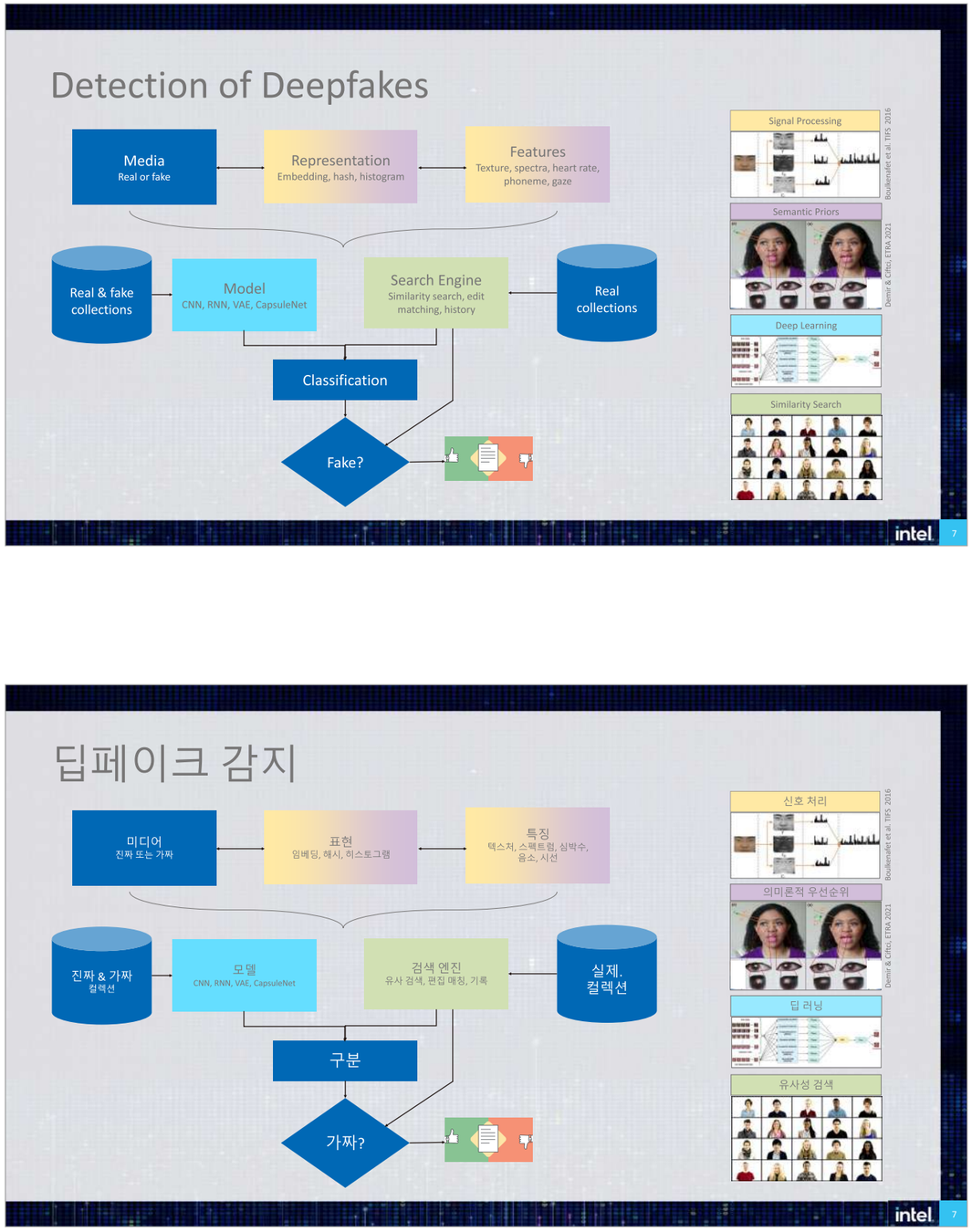
딥페이크 디스토피아

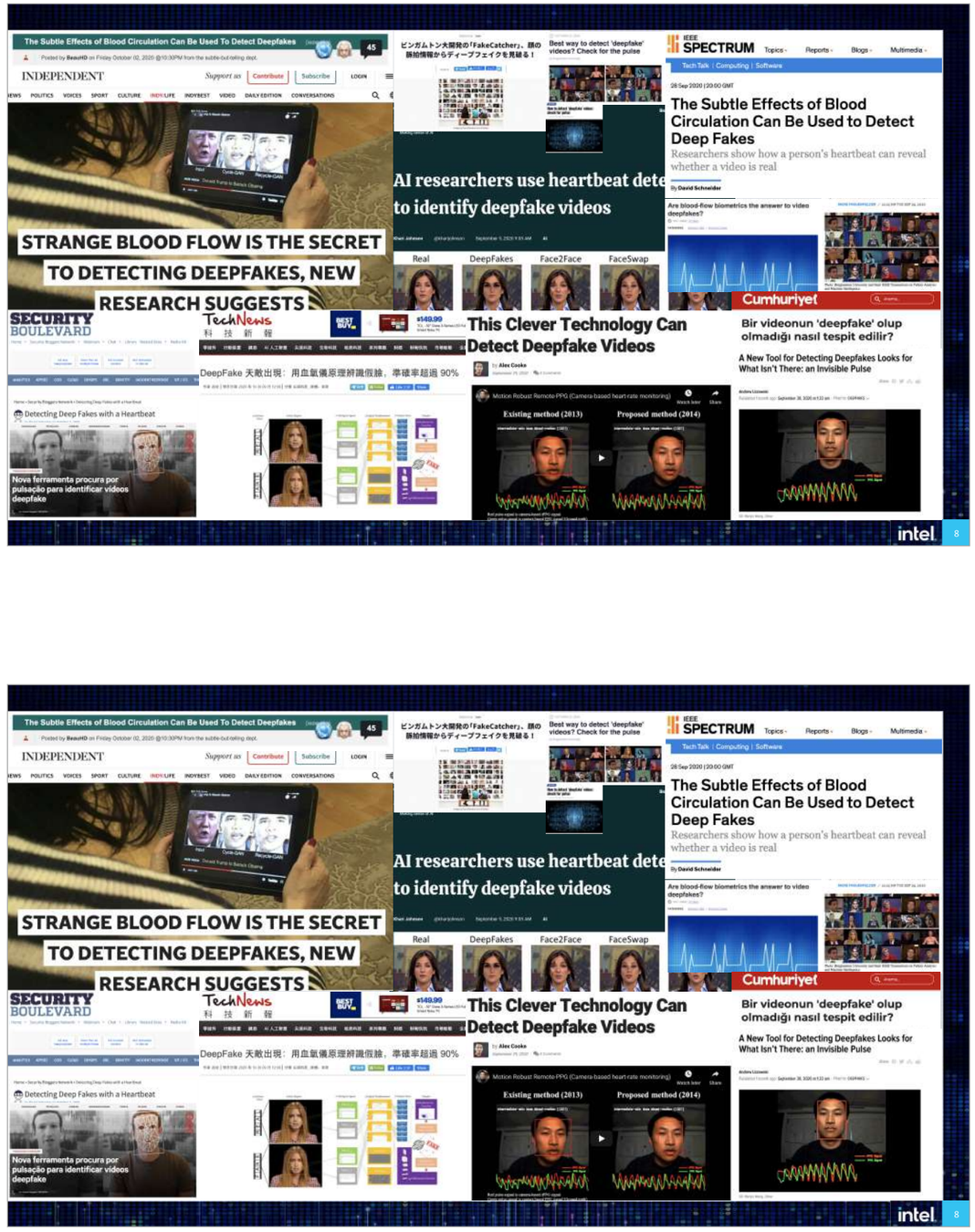


- **정치적 허위 정보**
  - 거짓 쿠데타 시도
- **성인 콘텐츠**
  - 봇이 68만 명의 여성을 누드화
  - 테일러 스위프트 사건
- **사칭 및 위조**
  - 합성 음성으로 $243K
  - CFO와 $25M 화상 통화
- **허위 법정 증거**
  - 가짜 응원단 -> 진짜
  - 실제 위협 -> 가짜
  - 구금 청문회에서의 가짜 녹음 파일

사회적 신뢰 침식!

[*] D. Chu, **I. Demir**, K. Eichensehr, J. G. Foster, M. L. Green, K. Lerman, F. Menczer et al. "White Paper: DEEP FAKERY—An Action Plan.", IPAM – UCLA, Technical Report.


Detection of Deepfakes
Media
Real or fake
Representation
Embedding, hash, histogram
Features
Texture, spectra, heart rate, phoneme, gaze
Real & fake collections
Model
CNN, RNN, VAE, CapsuleNet
Search Engine
Similarity search, edit matching, history
Real collections
Classification
Fake?
Signal Processing
Boulkenafet et al. IWFS - 2016
Semantic Priors
Demir & Ciftci, ETRA 2021
Deep Learning
Similarity Search
intel
7


The Subtle Effects of Blood Circulation Can Be Used To Detect Deepfakes
INDEPENDENT
STRANGE BLOOD FLOW IS THE SECRET
TO DETECTING DEEPFAKES, NEW
RESEARCH SUGGESTS
AI researchers use heartbeat dete
to identify deepfake videos
Real
DeepFakes
Face2Face
FaceSwap
IEEE SPECTRUM
The Subtle Effects of Blood Circulation Can Be Used to Detect Deep Fakes
Researchers show how a person's heartbeat can reveal whether a video is real
SECURITY BOULEVARD
TechNews
This Clever Technology Can Detect Deepfake Videos
Cumhuriyet
Bir videonun 'deepfake' olup olmadığı nasıl tespit edilir?
A New Tool for Detecting Deepfakes Looks for What Isn't There: an Invisible Pulse
Detecting Deep Fakes with a Heartbeat
Existing method (2013)
Proposed method (2014)
intel
8

딥페이크 감지
미디어
진짜 또는 가짜
표현
임베딩, 해시, 히스토그램
특징
텍스처, 스펙트럼, 심박수, 음소, 시선
진짜 & 가짜
컬렉션
모델
CNN, RNN, VAE, CapsuleNet
검색 엔진
유사 검색, 편집 매칭, 기록
실제.
컬렉션
구분
가짜?
신호 처리
Boulkenafet et al. IWFS - 2016
의미론적 우선순위
Demir & Ciftci, ETRA 2021
딥 러닝
유사성 검색
intel
7

The Subtle Effects of Blood Circulation Can Be Used To Detect Deepfakes
INDEPENDENT
STRANGE BLOOD FLOW IS THE SECRET
TO DETECTING DEEPFAKES, NEW
RESEARCH SUGGESTS
AI researchers use heartbeat dete
to identify deepfake videos
Real
DeepFakes
Face2Face
FaceSwap
IEEE SPECTRUM
The Subtle Effects of Blood Circulation Can Be Used to Detect Deep Fakes
Researchers show how a person's heartbeat can reveal whether a video is real
SECURITY BOULEVARD
TechNews
This Clever Technology Can Detect Deepfake Videos
Cumhuriyet
Bir videonun 'deepfake' olup olmadığı nasıl tespit edilir?
A New Tool for Detecting Deepfakes Looks for What Isn't There: an Invisible Pulse
Detecting Deep Fakes with a Heartbeat
Existing method (2013)
Proposed method (2014)
intel
8

FakeCatcher: Intuition

- Most detection approaches: Fakeness should have its faults
- Our approach: Is there a unique authenticity signature in real videos?

FakeCatcher: 직감

- 대부분의 탐지 접근 방식: 가짜는 결함이 있기 마련이다
- 우리의 접근 방식: 실제 영상에 고유한 진정성이 존재하는가?

FakeCatcher: Intuition

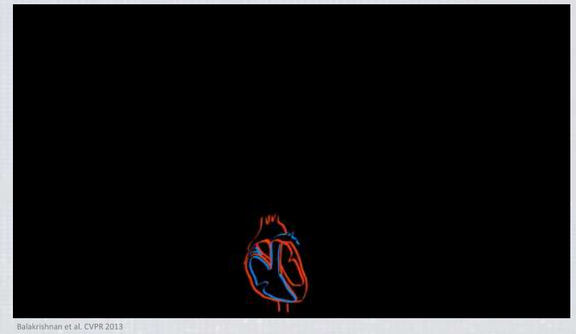
- Most detection approaches: Fakeness should have its faults
- Our approach: Is there a unique authenticity signature in real videos?

YES!

Blood flow creates subtle changes
- Invisible to the eye
- Visible computationally

Photoplethysmography (PPG): measuring blood pressure via skin color change

No generative model can create deepfakes with consistent PPG signals (yet)



Balakrishnan et al. CVPR 2013

FakeCatcher: 직관

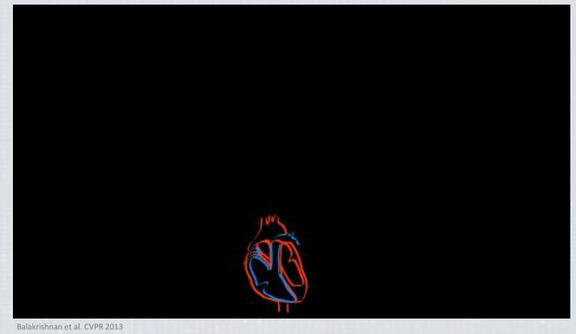
- 대부분의 탐지 접근 방식: 가짜는 단점이 있기 마련이다
- 우리의 접근 방식: 실제 영상에 고유한 진정성이 존재하는가?

그렇다!

혈류는 미묘한 변화 생성
- 육안으로 관찰 불가
- 컴퓨터로 관찰 가능

광전혈류측정법(PPG): 피부색 변화를 통해 혈압 측정

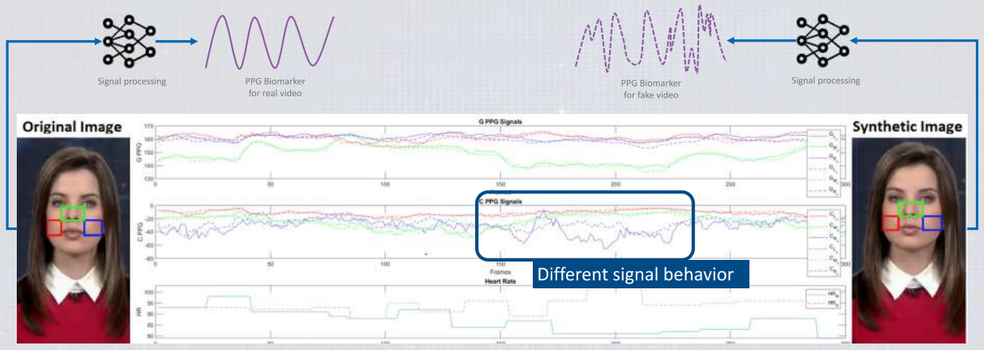
일관된 PPG 신호로 딥페이크를 생성할 수 있는 생성 모델은 아직 없음



Balakrishnan et al. CVPR 2013

FakeCatcher: Pairwise Separation

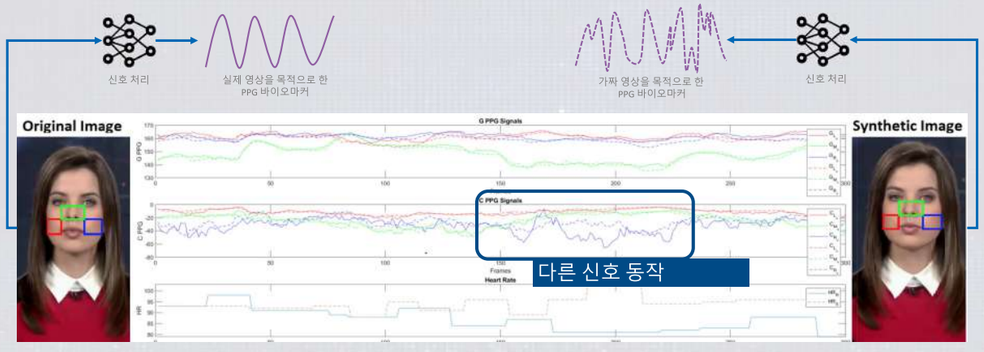
- Given N pairs of fake and real videos, can we find an implicit indicator of biological signals?



[*] U. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," in IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 2020

FakeCatcher: 쌍별 분리

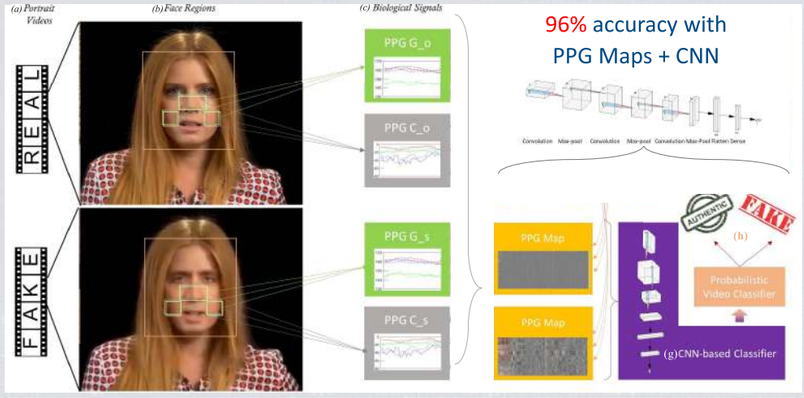
- N쌍의 가짜 및 진짜 비디오가 제공될 경우, 생물학적 신호의 암시적 지표를 발견할 수 있는가?



[*] U. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," in IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 2020

FakeCatcher: Generalization by Deep Learning

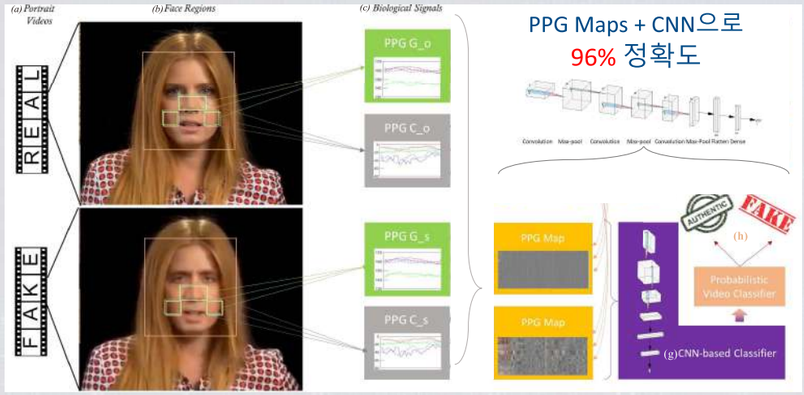
- Given N pairs of fake and real videos, can we learn the space of PPG features?



[*] U. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," in IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 2020

FakeCatcher: 딥러닝을 통한 일반화

- N 쌍의 가짜 비디오와 진짜 비디오가 주어질 시 PPG 특징의 공간을 배울 수 있는가?



[*] U. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," in IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 2020

FakeCatcher: Results

Train	Test	Video Accuracy
FF++ - Face2Face	Face2Face [37]	95.25%
FF++ - FaceSwap	FaceSwap [59]	96.25%
FF++ - Deepfakes	Deepfakes [57]	93.75%
FF++ - NeuralTextures	NeuralTextures [77]	81.25%

Cross-model results

Train	Test	Video Accuracy
Celeb-DF [4]	FF++ [3]	83.10%
FF++ [3]	Celeb-DF [4]	86.48%
FF++ [3]	Deep Fakes Dataset	84.51%
Celeb-DF [4]	Deep Fakes Dataset	82.39%
Deep Fakes	FF [2]	86.34%
FF [2]	Deep Fakes Dataset	67.61%
FF++ [3]	UADFV [56]	97.92%
Deep Fakes Dataset	FF++ [3]	80.60%
Deep Fakes Dataset	Celeb-DF [4]	85.13%

Cross-dataset results

[*] U. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," in IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 2020

FakeCatcher: 결과

Train	Test	Video Accuracy
FF++ - Face2Face	Face2Face [37]	95.25%
FF++ - FaceSwap	FaceSwap [59]	96.25%
FF++ - Deepfakes	Deepfakes [57]	93.75%
FF++ - NeuralTextures	NeuralTextures [77]	81.25%

교차 모델 결과

Train	Test	Video Accuracy
Celeb-DF [4]	FF++ [3]	83.10%
FF++ [3]	Celeb-DF [4]	86.48%
FF++ [3]	Deep Fakes Dataset	84.51%
Celeb-DF [4]	Deep Fakes Dataset	82.39%
Deep Fakes	FF [2]	86.34%
FF [2]	Deep Fakes Dataset	67.61%
FF++ [3]	UADFV [56]	97.92%
Deep Fakes Dataset	FF++ [3]	80.60%
Deep Fakes Dataset	Celeb-DF [4]	85.13%

교차 데이터 세트 결과

[*] U. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," in IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 2020

FakeCatcher: Results

Operation	Kernel	Accuracy
Original	N/A	91.50%
Gaussian blur	3x3	91.31%
Gaussian blur	5x5	88.61%
Gaussian blur	7x7	85.13%
Gaussian blur	9x9	70.84%

Extreme noise


Train	Test	Video Accuracy
FF++ - Face2Face	Face2Face [37]	95.25%
FF++ - FaceSwap	FaceSwap [59]	96.25%
FF++ - Deepfakes	Deepfakes [57]	93.75%
FF++ - NeuralTextures	NeuralTextures [77]	81.25%

Cross-model results

Train	Test	Video Accuracy
Celeb-DF [4]	FF++ [3]	83.10%
FF++ [3]	Celeb-DF [4]	86.48%
FF++ [3]	Deep Fakes Dataset	84.51%
Celeb-DF [4]	Deep Fakes Dataset	82.39%
Deep Fakes	FF [2]	86.34%
FF [2]	Deep Fakes Dataset	67.61%
FF++ [3]	UADFV [56]	97.92%
Deep Fakes Dataset	FF++ [3]	80.60%
Deep Fakes Dataset	Celeb-DF [4]	85.13%

Cross-dataset results

[*] U. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," in IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 2020

FakeCatcher: 결과

Operation	Kernel	Accuracy
Original	N/A	91.50%
Gaussian blur	3x3	91.31%
Gaussian blur	5x5	88.61%
Gaussian blur	7x7	85.13%
Gaussian blur	9x9	70.84%

극심한 노이즈


Train	Test	Video Accuracy
FF++ - Face2Face	Face2Face [37]	95.25%
FF++ - FaceSwap	FaceSwap [59]	96.25%
FF++ - Deepfakes	Deepfakes [57]	93.75%
FF++ - NeuralTextures	NeuralTextures [77]	81.25%

교차 모델 결과

Train	Test	Video Accuracy
Celeb-DF [4]	FF++ [3]	83.10%
FF++ [3]	Celeb-DF [4]	86.48%
FF++ [3]	Deep Fakes Dataset	84.51%
Celeb-DF [4]	Deep Fakes Dataset	82.39%
Deep Fakes	FF [2]	86.34%
FF [2]	Deep Fakes Dataset	67.61%
FF++ [3]	UADFV [56]	97.92%
Deep Fakes Dataset	FF++ [3]	80.60%
Deep Fakes Dataset	Celeb-DF [4]	85.13%

교차 데이터 세트 결과

[*] U. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," in IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 2020

FakeCatcher: Results

Operation	Kernel	Accuracy
Original	N/A	91.50%
Gaussian blur	3x3	91.31%
Gaussian blur	5x5	88.61%
Gaussian blur	7x7	85.13%
Gaussian blur	9x9	70.84%

Extreme noise

Dataset	Gen. Model	%Dataset	%Real	%Synthetic
UADFV [56]	FakeApp [75]	97.36%	94.93%	100%
FF++ [3]	Face2Face [37]	96.37%	96.00%	96.75%
FF [2]	Face2Face [37]	96%	94.24%	97.75%
FF++ [3]	FaceSwap [59]	95.75%	94.75%	96.75%
FF++ [3]	Deepfakes [57]	94.87%	93.25%	96.50%
FF++ [3]	All	94.65%	88.25%	96.25%
Celeb-DF [4]	Default	91.50%	76.40%	99.41%
DF (ours)	Mixed	91.07%	85.26%	96.89%
FF++ [3]	Neural Textures [77]	89.12%	86.75%	91.50%

In the wild

Train	Test	Video Accuracy
FF++ - Face2Face	Face2Face [37]	95.25%
FF++ - FaceSwap	FaceSwap [59]	96.25%
FF++ - Deepfakes	Deepfakes [57]	93.75%
FF++ - NeuralTextures	NeuralTextures [77]	81.25%

Cross-model results

Train	Test	Video Accuracy
Celeb-DF [4]	FF++ [3]	83.10%
FF++ [3]	Celeb-DF [4]	86.48%
FF++ [3]	Deep Fakes Dataset	84.51%
Celeb-DF [4]	Deep Fakes Dataset	82.39%
Deep Fakes	FF [2]	86.34%
FF [2]	Deep Fakes Dataset	67.61%
FF++ [3]	UADFV [56]	97.92%
Deep Fakes Dataset	FF++ [3]	80.60%
Deep Fakes Dataset	Celeb-DF [4]	85.13%

Cross-dataset results

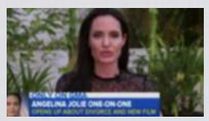
[*] U. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," in IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 2020

intel 15

FakeCatcher: 결과

Operation	Kernel	Accuracy
Original	N/A	91.50%
Gaussian blur	3x3	91.31%
Gaussian blur	5x5	88.61%
Gaussian blur	7x7	85.13%
Gaussian blur	9x9	70.84%

극심한 노이즈



Dataset	Gen. Model	%Dataset	%Real	%Synthetic
UADFV [56]	FakeApp [75]	97.36%	94.93%	100%
FF++ [3]	Face2Face [37]	96.37%	96.00%	96.75%
FF [2]	Face2Face [37]	96%	94.24%	97.75%
FF++ [3]	FaceSwap [59]	95.75%	94.75%	96.75%
FF++ [3]	Deepfakes [57]	94.87%	93.25%	96.50%
FF++ [3]	All	94.65%	88.25%	96.25%
Celeb-DF [4]	Default	91.50%	76.40%	99.41%
DF (ours)	Mixed	91.07%	85.26%	96.89%
FF++ [3]	Neural Textures [77]	89.12%	86.75%	91.50%

자연값

Train	Test	Video Accuracy
FF++ - Face2Face	Face2Face [37]	95.25%
FF++ - FaceSwap	FaceSwap [59]	96.25%
FF++ - Deepfakes	Deepfakes [57]	93.75%
FF++ - NeuralTextures	NeuralTextures [77]	81.25%

교차 모델 결과

Train	Test	Video Accuracy
Celeb-DF [4]	FF++ [3]	83.10%
FF++ [3]	Celeb-DF [4]	86.48%
FF++ [3]	Deep Fakes Dataset	84.51%
Celeb-DF [4]	Deep Fakes Dataset	82.39%
Deep Fakes	FF [2]	86.34%
FF [2]	Deep Fakes Dataset	67.61%
FF++ [3]	UADFV [56]	97.92%
Deep Fakes Dataset	FF++ [3]	80.60%
Deep Fakes Dataset	Celeb-DF [4]	85.13%

교차 데이터 세트 결과

[*] U. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," in IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 2020

intel 15


Deepfake Source Detection
Which GAN created this deepfake?

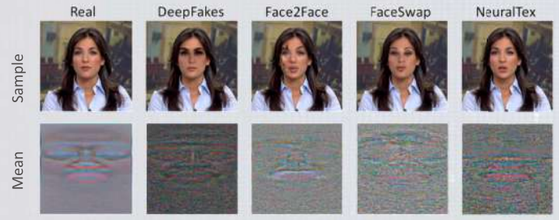

Find the
Real One!
intel
18


딥페이크 소스 감지
이 딥페이크를 생성한 GAN은 무엇인가?


진짜를
찾아라!
intel
18

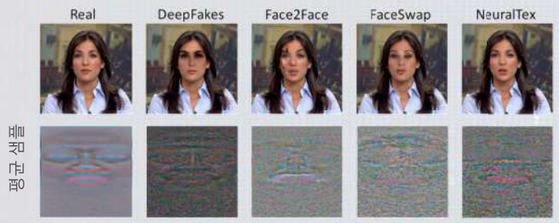
Source Detection: Intuition

- Instead of looking for an authenticity signature, can we interpret the generative noise of specific models by biological signals?



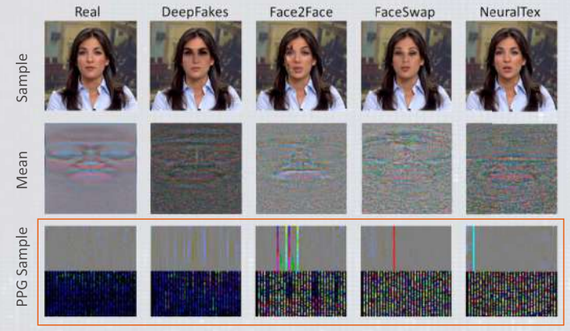
소스 감지: 직관

- 진위 서명을 찾지 않고, 특정 모델의 생성 노이즈를 생물학적 신호로 해석할 수 있는가?



Source Detection: Intuition

- Instead of looking for an authenticity signature, can we interpret the generative noise of specific models by biological signals?

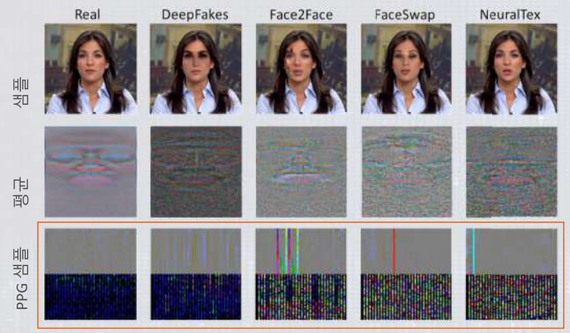


- Interpret residuals of specific models by projecting to biological signal domain.

[*] U. Ciftci, I. Demir, and L. Yin, "How Do the Hearts of Deepfakes Beat? Deepfake Source Detection via Interpreting Residuals with Biological Signals," in IEEE/IAPR International Joint Conference on Biometrics (IJCB). 2020

소스 감지: 직관

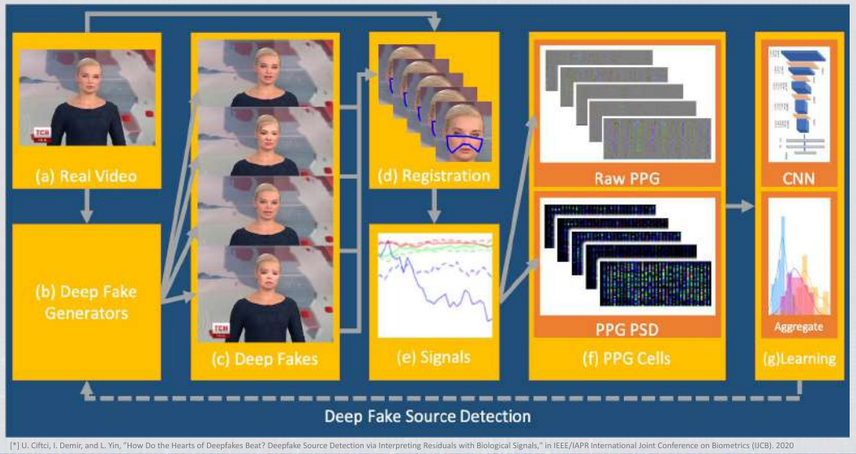
- 진위 서명을 찾지 않고, 특정 모델의 생성 노이즈를 생물학적 신호로 해석할 수 있는가?



- 특정 모델의 잔여 값을 생물학적 신호 도메인에 투영하여 해석

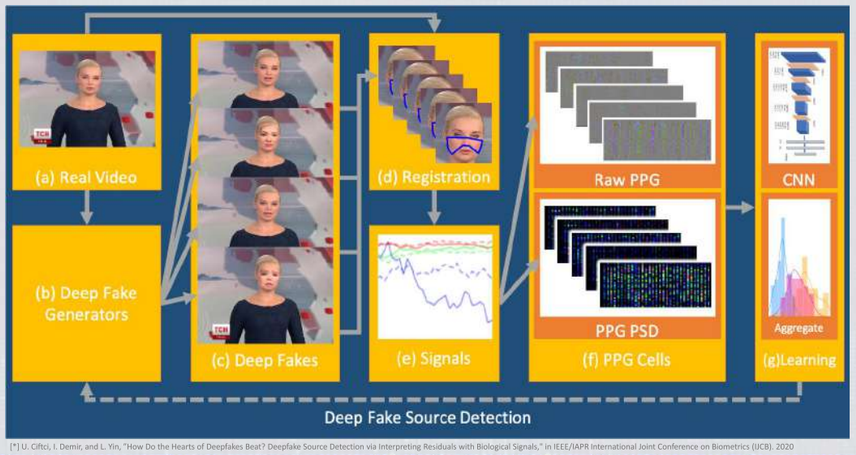
[*] U. Ciftci, I. Demir, and L. Yin, "How Do the Hearts of Deepfakes Beat? Deepfake Source Detection via Interpreting Residuals with Biological Signals," in IEEE/IAPR International Joint Conference on Biometrics (IJCB). 2020

Source Detection for ~~GANerated Images~~ Deepfakes?



[*] U. Ciftci, I. Demir, and L. Yin, "How Do the Hearts of Deepfakes Beat? Deepfake Source Detection via Interpreting Residuals with Biological Signals," in IEEE/IAPR International Joint Conference on Biometrics (IJCB). 2020

GAN으로 생성된 ~~이미지 딥페이크 소스~~ 감지?



[*] U. Ciftci, I. Demir, and L. Yin, "How Do the Hearts of Deepfakes Beat? Deepfake Source Detection via Interpreting Residuals with Biological Signals," in IEEE/IAPR International Joint Conference on Biometrics (IJCB). 2020

Source Detection: Results

- 1000 fakes of CelebDF = sixth class
- 93.69% source detection accuracy
- 96.89% fake detection accuracy
- 92.17% accuracy on the new class!

Source	Video SD Accuracy
CelebDF	92.17%
DeepFakes	94.66%
Face2Face	91.66%
FaceSwap	92.66%
NeuralTex	86.62%
Real	96.89%
Total	93.69%

- FakeCatcher: 96% accuracy
- Source Detection: 97.29% fake detection accuracy
  93.39% source detection accuracy

continuous detection and integration!

[*] U. Ciftci, I. Demir, and L. Yin, "How Do the Hearts of Deepfakes Beat? Deepfake Source Detection via Interpreting Residuals with Biological Signals," in IEEE/IAPR International Joint Conference on Biometrics (IJCB). 2020

소스 감지: 결과

- 1000개의 CelebDF = 6등급
- 93.69% 소스 감지 정확도
- 96.89% 가짜 감지 정확도
- 새로운 등급에서 92.17% 정확도

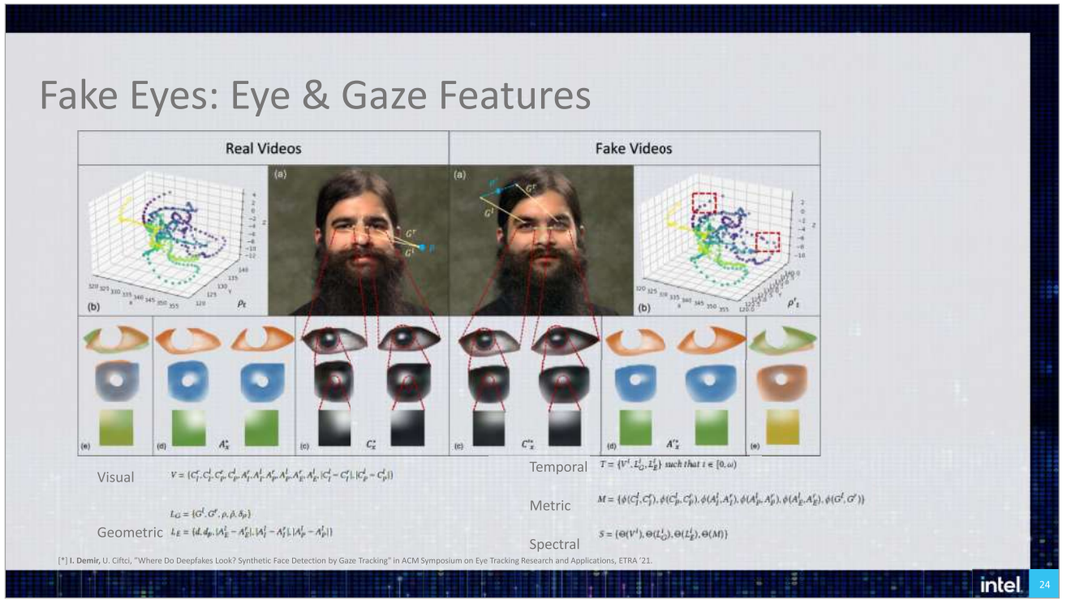
Source	Video SD Accuracy
CelebDF	92.17%
DeepFakes	94.66%
Face2Face	91.66%
FaceSwap	92.66%
NeuralTex	86.62%
Real	96.89%
Total	93.69%

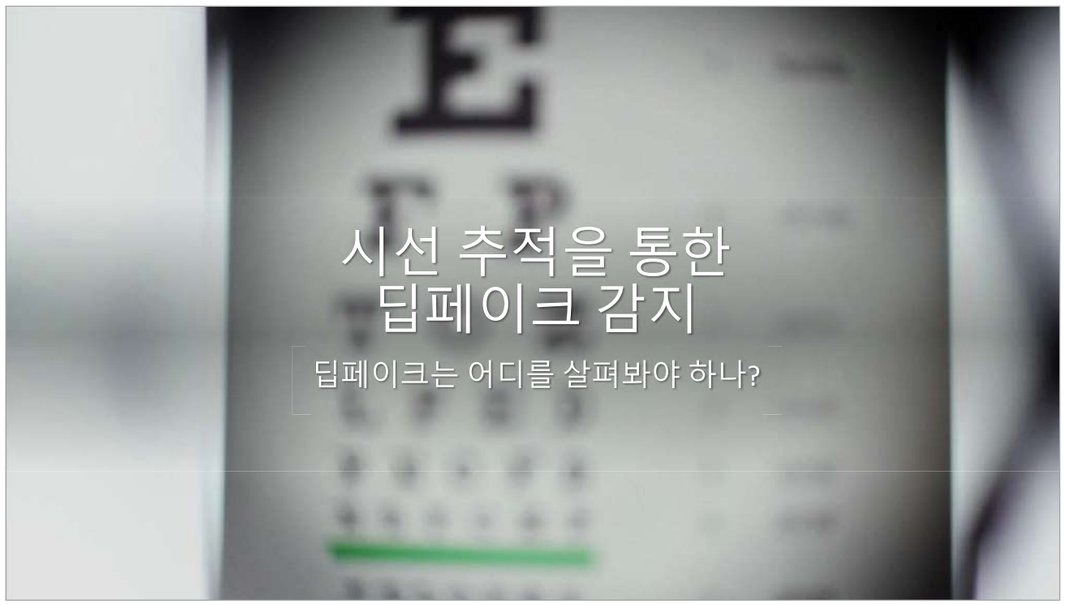
- FakeCatcher: 정확도 96%
- 소스 감지: 97.29% 가짜 감지 정확도
  93.39% 소스 감지 정확도

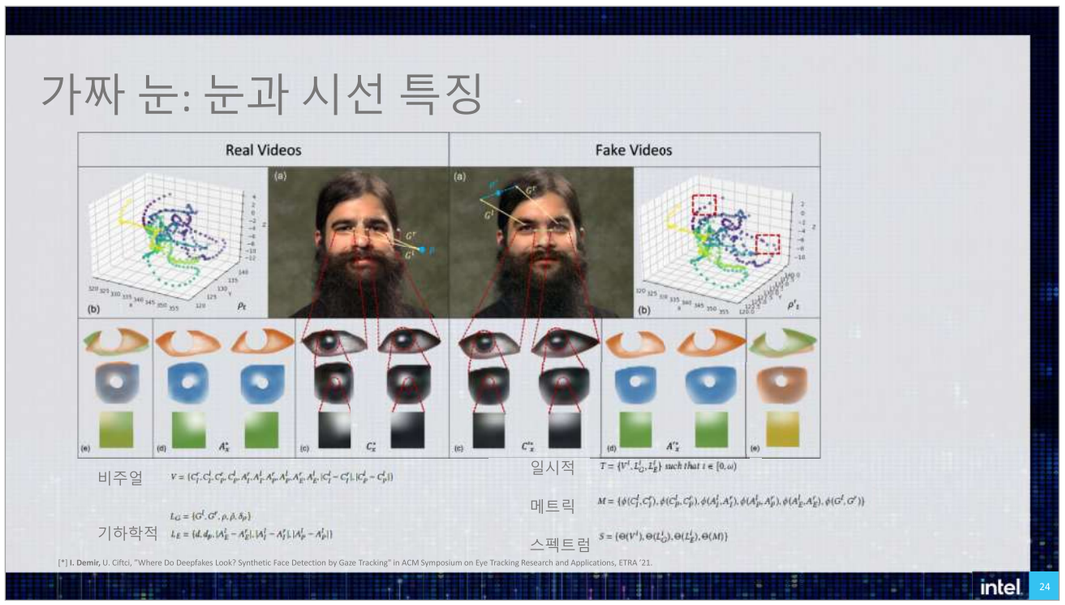
지속적인 감지 및 통합!

[*] U. Ciftci, I. Demir, and L. Yin, "How Do the Hearts of Deepfakes Beat? Deepfake Source Detection via Interpreting Residuals with Biological Signals," in IEEE/IAPR International Joint Conference on Biometrics (IJCB). 2020


Deepfake Detection
via Gaze Tracking
Where do deepfakes look?


Fake Eyes: Eye & Gaze Features
Real Videos
Fake Videos
Visual
Geometric
Temporal
Metric
Spectral
[*] I. Demir, U. Ciftci, "Where Do Deepfakes Look? Synthetic Face Detection by Gaze Tracking" in ACM Symposium on Eye Tracking Research and Applications, ETRA '21.
intel
24


시선 추적을 통한
딥페이크 감지
딥페이크는 어디를 살펴봐야 하나?


가짜 눈: 눈과 시선 특징
Real Videos
Fake Videos
비주얼
기하학적
일시적
메트릭
스펙트럼
[*] I. Demir, U. Ciftci, "Where Do Deepfakes Look? Synthetic Face Detection by Gaze Tracking" in ACM Symposium on Eye Tracking Research and Applications, ETRA '21.
intel
24

Fake Eyes: Deepfake Detection by Gaze Tracking

- Gaze signatures of size (40, ω, 3)
- Normalized features
- Equal number of reals and fakes for training
- 70/30 random train/test split
- Training 100 epochs < 1 hour

Powerful representation → simple network

[*] **I. Demir,** U. Ciftci, "Where Do Deepfakes Look? Synthetic Face Detection by Gaze Tracking" in ACM Symposium on Eye Tracking Research and Applications, ETRA '21.

가짜 눈: 시선 추적을 통한 딥페이크 감지

- size (40, ω, 3) 사이즈의 시선 신호
- 정규화된 특징
- 훈련을 위한 동일한 수의 진짜와 가짜
- 70/30 무작위 훈련/테스트 분할
- 100 epochs 훈련< 1시간

강력한 표현 → 단순 네트워크

[*] **I. Demir,** U. Ciftci, "Where Do Deepfakes Look? Synthetic Face Detection by Gaze Tracking" in ACM Symposium on Eye Tracking Research and Applications, ETRA '21.

Fake Eyes: Results

Source	S. Acc.	V. Acc.
Real	80.36	91.30
DeepFakes	81.02	93.28
FaceSwap	80.63	91.62
CDF [Li et al. 2020b]	85.76	88.35
DFor [Jiang et al. 2020]	95.57	99.27
FF++ [Rossler et al. 2019]	83.12	92.48
DF [Ciftci et al. 2020a]	79.84	80

99.27% on DeeperForensics
92.48% on FaceForensics++

Approach	FF/DF	FF/FS	DF
Blink [Li et al. 2018]	67.14	54.15	57.69
Head pose	55.69	50.08	67.85
PPG [Ciftci et al. 2020a]	94.87	95.75	91.07
Inception [Szegedy et al. 2016]	65.5*	54.4*	68.88
Xception [Chollet 2017]	74.5*	70.9*	75.55
Ours	93.28	91.62	80.00

Second best after FakeCatcher

Approach	Dataset	Acc.
3-layer CNN	DF	48.88
ConvLSTM	DF	48.83
[Tariq et al. 2018] V3	DF	73.33
[Tariq et al. 2018] ensemble	DF	80.00
MesoNet [Afchar et al. 2018]	FF/DF	87.3*
MesoNet [Afchar et al. 2018]	FF/FS	61.2*

Better than complex, deep, ensemble, and time-series networks

[*] **I. Demir,** U. Ciftci, "Where Do Deepfakes Look? Synthetic Face Detection by Gaze Tracking" in ACM Symposium on Eye Tracking Research and Applications, ETRA '21.

가짜 눈: 결과

Source	S. Acc.	V. Acc.
Real	80.36	91.30
DeepFakes	81.02	93.28
FaceSwap	80.63	91.62
CDF [Li et al. 2020b]	85.76	88.35
DFor [Jiang et al. 2020]	95.57	99.27
FF++ [Rossler et al. 2019]	83.12	92.48
DF [Ciftci et al. 2020a]	79.84	80

DeeperForensics에서 99.27%
FaceForensics++에서 92.48%

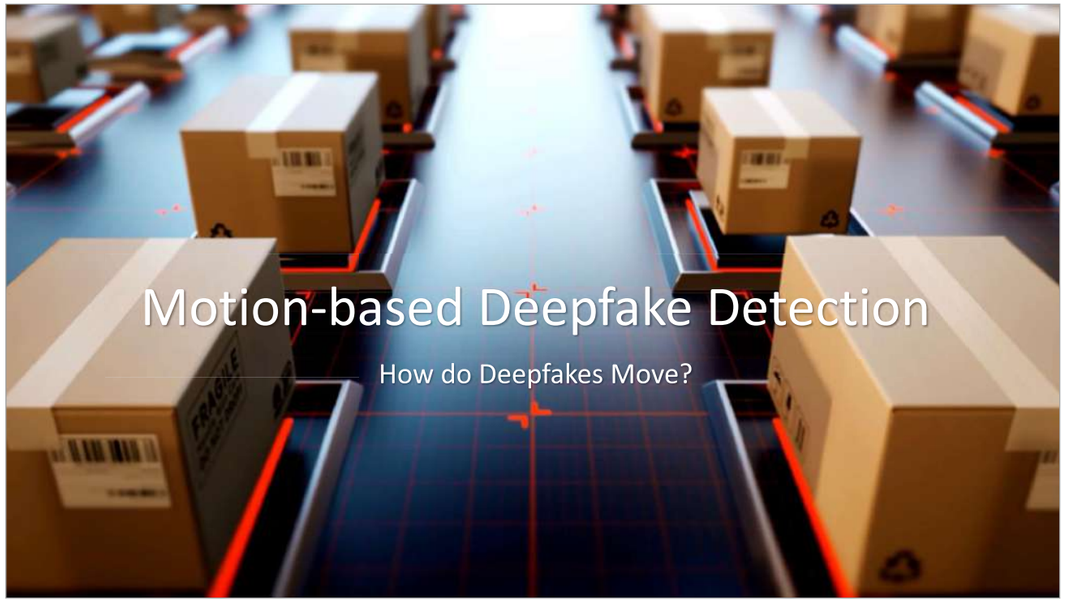
Approach	FF/DF	FF/FS	DF
Blink [Li et al. 2018]	67.14	54.15	57.69
Head pose	55.69	50.08	67.85
PPG [Ciftci et al. 2020a]	94.87	95.75	91.07
Inception [Szegedy et al. 2016]	65.5*	54.4*	68.88
Xception [Chollet 2017]	74.5*	70.9*	75.55
Ours	93.28	91.62	80.00

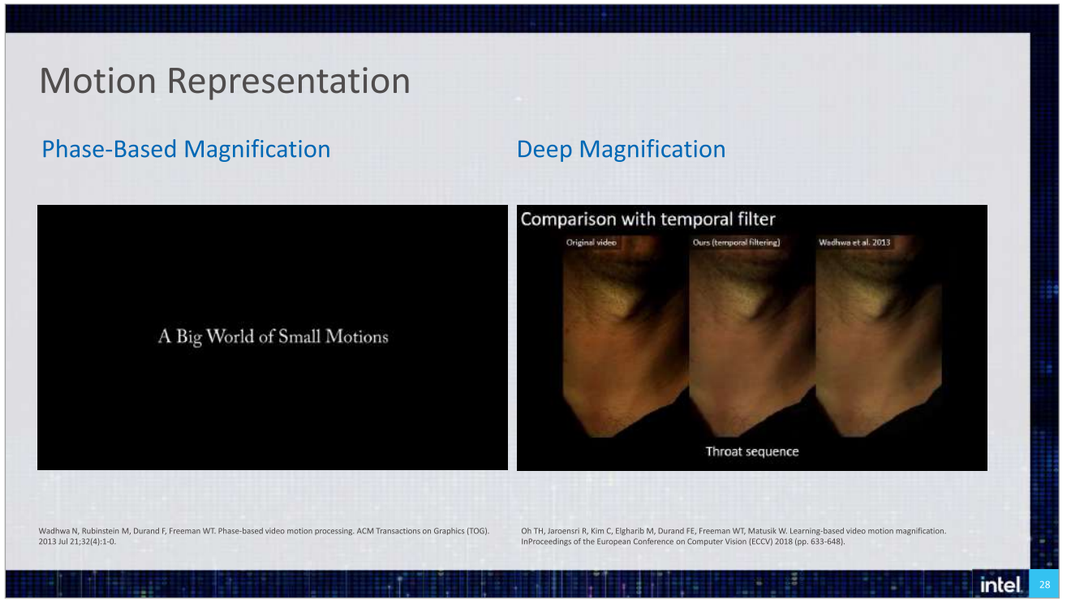
FakeCatcher 다음으로 좋은

Approach	Dataset	Acc.
3-layer CNN	DF	48.88
ConvLSTM	DF	48.83
[Tariq et al. 2018] V3	DF	73.33
[Tariq et al. 2018] ensemble	DF	80.00
MesoNet [Afchar et al. 2018]	FF/DF	87.3*
MesoNet [Afchar et al. 2018]	FF/FS	61.2*

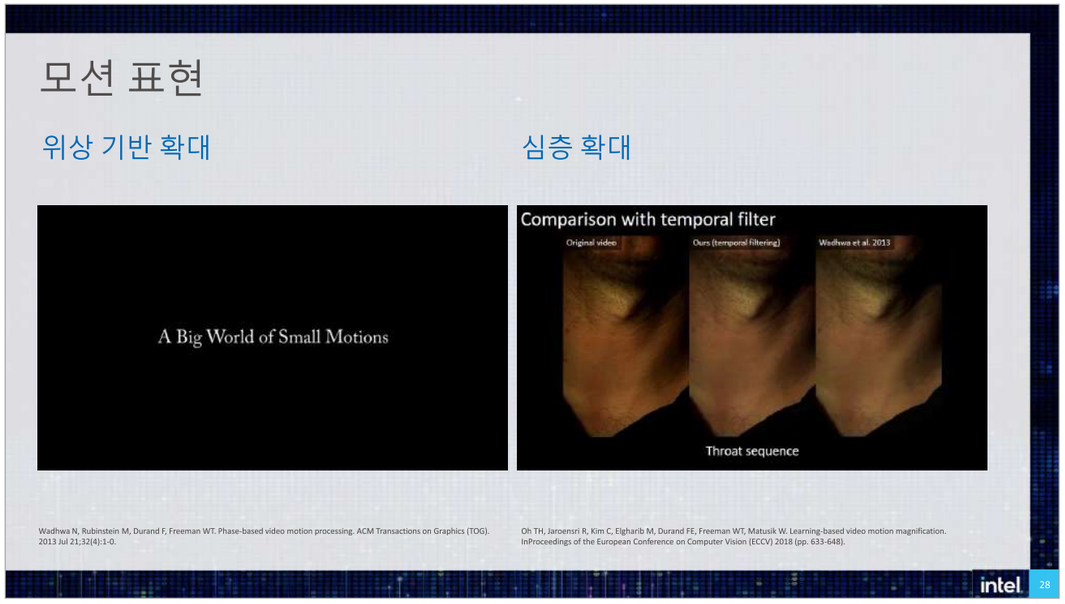
복잡한 네트워크, 딥 네트워크, 앙상블 네트워크, 시계열 네트워크보다 더 나은

[*] **I. Demir,** U. Ciftci, "Where Do Deepfakes Look? Synthetic Face Detection by Gaze Tracking" in ACM Symposium on Eye Tracking Research and Applications, ETRA '21.

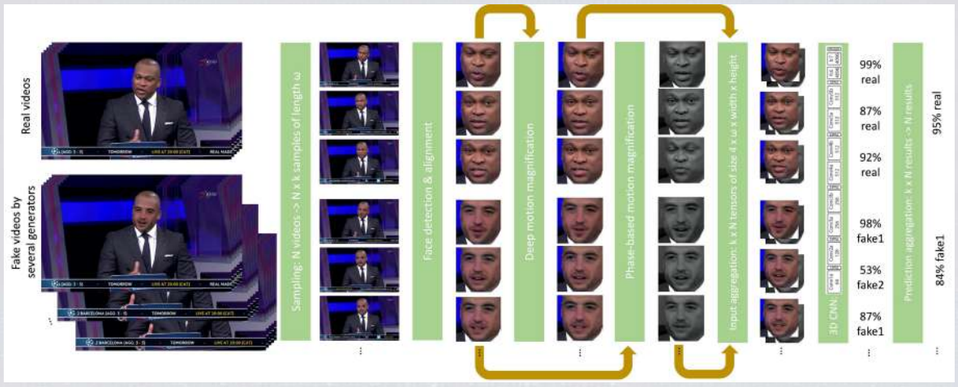

Motion-based Deepfake Detection
How do Deepfakes Move?


Motion Representation
Phase-Based Magnification
Deep Magnification
A Big World of Small Motions
Comparison with temporal filter
Original video
Ours (temporal filtering)
Wadhwa et al. 2013
Throat sequence
Wadhwa N, Rubinstein M, Durand F, Freeman WT. Phase-based video motion processing. ACM Transactions on Graphics (TOG). 2013 Jul 21;32(4):1-0.
Oh TH, Jaroensri R, Kim C, Elgharib M, Durand FE, Freeman WT, Matusik W. Learning-based video motion magnification. InProceedings of the European Conference on Computer Vision (ECCV) 2018 (pp. 633-648).
intel
28


동작 기반 딥페이크 감지
딥페이크는 어떻게 움직이는가?


모션 표현
위상 기반 확대
심층 확대
A Big World of Small Motions
Comparison with temporal filter
Original video
Ours (temporal filtering)
Wadhwa et al. 2013
Throat sequence
Wadhwa N, Rubinstein M, Durand F, Freeman WT. Phase-based video motion processing. ACM Transactions on Graphics (TOG). 2013 Jul 21;32(4):1-0.
Oh TH, Jaroensri R, Kim C, Elgharib M, Durand FE, Freeman WT, Matusik W. Learning-based video motion magnification. InProceedings of the European Conference on Computer Vision (ECCV) 2018 (pp. 633-648).
intel
28

Motion-based Detection: System Overview



[*] I. Demir, U. Ciftci, "How Do Deepfakes Move? Motion Magnification for Deepfake Source Detection," WACV, 2024

Motion-based Detection: Results

	FaceSwap	FSGAN	W2L	Real
FaceSwap	97.53	1.23	1.23	0.00
FSGAN	0.00	100.00	0.00	0.00
W2L	5.05	5.05	84.85	5.05
Real	1.43	1.43	5.71	91.43

	Deepfakes	Face2Face	FaceShifter	FaceSwap	NeuralTex	Real
Deepfakes	100.00	0.00	0.00	0.00	0.00	0.00
Face2Face	0.00	99.33	0.00	0.00	0.00	0.67
FaceShifter	0.00	0.00	99.67	0.00	0.00	0.33
FaceSwap	0.00	0.00	0.00	100.00	0.00	0.00
NeuralTex	0.00	1.33	0.00	0.00	93.00	5.67
Real	0.33	3.33	1.33	1.33	2.67	91.00

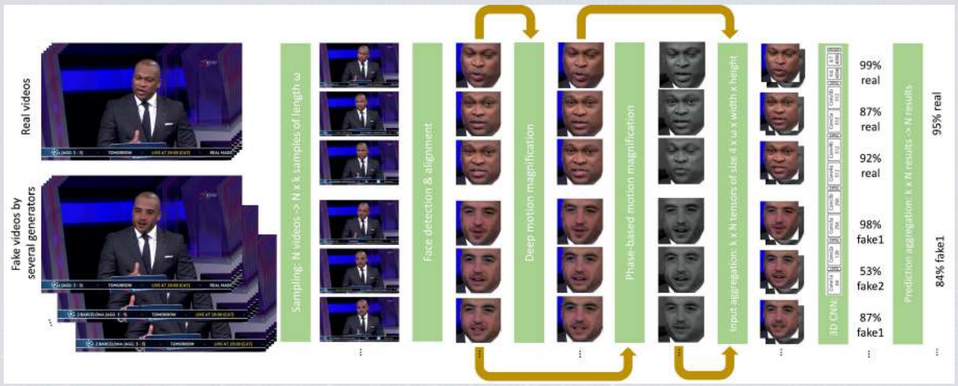
FaceForensics dataset:
- 97.77% video source detection accuracy,
- 95.92% sample source detection accuracy,
- 91% real class accuracy.

FakeAVCeleb dataset:
- 94.03% video source detection accuracy,
- 89.67% sample source detection accuracy,
- 91.43% real class accuracy.

Per-class accuracies > real class accuracy
- model learns motion of generative residue
- real class becomes “chaotic” class

[*] I. Demir, U. Ciftci, "How Do Deepfakes Move? Motion Magnification for Deepfake Source Detection," WACV, 2024

동작 기반 감지: 시스템 개요



[*] I. Demir, U. Ciftci, "How Do Deepfakes Move? Motion Magnification for Deepfake Source Detection," WACV, 2024

동작 기반 감지: 결과

	FaceSwap	FSGAN	W2L	Real
FaceSwap	97.53	1.23	1.23	0.00
FSGAN	0.00	100.00	0.00	0.00
W2L	5.05	5.05	84.85	5.05
Real	1.43	1.43	5.71	91.43

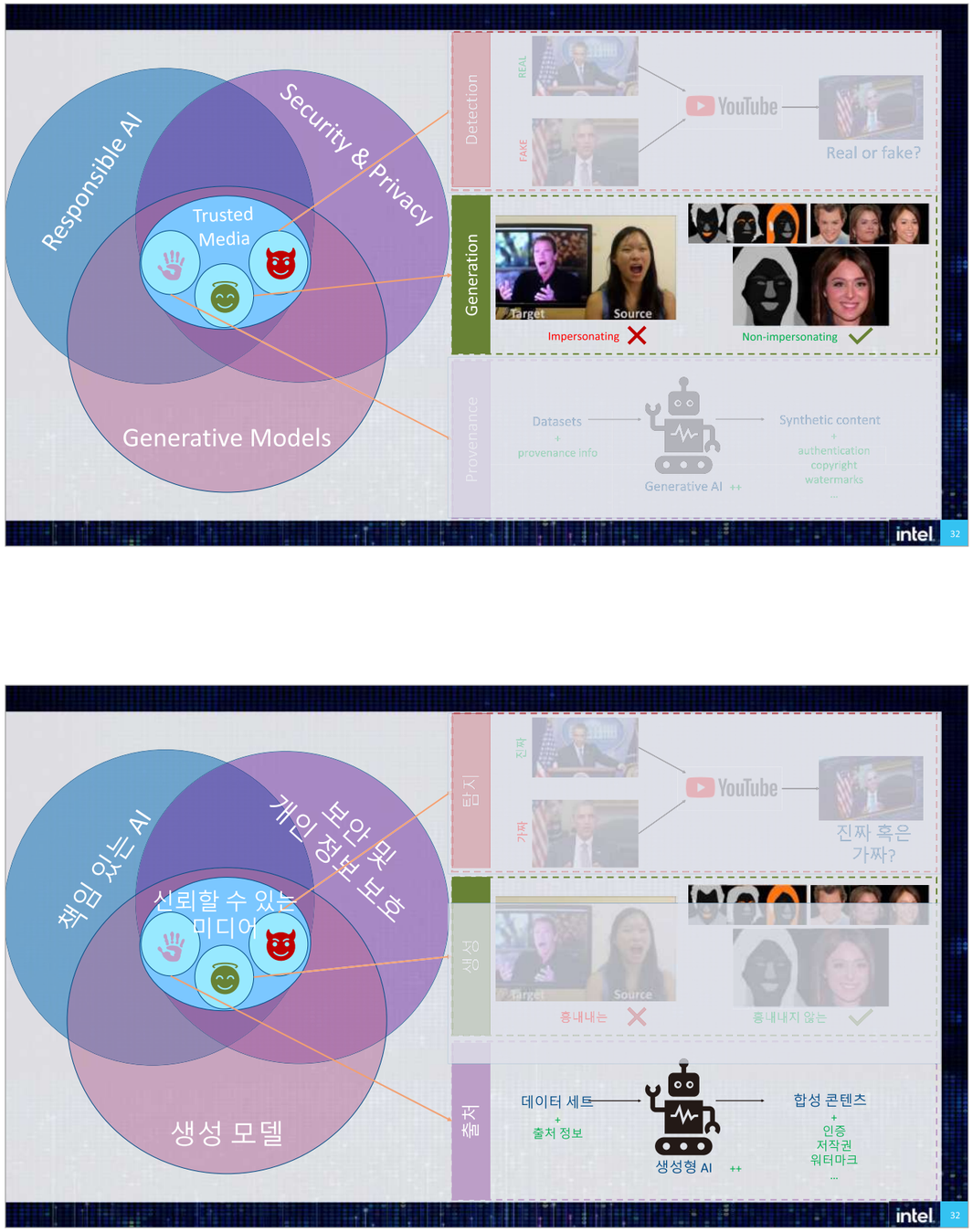
	Deepfakes	Face2Face	FaceShifter	FaceSwap	NeuralTex	Real
Deepfakes	100.00	0.00	0.00	0.00	0.00	0.00
Face2Face	0.00	99.33	0.00	0.00	0.00	0.67
FaceShifter	0.00	0.00	99.67	0.00	0.00	0.33
FaceSwap	0.00	0.00	0.00	100.00	0.00	0.00
NeuralTex	0.00	1.33	0.00	0.00	93.00	5.67
Real	0.33	3.33	1.33	1.33	2.67	91.00

FaceForensics 데이터 세트:
- 97.77% 비디오 소스 감지 정확도
- 95.92% 샘플 소스 감지 정확도
- 91% 실제 클래스 정확도

FakeAVCeleb 데이터 세트:
- 94.03% 비디오 소스 감지 정확도
- 89.67% 샘플 소스 감지 정확도
- 91.43% 실제 클래스 정확도

클래스별 정확도 > 실제 클래스 정확도
- 모델은 생성 잔류물의 움직임을 학습
- 실제 클래스= "혼돈스러운" 클래스

[*] I. Demir, U. Ciftci, "How Do Deepfakes Move? Motion Magnification for Deepfake Source Detection," WACV, 2024


Responsible AI
Security & Privacy
Trusted Media
Generative Models
Detection
REAL
FAKE
YouTube
Real or fake?
Generation
Target
Source
Impersonating
Non-impersonating
Provenance
Datasets
+
provenance info
Generative AI ++
Synthetic content
+
authentication
copyright
watermarks
...
intel
32


Responsible Deep Fakes:
Multi-source Generation
Can we mix multiple faces to prevent impersonation?

책임 있는 AI
보안 및
개인 정보 보호
신뢰할 수 있는
미디어
생성 모델
탐지
진짜
가짜
YouTube
진짜 혹은
가짜?
생성
Target
Source
흉내내는
흉내내지 않는
출처
데이터 세트
+
출처 정보
생성형 AI ++
합성 콘텐츠
+
인증
저작권
워터마크
...
intel
32


책임 있는 딥페이크:
다중 소스 생성
사칭을 막기 위해 여러 얼굴을 섞을 수 있나?

Deepfake Generation – in a Responsible Way?

- Source-target
- Mask dependent
- Discrete masks

intel 34

딥페이크 생성 - 책임감 있는 방식으로?

- 소스-타겟
- 마스크 의존
- 불연속 마스크

intel 34

Multi-source Face Synthesis

- ~~Source-target~~
- ~~Mask dependent~~
- ~~Discrete masks~~
- Multi-source
- Learning both composition & style
- Flexible compositions

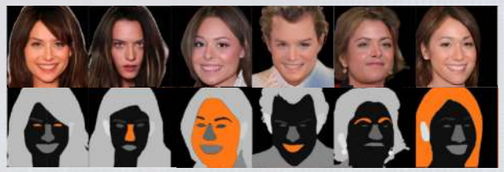
- How to mix and merge the masked regions?

- Copy-paste?

intel 35

다중 소스 얼굴 합성

- ~~소스타켁~~
- ~~마스크 의존~~
- ~~불연속 마스크~~
- 다중 소스
- 구성과 스타일을 모두 학습
- 다양한 구성 가능



- 마스크 영역 혼합 및 병합 방법은?

- 복사-붙여넣기?

intel 35

Multi-source Face Synthesis

- ~~Source-target~~
- ~~Mask dependent~~
- ~~Discrete masks~~

- Multi-source
- Learning both composition & style
- Flexible compositions

- How to mix and merge the masked regions?

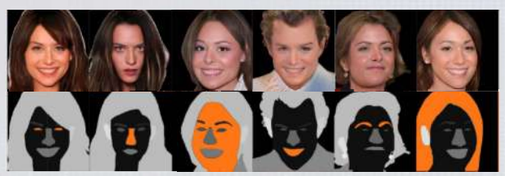
- Learn composition + style jointly!

[*] **I. Demir,** U. Ciftci, "MixSyn: Compositional Image Synthesis with Fuzzy Masks and Style Fusion", CVPRW, 2024.

다중 소스 얼굴 합성

- ~~소스타겟~~
- ~~마스크 의존~~
- ~~불연속 마스크~~

- 다중 소스
- 구성과 스타일을 모두 학습
- 다양한 구성 가능

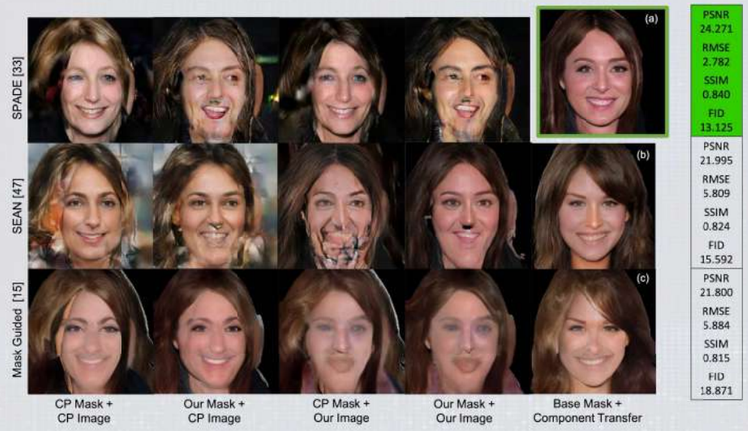


- 마스크 영역 혼합 및 병합 방법은?

- 구성 + 스타일 함께 배우기!

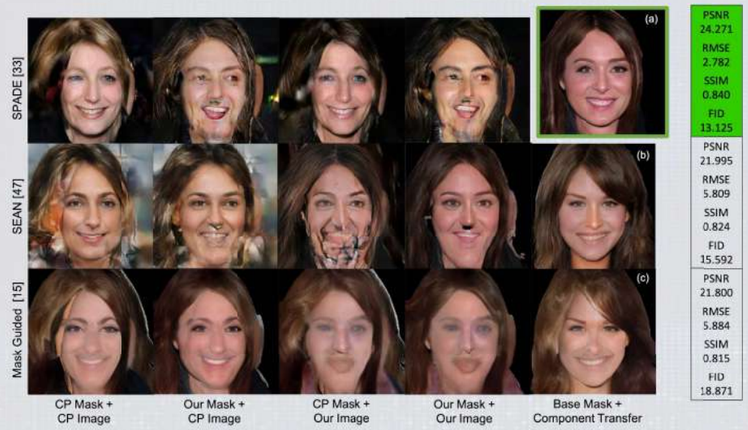
[*] **I. Demir,** U. Ciftci, "MixSyn: Compositional Image Synthesis with Fuzzy Masks and Style Fusion", CVPRW, 2024.

MixSyn: Comparison



[*] **I. Demir,** U. Ciftci, "MixSyn: Compositional Image Synthesis with Fuzzy Masks and Style Fusion", CVPRW, 2024.

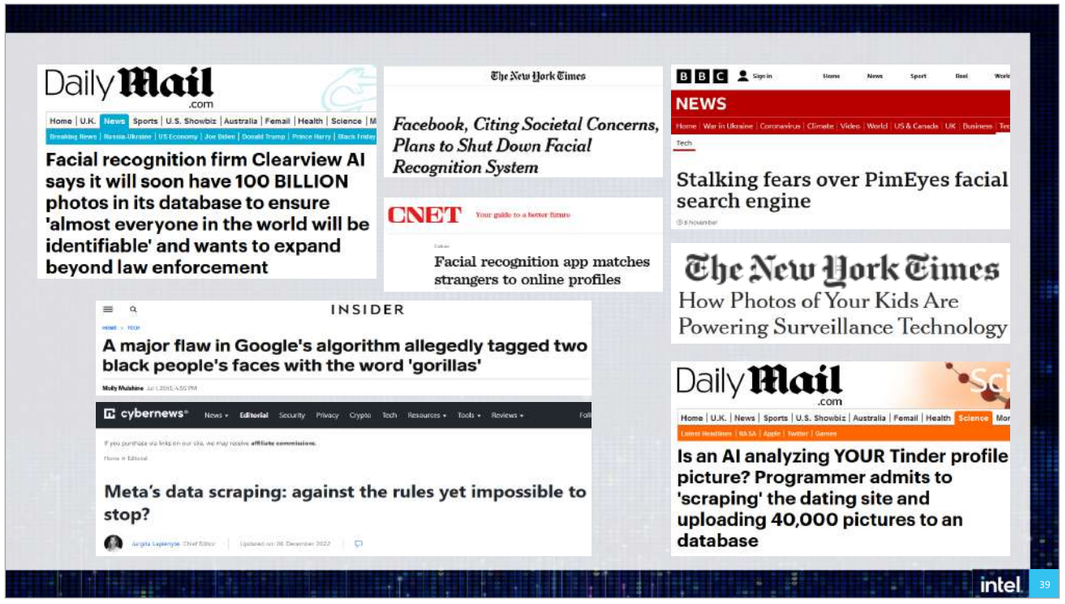
MixSyn: 비교



[*] **I. Demir,** U. Ciftci, "MixSyn: Compositional Image Synthesis with Fuzzy Masks and Style Fusion", CVPRW, 2024.

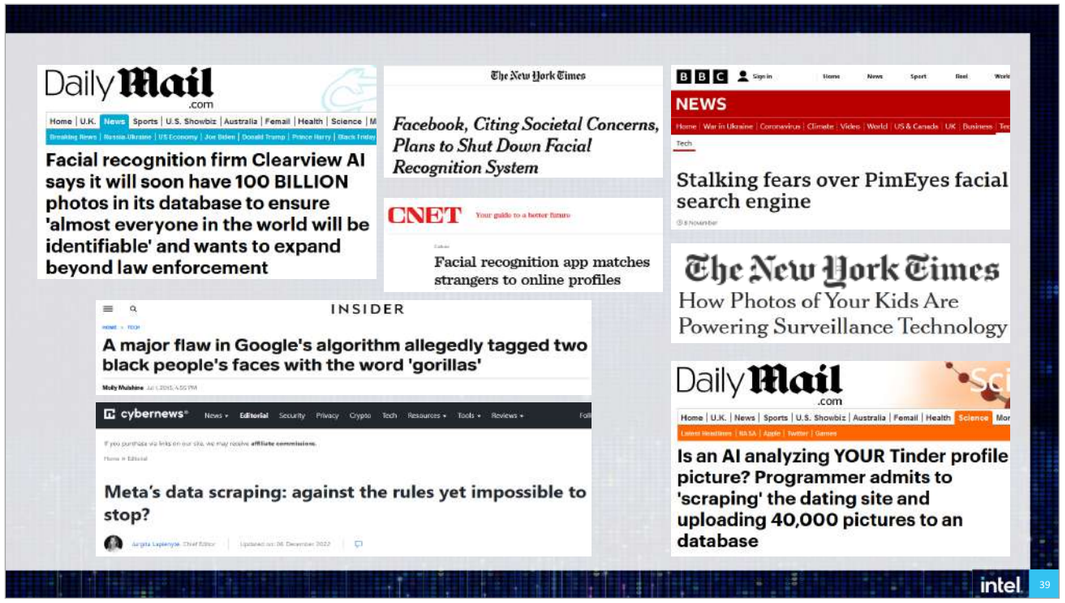
Deepfakes for Good: Privacy Enhancement for Anonymization

Can we mask faces without obvious artifacts?



딥페이크의 긍정적인 활용: 익명화를 위한 개인 정보 보호 강화

명백한 결함 없이 얼굴을 마스킹할 수 있을까?



My Face My Choice

- Empowering people to have control over faces
- Unauthorized access = deepfakes

- New faces
  - quantitatively dissimilar to the original
  - not similar to any other face
  - approximate the original age and gender
  - preserve head-pose and expression.

[*] U. Ciftci, G. Yuksek, **I. Demir**, "My Face My Choice: Privacy Enhancing Deepfakes for Social Media Anonymization", IEEE/CVF Winter Conference on Applications of Computer Vision. Jan 2023.

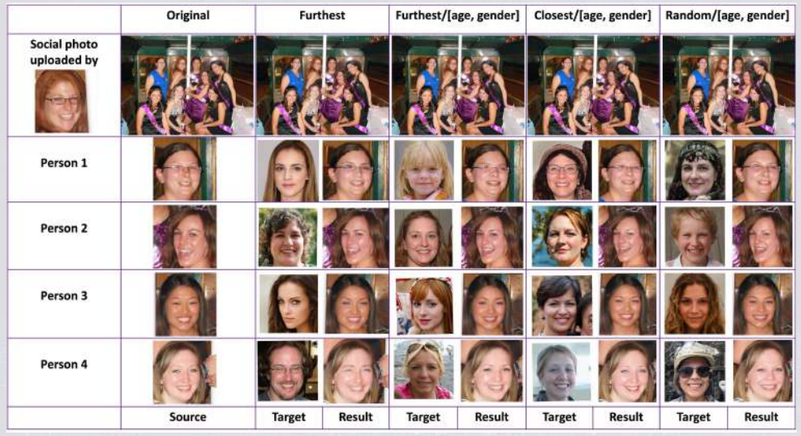
내 얼굴, 내 선택

- 사람들이 얼굴을 통제할 수 있도록 권한 부여
- 무단 액세스 = 딥페이크

- 뉴페이스
  .원래 얼굴과 정량적으로 다른 얼굴
  .다른 얼굴과 유사하지 않음
  .원래 나이와 성별에 근접
  .머리 자세와 표정 유지

[*] U. Ciftci, G. Yuksek, **I. Demir**, "My Face My Choice: Privacy Enhancing Deepfakes for Social Media Anonymization", IEEE/CVF Winter Conference on Applications of Computer Vision. Jan 2023.

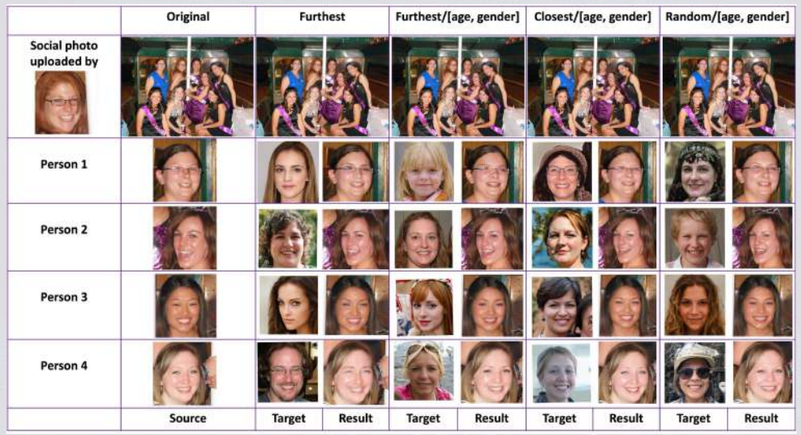
Face Anonymization in Social Networks



- Source image-> Non-existing faces
- Three privacy-preserving access designs for face-ownership.
- Evaluated on StyleGAN and balanced source datasets, with four GANs, five similarity spaces, and various distance configs.

[*] U. Ciftci, G. Yuksek, **I. Demir**, "My Face My Choice: Privacy Enhancing Deepfakes for Social Media Anonymization", IEEE/CVF Winter Conference on Applications of Computer Vision. Jan 2023.

SNS에서의 얼굴 익명화



- 소스 이미지> 존재하지 않는 얼굴
- 얼굴 소유권을 위한 3가지 프라이버시 보호 액세스 디자인.
- StyleGAN과 균형 잡힌 소스 데이터 세트에서 평가, 4개의 GAN, 5개의 유사성 공간, 다양한 거리 구성

[*] U. Ciftci, G. Yuksek, **I. Demir**, "My Face My Choice: Privacy Enhancing Deepfakes for Social Media Anonymization", IEEE/CVF Winter Conference on Applications of Computer Vision. Jan 2023.

Results: Face Recognition Accuracy

- 61% reduction on face recognition accuracy!
  - 65% if we lift the randomness
- Source-target distances:
  - How much recognizable is the selected target?
  - Almost zero means target query works!
- Source-result distances:
  - How much recognizable is the created face?
  - More than the target (obviously) but still high!
- Per detector thresholds on the cosine distance
  - See supp for L2
  - See supp for SSIM and RMSE instead of face embedding distance

Face Detector	Source vs. Target		Source vs. Result	
	Furthest	Random	Furthest	Random
FaceNet512	0.001	0.0	0.14	0.16
OpenFace	0.001	0.002	0.2	0.23
FaceNet	0.03	0.02	0.32	0.34
DLib	0.02	0.05	0.35	0.45
ArcFace	0.06	0.04	0.36	0.45
DeepID	0.006	0.01	0.54	0.55
DeepFace	0.04	0.06	0.57	0.55
Average	0.02	0.02	0.35	0.39

결과: 얼굴 인식 정확도

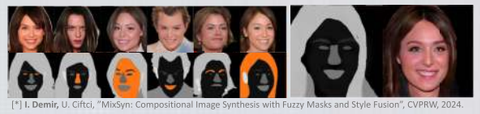
- 얼굴 인식 정확도 61% 감소!
  - 무작위성 제거 시 65% 감소
- 소스-대상 거리:
  - 선택한 대상을 얼마나 인식할 수 있나?
  - 거의 0이면 대상 쿼리가 작동!
- 소스-결과 거리:
  - 생성된 얼굴을 얼마나 인식할 수 있는가?
  - 대상보다 더 크지만(분명히) 여전히 높음!
- 코사인 거리에 대한 검출기당 임계값
  - L2에 대한 보충 설명 참조
  - 얼굴 임베딩 거리 대신 SSIM 및 RMSE에 대한 보충 설명 참조

Face Detector	Source vs. Target		Source vs. Result	
	Furthest	Random	Furthest	Random
FaceNet512	0.001	0.0	0.14	0.16
OpenFace	0.001	0.002	0.2	0.23
FaceNet	0.03	0.02	0.32	0.34
DLib	0.02	0.05	0.35	0.45
ArcFace	0.06	0.04	0.36	0.45
DeepID	0.006	0.01	0.54	0.55
DeepFace	0.04	0.06	0.57	0.55
Average	0.02	0.02	0.35	0.39

My Face My Choice: Privacy Enhancing Deepfakes for Social Media

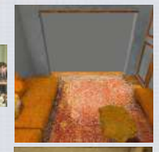
[*] U. Ciftci, G. Yuksek, **I. Demir**, "My Face My Choice: Privacy Enhancing Deepfakes for Social Media Anonymization", IEEE WACV 2023.

Multi-Source Synthesis by Composition and Style



[*] **I. Demir**, U. Ciftci, "MixSyn: Compositional Image Synthesis with Fuzzy Masks and Style Fusion", CVPRW, 2024.

Compositional 3D Scene Generation



[*] Coming soon!

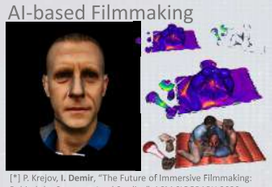
My Body My Choice

[*] U. Ciftci, A. K. Tanriverdi, **I. Demir**, "My Body My choice: Human Centric Full Body Anonymization, CVPRW, 2024.

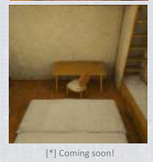
Unsupervised Content Removal

[*] Coming soon!

AI-based Filmmaking

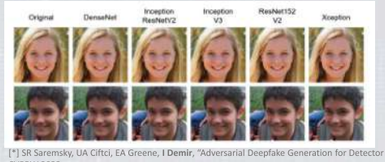


[*] P. Krejov, **I. Demir**, "The Future of Immersive Filmmaking: Behind the Scenes at Intel Studios", ACM SIGGRAPH 2020.



[*] Coming soon!

Detector Misclassification by Adversarial Generation

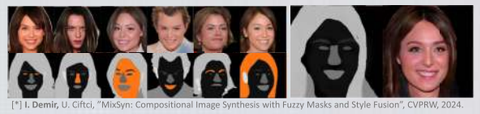


[*] SR Saremsky, UA Ciftci, EA Greene, **I Demir**, "Adversarial Deepfake Generation for Detector Misclassification", CVPRW 2022

내 얼굴 내 선택: 소셜 미디어를 위한 프라이버시 강화 딥페이크

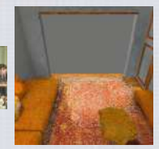
[*] U. Ciftci, G. Yuksek, **I. Demir**, "My Face My Choice: Privacy Enhancing Deepfakes for Social Media Anonymization", IEEE WACV 2023.

구성 및 스타일별 다중 소스 합성



[*] **I. Demir**, U. Ciftci, "MixSyn: Compositional Image Synthesis with Fuzzy Masks and Style Fusion", CVPRW, 2024.

합성 3D 장면 생성

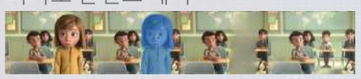


[*] Coming soon!

내 몸 내 선택

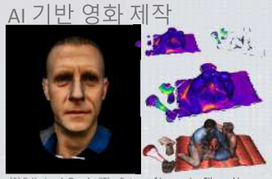
[*] U. Ciftci, A. K. Tanriverdi, **I. Demir**, "My Body My choice: Human Centric Full Body Anonymization, CVPRW, 2024.

비지도 콘텐츠 제거



[*] Coming soon!

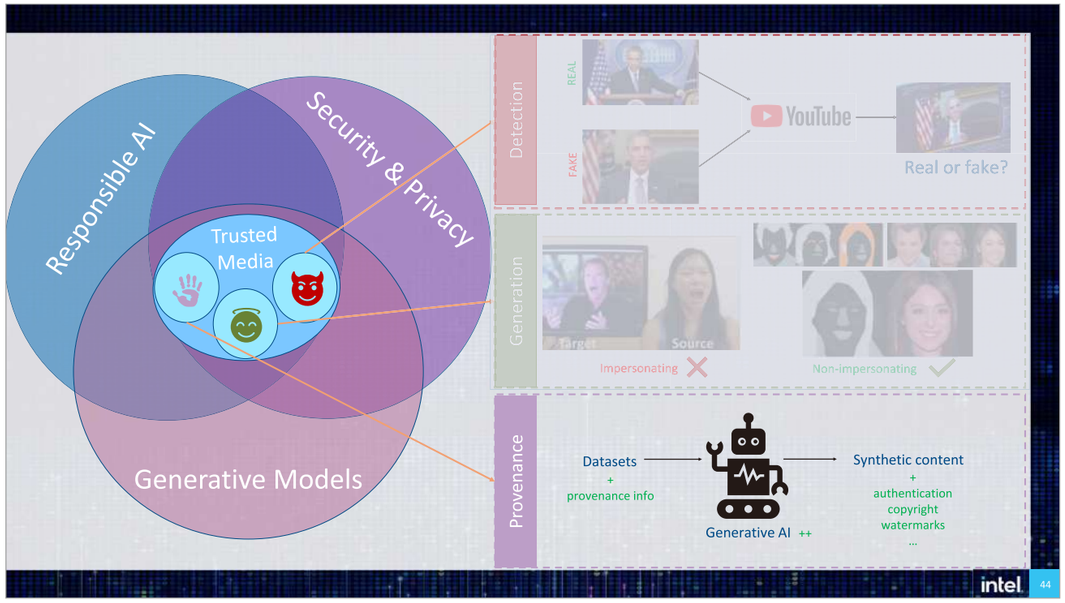
AI 기반 영화 제작

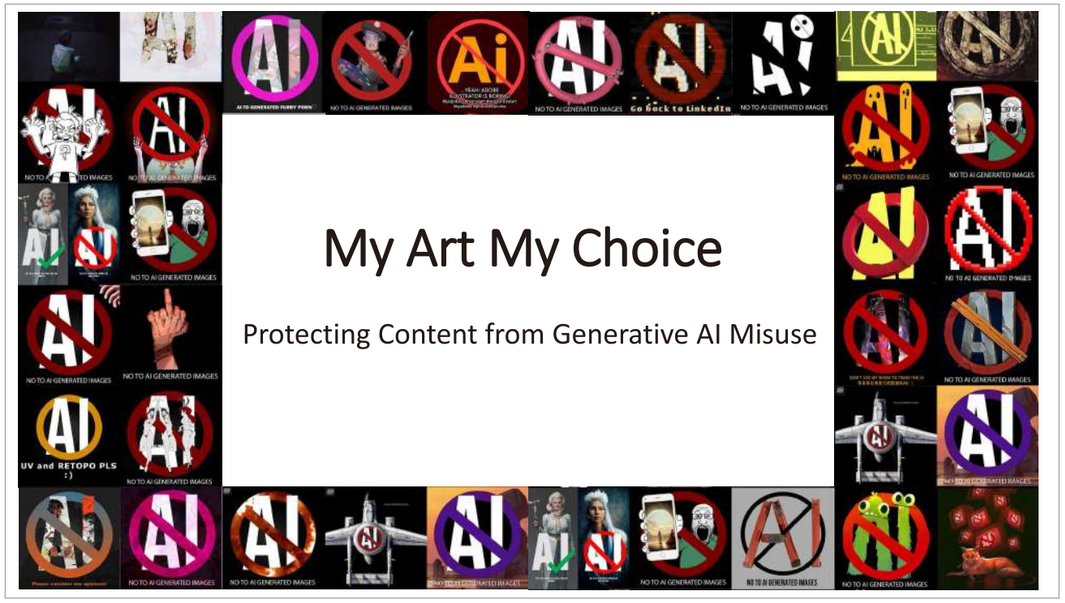


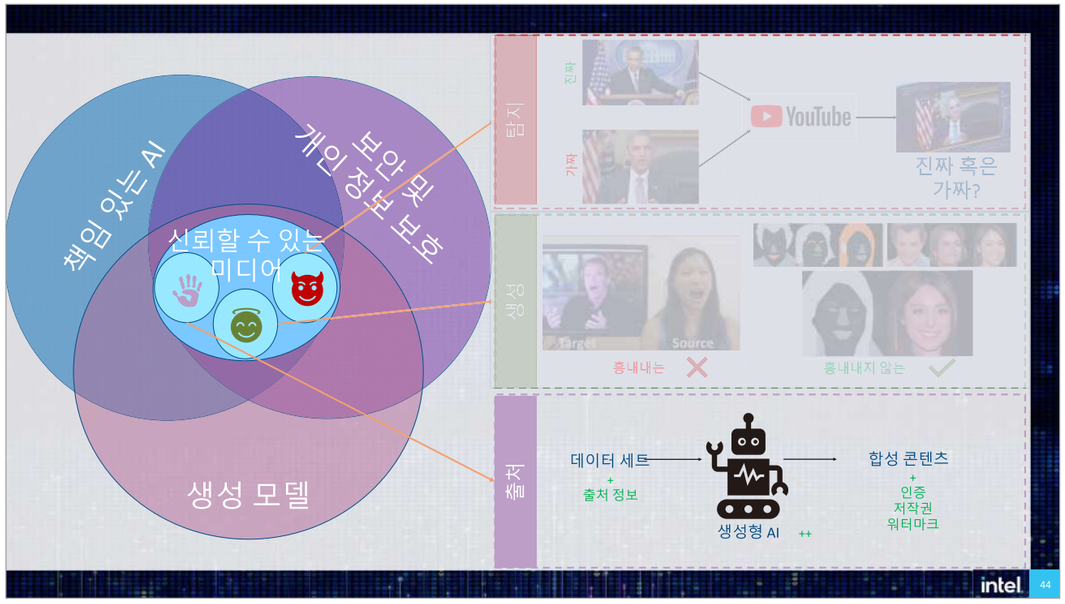
[*] P. Krejov, **I. Demir**, "The Future of Immersive Filmmaking: Behind the Scenes at Intel Studios", ACM SIGGRAPH 2020.

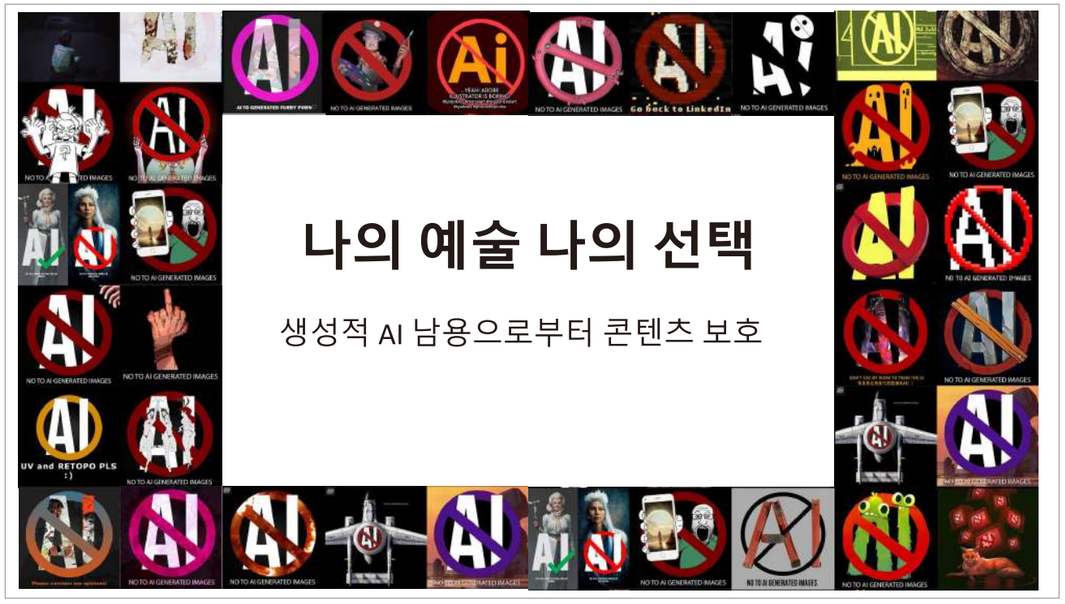
적대적 생성에 의한 감지기 오분류

[*] SR Saremsky, UA Ciftci, EA Greene, **I Demir**, "Adversarial Deepfake Generation for Detector Misclassification", CVPRW 2022


Responsible AI
Security & Privacy
Trusted Media
Generative Models
Detection
REAL
FAKE
YouTube
Real or fake?
Generation
Target
Source
Impersonating
Non-impersonating
Provenance
Datasets
+
provenance info
Generative AI ++
Synthetic content
+
authentication
copyright
watermarks
...
intel
44

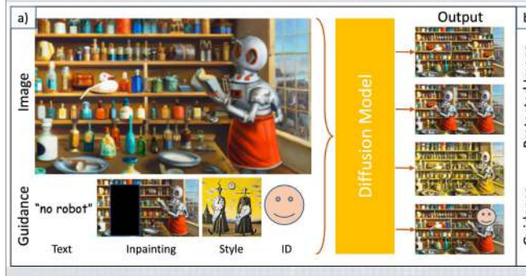

My Art My Choice
Protecting Content from Generative AI Misuse


책임 있는 AI
보안 및
개인 정보 보호
신뢰할 수 있는
미디어
생성 모델
탐지
진짜
가짜
YouTube
진짜 혹은
가짜?
생성
Target
Source
흉내내는
흉내내지 않는
출처
데이터 세트
+
출처 정보
생성형 AI ++
합성 콘텐츠
+
인증
저작권
워터마크
intel
44


나의 예술 나의 선택
생성적 AI 남용으로부터 콘텐츠 보호

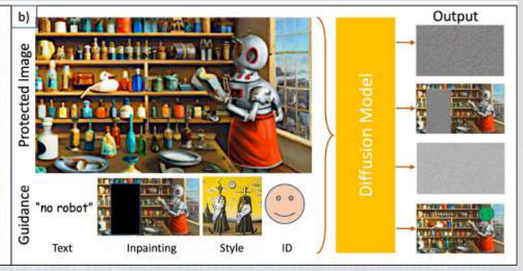
Generative AI Use Cases

Current



Diffusion models can replicate style, content, face, etc. properties of an image, however, these are not always permitted by the owner of the content.

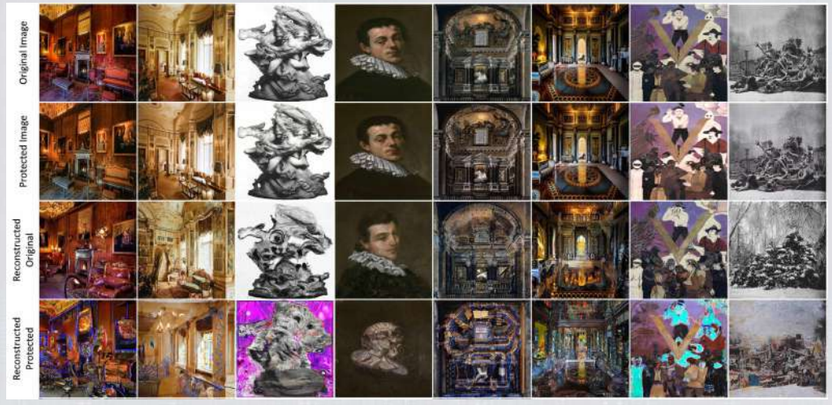
Motivation



Can we attack diffusion models in a black box manner to learn how we can break them by adversarial generation?

My Art My Choice

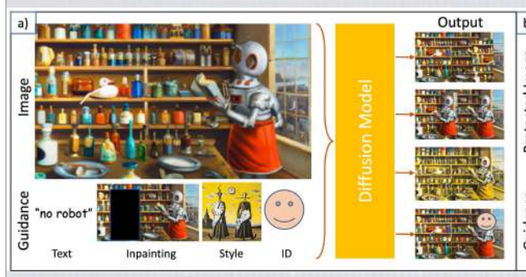
- Empower content owners by protecting their copyrighted materials from being utilized by diffusion models
- Let perturbation amount be decided by the artist to balance distortion vs. protection of the content.



[*] A. Rhodes, R. Bhagat, U. A. Ciftci, and **I. Demir**. "My Art My Choice: Adversarial Protection Against Unruly AI". CVPRW 2024.

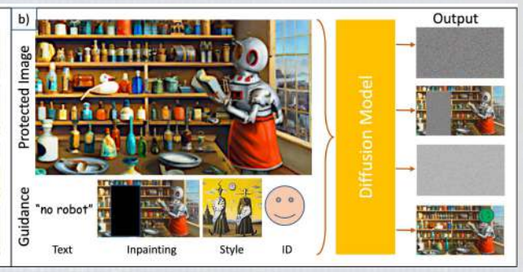
생성 AI 사용 사례

현재



확산 모델은 이미지의 스타일, 내용, 얼굴 등 속성을 복제할 수 있지만, 이러한 작업이 항상 콘텐츠 소유자의 허가를 받지는 않는다

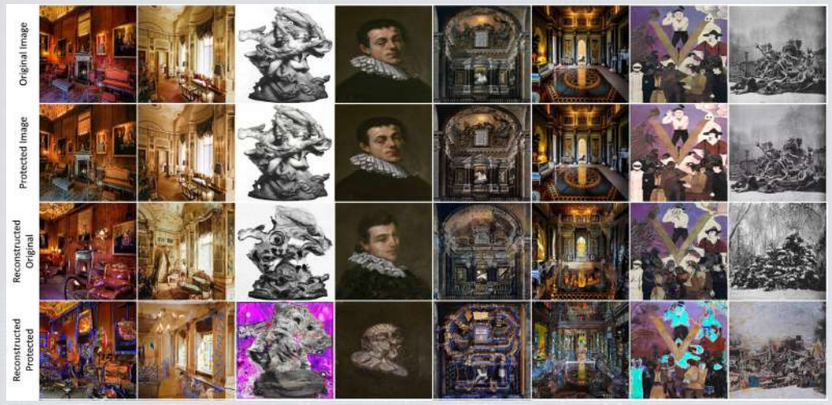
동기 부여



확산 모델을 블랙 박스 방식으로 공격하여 적대적 생성으로 이를 어떻게 파괴할 수 있는지 알 수 있을까?

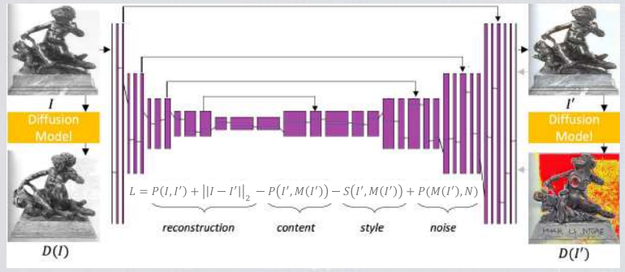
나의 예술 나의 선택

- 콘텐츠 소유자의 권리를 강화하여 저작권 자료가 확산 모델에서 사용되지 않도록 보호한다.
- 변형량은 아티스트가 결정하여 왜곡과 콘텐츠 보호 간의 균형을 맞춘다.



[*] A. Rhodes, R. Bhagat, U. A. Ciftci, and **I. Demir**. "My Art My Choice: Adversarial Protection Against Unruly AI". CVPRW 2024.

MAMC Model



$L_R = \alpha_{R1}\mathcal{P}(I,I') + \alpha_{R2}||I - I'||_2^2$ Reconstruction Loss

$L_C = -\alpha_C \mathcal{P}(I', M(I'))$ Content loss

$L_N = \alpha_{N1}\mathcal{P}(M(I'), \mathcal{N})$ Noise loss

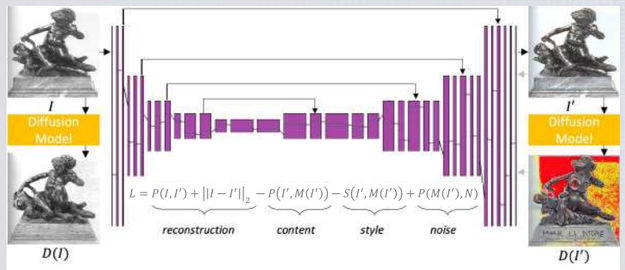
$L_S = -\alpha_S \frac{1}{|j|}\sum_j ||\Omega_j(I') - \Omega_j(M(I'))||$ Style loss

$L = \alpha_R L_R - \alpha_C L_C - \alpha_S L_S + \alpha_N L_N$ Final loss

- Simple U-Net architecture
- Standard pre-trained diffusion model
- Combination of four losses to direct fidelity of the protected image and atrophy of diffusion output

[*] A. Rhodes, R. Bhagat, U. A. Ciftci, and **I. Demir**. "My Art My Choice: Adversarial Protection Against Unruly AI". CVPRW 2024.

MAMC 모델



$L_R = \alpha_{R1}\mathcal{P}(I,I') + \alpha_{R2}||I - I'||_2^2$ 재구성 손실

$L_C = -\alpha_C \mathcal{P}(I', M(I'))$ 콘텐츠 손실

$L_N = \alpha_{N1}\mathcal{P}(M(I'), \mathcal{N})$ 노이즈 손실

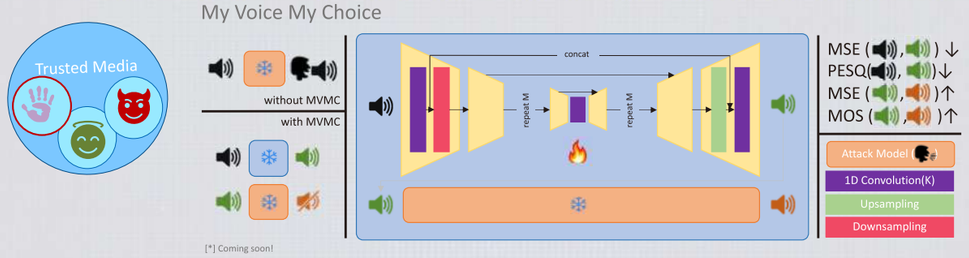
$L_S = -\alpha_S \frac{1}{|j|}\sum_j ||\Omega_j(I') - \Omega_j(M(I'))||$ 스타일 손실

$L = \alpha_R L_R - \alpha_C L_C - \alpha_S L_S + \alpha_N L_N$ 최종 손실

- 간단한 U-Net 아키텍처
- 표준 사전 훈련된 확산 모델
- 보호된 이미지의 정확성과 확산 출력의 감쇠를 위한 네 가지 손실 조합

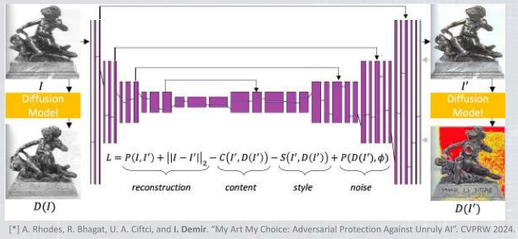
[*] A. Rhodes, R. Bhagat, U. A. Ciftci, and **I. Demir**. "My Art My Choice: Adversarial Protection Against Unruly AI". CVPRW 2024.

My Voice My Choice



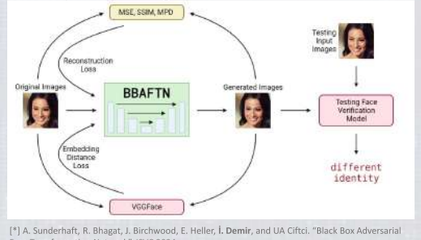
[*] Coming soon!

My Art My Choice



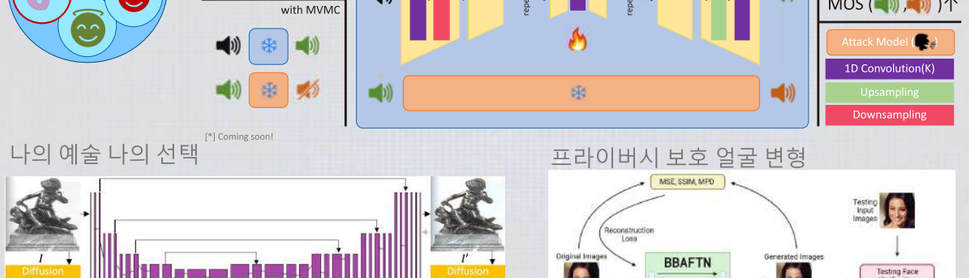
[*] A. Rhodes, R. Bhagat, U. A. Ciftci, and **I. Demir**. "My Art My Choice: Adversarial Protection Against Unruly AI". CVPRW 2024.

Privacy Preserving Face Transformation



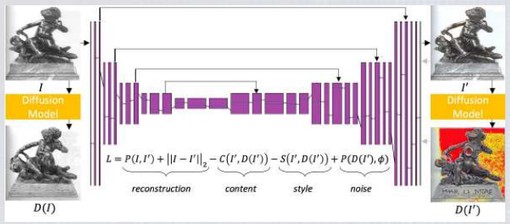
[*] A. Sunderhaft, R. Bhagat, J. Birchwood, E. Heller, **İ. Demir**, and UA Ciftci. "Black Box Adversarial Face Transformation Network", ISVC 2024

내 목소리 내 선택



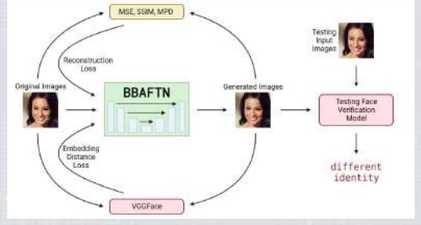
[*] Coming soon!

나의 예술 나의 선택

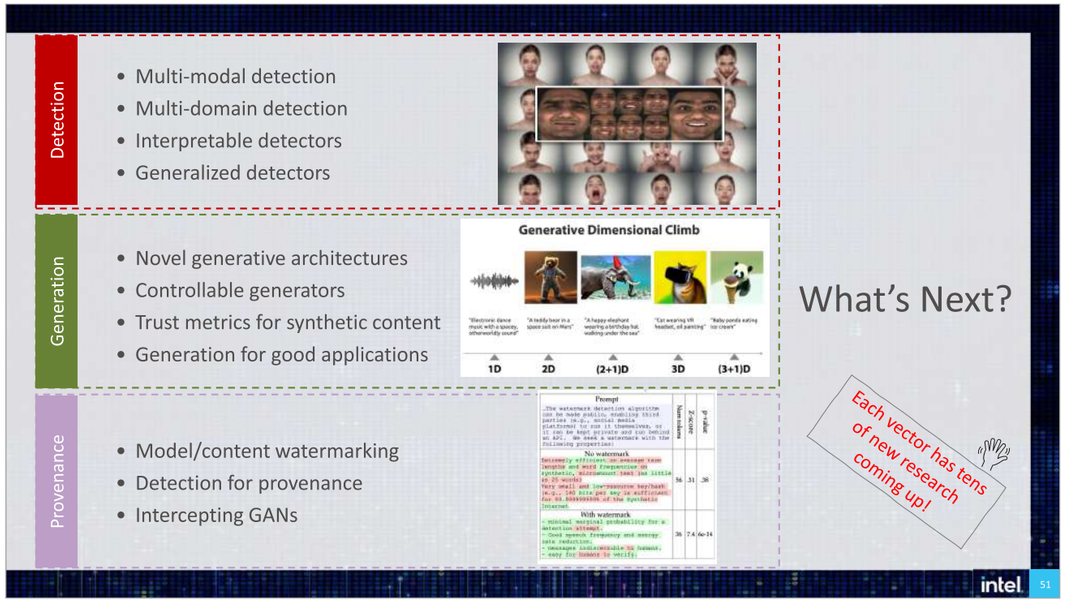


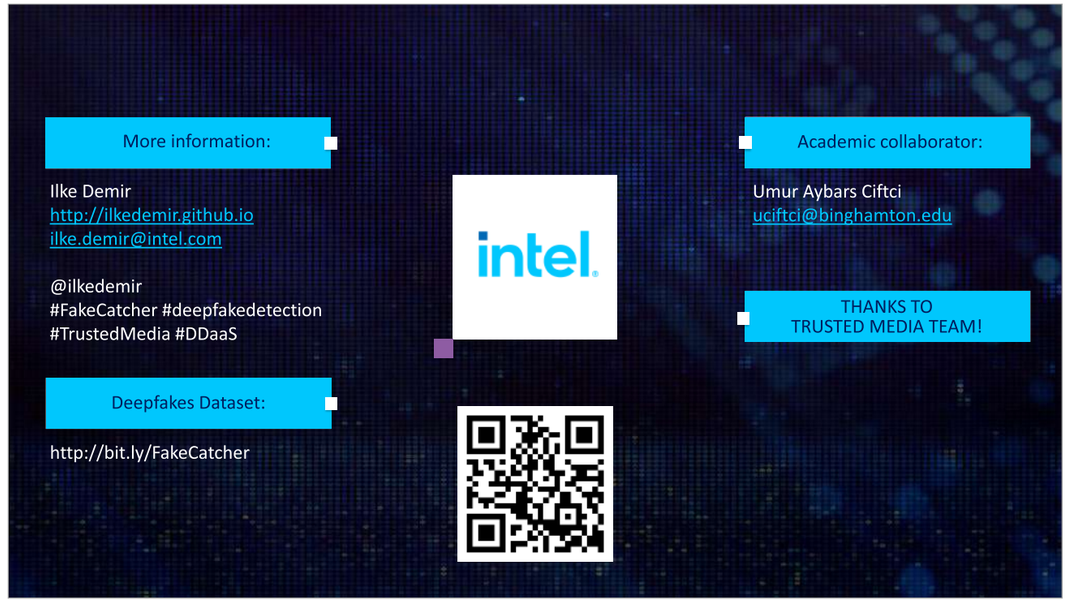
[*] A. Rhodes, R. Bhagat, U. A. Ciftci, and **I. Demir**. "My Art My Choice: Adversarial Protection Against Unruly AI". CVPRW 2024.

프라이버시 보호 얼굴 변형



[*] A. Sunderhaft, R. Bhagat, J. Birchwood, E. Heller, **İ. Demir**, and UA Ciftci. "Black Box Adversarial Face Transformation Network", ISVC 2024


Detection
• Multi-modal detection
• Multi-domain detection
• Interpretable detectors
• Generalized detectors
Generation
• Novel generative architectures
• Controllable generators
• Trust metrics for synthetic content
• Generation for good applications
Generative Dimensional Climb
1D
2D
(2+1)D
3D
(3+1)D
Provenance
• Model/content watermarking
• Detection for provenance
• Intercepting GANs
Prompt
No watermark
With watermark
What's Next?
Each vector has tens of new research coming up!
intel
51


More information:
Ilke Demir
http://ilkedemir.github.io
ilke.demir@intel.com
@ilkedemir
#FakeCatcher #deepfakedetection
#TrustedMedia #DDaaS
Deepfakes Dataset:
http://bit.ly/FakeCatcher
intel
Academic collaborator:
Umur Aybars Ciftci
uciftci@binghamton.edu
THANKS TO
TRUSTED MEDIA TEAM!


감지
• 멀티모달 감지
• 멀티 도메인 도메인 감지
• 해석 가능한 감지기
• 일반화된 감지기
생성
• 새로운 생성 아키텍처
• 제어 가능한 생성
• 기합성 콘텐츠에 대한 신뢰 메트릭
• 선한 목적의 생성 응용 프로그램
Generative Dimensional Climb
1D
2D
(2+1)D
3D
(3+1)D
출처
• 모델 및 콘텐츠 워터마킹
• 출처 탐지
• GANs 차단
Prompt
No watermark
With watermark
다음은?
각 벡터마다 수십 개의 새로운 연구가 진행중
intel
51


더 많은 정보는 아래 참조
Ilke Demir
http://ilkedemir.github.io
ilke.demir@intel.com
@ilkedemir
#FakeCatcher #deepfakedetection
#TrustedMedia #DDaaS
딥페이크 데이터 세트:
http://bit.ly/FakeCatcher
intel
학술 협력자 :
Umur Aybars Ciftci
uciftci@binghamton.edu
THANKS TO
TRUSTED MEDIA TEAM!

Session 1 디지털 혁신 속 저작권 보호 기술

III 30년간 비가시성 워터마크를 연구한 기업이 말해주는 "다양한 비가시성 워터마크 활용" 및 "생성형 AI 콘텐츠 위한 고속 비가시성 워터마킹 기술" 이야기

최고
마크애니 대표

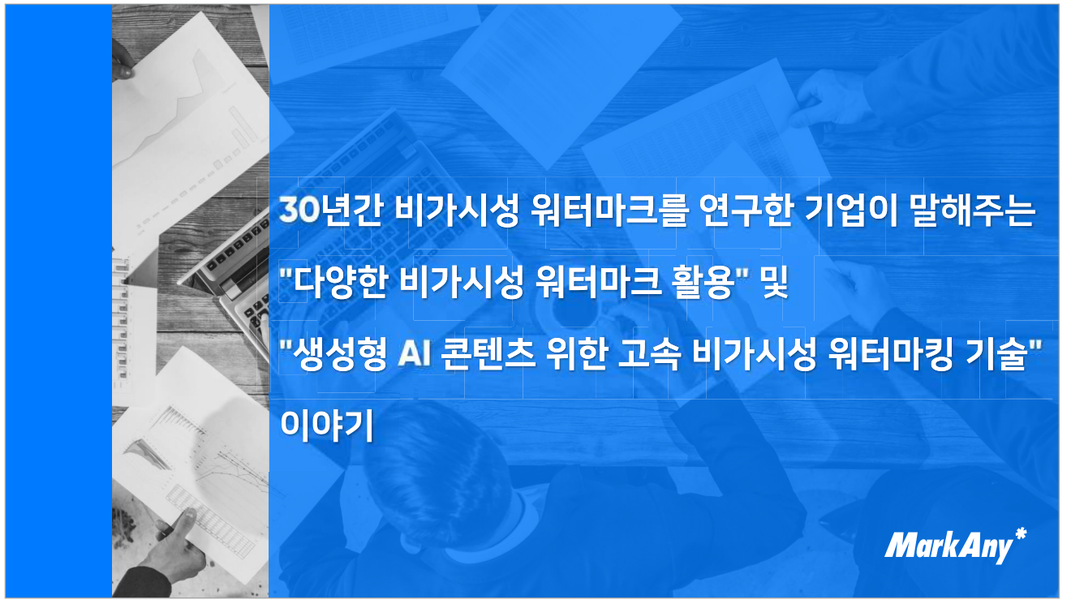
연사 이력

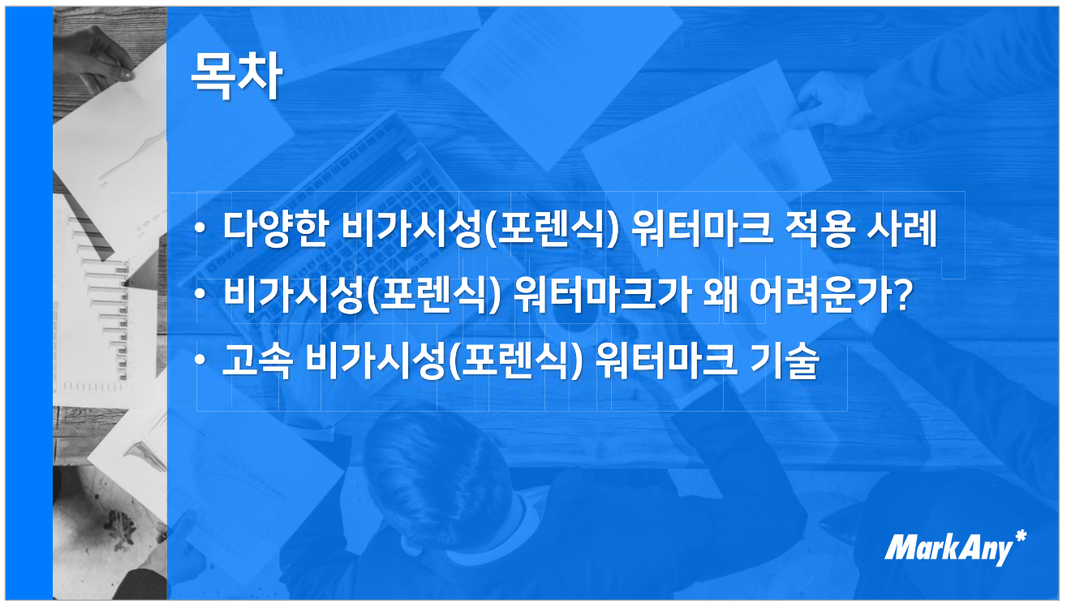
- 마크애니 대표이사(2018~)
- Mobvoi 사업개발 임원(2017~2018)
- DocuSign 마케터(2016)
- 삼성 SDS 책임(2013~2015)
- Woodall Tech CTO(2012~2013)

발표 내용

저작권 보호를 목적으로 탄생한 비가시성 워터마크의역할이 확대되고 있습니다. 기업 내부 정보 유출 방지를 위한 목적과 정품인증을 위한 목적으로도 활용되고 있고, 최근에는 생성형 AI로 인해 화두가 되고 있는 딥페이크에 대항하기 위한 용도로도 많이 거론되고 있습니다. 생성형 AI에 워터마크를 적용하기 위해선 고속 기술이 필요합니다. 이 세션을 통해 다양한 워터마크 활용 Use case를 소개하고, 생성형 AI에 적용하기 위해 필수적인 고속 기술에 대해 설명합니다.

The role of invisible watermark, originally designed for copyright protection, is expanding. For example, invisible watermarks are utilized to "Manage Insider Risk" and prove “Product Authenticity”. Today, it is frequently discussed as a defense mechanism against deepfake technology. This discussion explores how invisible watermarks are applied: protecting copyrights, safeguarding internal information, authenticating genuine products. In addition, we will highlight one of the most essential technologies required for applying watermarks in generative AI: high-speed watermarking.





WHO WE ARE

MarkAny

국가와 국민과 고객을 보호한다

창립
1999 (1995)

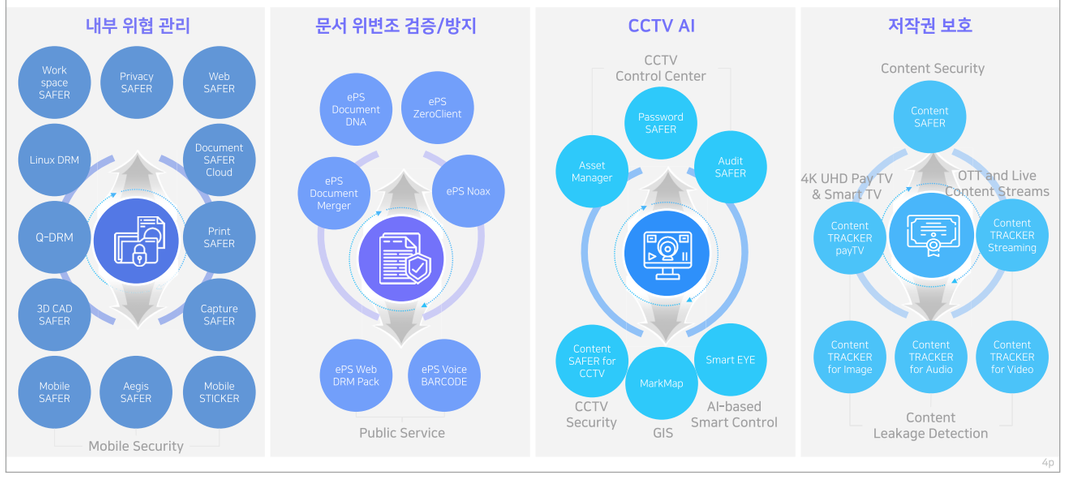
포렌식 워터마크 경력
30년

직원 수
220+

R&D 투자
연 매출의 20% / 500+ 특허 출원


Ssangrim building

사업 라인업 (정보보안 & AI)

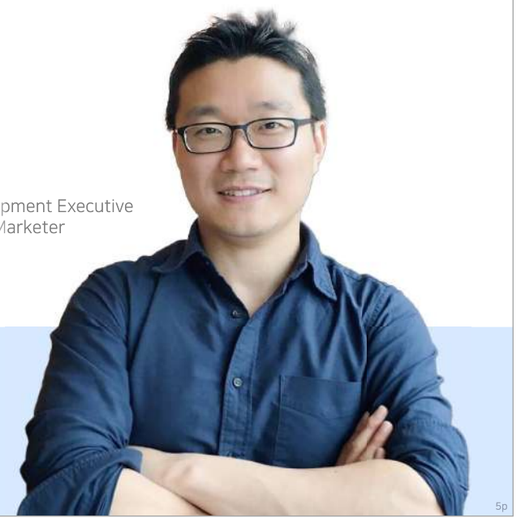


4p

발표자 소개

- 하버드 (Harvard) MBA
- 퍼듀 (Purdue) 전자공학 박사
- 퍼듀 (Purdue) 전자공학 학사, 응용물리 학사

- 마크애니 대표이사(현재)
- Mobvoi (구글에서 3,600억원 투자한 AI 기업) Business Development Executive
- DocuSign (세계 전자서명 70%, 미국 95% 마켓셰어) Product Marketer
- 삼성 SDS 사업기획실 책임
- WoodallTech (베터리 스타트업) R&D Director & Co-Founder



MarkAny CEO/CSO Go Choi

5p

01 사례

콘텐츠 저작권 보호 (이미지, 동영상, 음원 + 문서)

Use case (동영상)

- 동영상을 직접 리핑하거나 화면을 모바일폰/캠코더로 촬영
- 리핑하거나 촬영한 동영상을 Dark web 혹은 웹하드에 유출
- 유출 된 동영상을 포렌식 워터마크 검출 SaaS에 업로드
- 포렌식 워터마크 검출 SaaS에서 동영상에 숨겨진 포렌식 워터마크 메시지를 분석
- 포렌식 워터마크 검출 SaaS 분석 결과로 유출자 ID를 출력

6p

01 사례

콘텐츠 저작권 보호 (이미지, 동영상, 음원 + 문서)

유출한 End-user 추적

•OTT, 웹툰, 음악 streaming, 스톡/셀럽 이미지, 이력서/개인정보, 교육 플랫폼

유출한 협력사 추적

•음악 스튜디오, 엔터테인먼트 회사 (뮤직비디오), 브랜드 회사

7p

01 사례

내부 위협 관리 (모니터 촬영 및 캡쳐를 통한 유출)

모니터 촬영 및 캡쳐 통한 정보 유출

• 촬영 행위 자체를 근본적으로 막을 수 있는 방안은 없음
• 개인 폰 혹은 카메라로 촬영을 하였을 때 촬영자가 누구인지 알 수 있는 방법이 없음

재택/원격 근무 환경 제어 불가

• 거듭된 팬데믹과 MZ세대의 니즈를 반영하여 재택/원격 근무 환경은 늘어날 것으로 예상
• 주로 회사 PC 대신 개인 PC로 접속하여 업무를 할 수 있도록 함
• 집에서 핸드폰 혹은 카메라로 모니터를 촬영 하더라도 사실 여부를 확인 할 수 있는 증인도 없고, 당시 현장을 촬영한 CCTV도 없음

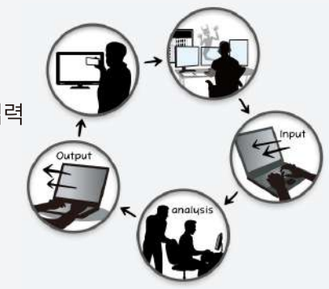
8p

01 사례

포렌식 워터마크 통한 유출자 추적

Use case

•모니터를 개인 폰 혹은 카메라로 촬영
•촬영한 영상 혹은 이미지를 블라인드 혹은 외부에 전달
•유출 된 영상 혹은 이미지를 포렌식 워터마크 검출 소프트웨어에 입력
•포렌식 워터마크 검출 소프트웨어가 영상 혹은 이미지에 숨겨진 포렌식 워터마크 메시지를 분석
•포렌식 워터마크 검출 소프트웨어가 분석 결과로 유출자 ID를 출력



9p

01 사례

프린트 문서 적용

Use case

•모니터 보안 케이스와 작동원리는 동일하나, 모니터에 Overlay로 포렌식 워터마크를 입력하는 대신 프린트 시 프린트하는 문서에 적용

•어려움: 이론적으로 모든 프린터에서 동일하게 작동해야 하나, 실제로 잘 작동하는지 전세계 수십만개가 넘는 프린터 모델들을 테스트하기는 어려움

10p

01 사례

딥페이크

- 포렌식 워터마크 의무화 논의/추진 중
  - •미국 캘리포니아
  - •EU
- 주의점!
  - •고속 워터마크 기술... (뒤에서 설명)

11p

02 어떻게 해결하나?

SaForus Learning partners

- 7 Learning partners
  - •콘텐츠 관리/배포/공유 플랫폼 3
  - •SNS 플랫폼 1
  - •교육 플랫폼 1
  - •의류 브랜드 1
  - •E-commerce 플랫폼 1

더 많은 Learning Partner들과 함께 하고 싶습니다

13p

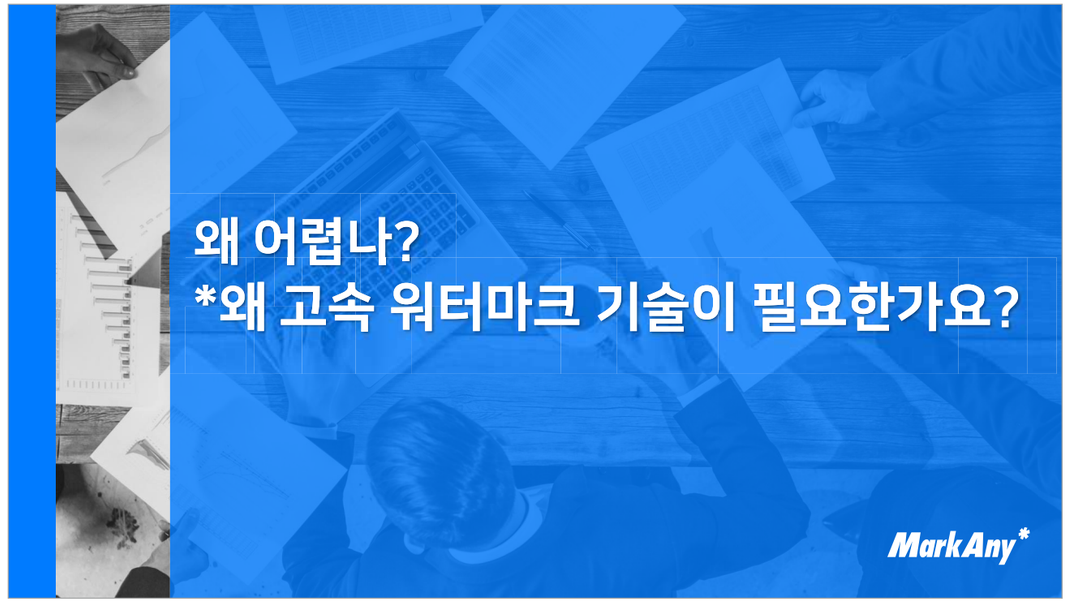
02 어떻게 해결하나?

SaForus / SecuAlign

- Pain point
  - •콘텐츠 무단 사용
  - •오리지날 제작자 증명 불가

너무 비싸다!!!!

12p



03 왜 어렵나?

왜 어렵나?

원리는 쉬운데 잘 하기는 왜 어렵나?

- 빛 산란, 각도, 흔들림, 편집 등으로 인한 "공격"

▼ 영상의 흔들림 샘플

마크애니는 왜 잘 하는가?

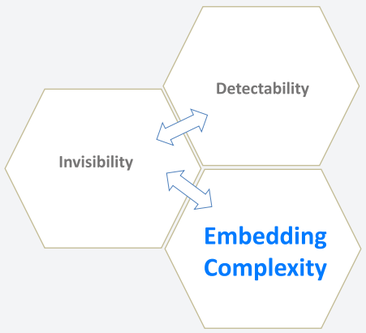
- 창업부터 꾸준히 포렌식 워터마크 연구
- 할리우드 6대 스튜디오가 인정한 세계 6개 기업 중 1
  - ✓ 총 97개 공격 테스트 해야 인정
  - ✓ 유럽 3 (실제로 1개 업체는 마크애니 라이선스 사용)
  - ✓ 미국 2
  - ✓ 한국 1

▼ 영상의 색변조 샘플

15p

03 왜 어렵나?

생성형 AI에서는 왜 더 어렵나?



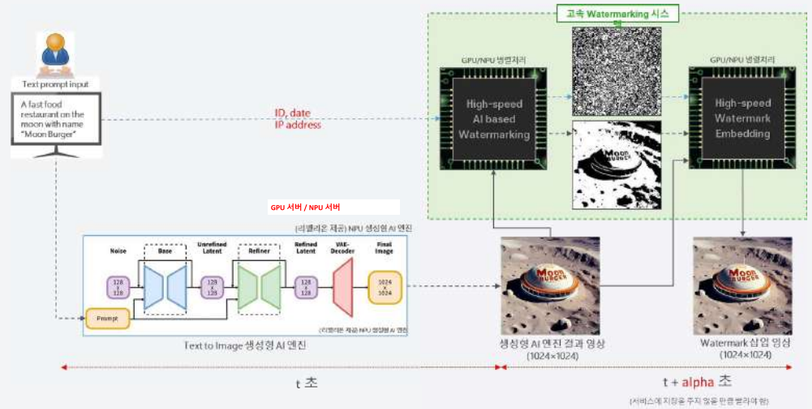
생성형 AI Watermarking

- Watermark 적용 시간이 Zero로 수렴
- 적용 시간이 너무 오래 걸리면 Text 입력 후 이미지, 동영상, 오디오 생성까지 너무 오래 시간이 걸림 (비용, UX 문제)

16p

04 고속 워터마크 기술

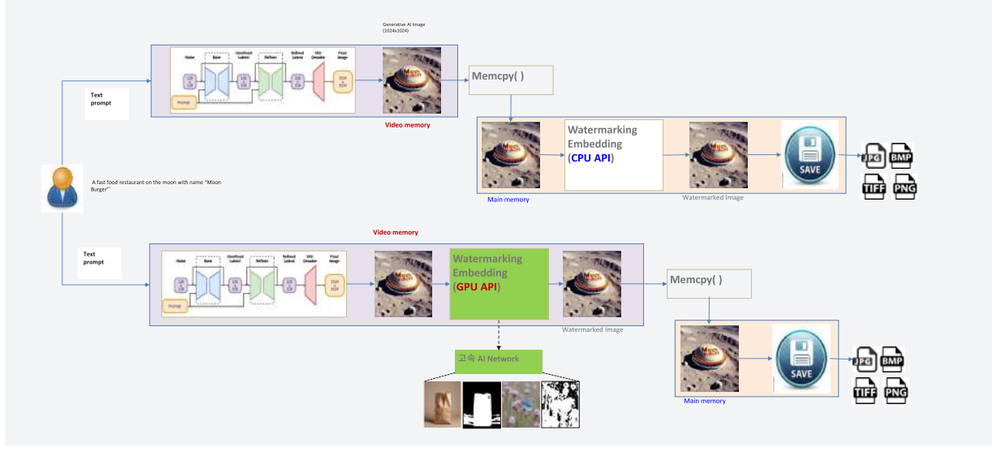
생성형 AI 고속 포렌식 워터마크 시스템(GPU/NPU)



18p

04 고속 워터마크 기술

생성형 AI 고속 포렌식 워터마크 시스템(GPU/NPU)



19p

04 고속 워터마크 기술

생성형 AI 고속 포렌식 워터마크 시스템(GPU/NPU)

	Screen watermarking (low CPU)	**Image** watermarking (CPU)	**OTT Video** Watermarking (CPU)	**Image AI** **Video AI** Watermarking (Single GPU)	**Video AI + Audio** watermarking (Single GPU)	**Video AI Watermarking** for Generative AI Image/video
Embedding Throughput	480 FPS @ FHD 120 FPS @ UHD	40 FPS @ FHD	40 FPS @ FHD	800 FPS@ FHD		Over thousands FPS @ FHD
	Realtime on Screen	1.5M 웹툰 영상 10장에 0.3초 소요 4M 영상 10장에 0.5초 소요	30FPS/FHD 1분 동영상에 45초 소요 2시간 영화 1편에 1.5 시간 소요	1.5M AI 이미지 10장에 0.009초 소요 4M 이미지 10장에 0.025초 소요 30FPS/FHD 1분 동영상에 2.25초 소요 FHD 1시간 동영상에 1분 15초 소요		다중 GPU를 사용하는 만큼 추가 시간 단축 가능
정지영상 시인성	Very little	None	None	None		None
동영상 시인성	Little	Very little	None	None		None
	 Desktop/CPU	 Desktop/CPU		 Desktop /Single GPU		 Clouding Multi GPU/NPU

20p

Thank you!

MarkAny

21p

Session 1 디지털 혁신 속 저작권 보호 기술

Ⅳ 저작물 방송 송출을 위한 콘텐츠 보안 적용

로날드 휠러
A3SA 전무이사

연사 이력

- A3SA 전무이사 (2019년 8월~현재)
- 영화 라이선스 회사(MPLC) 사업 및 법무 부문 총괄 (2019년 10월~현재)
- 20세기 폭스 콘텐츠 보호 및 기술 전략 부문 수석 부사장 (2017년 2월~2019년 7월)
- 폭스 그룹 법무부 콘텐츠 보호 부문 수석 부사장 (2001년 10월~2017년 1월)
- 20세기 폭스 법무 부문 부사장 (1998년 10월~2001년 9월)
- 20세기 폭스 선임 고문 (1997년 10월~1998년 9월)
- 20세기 폭스 고문 (1994년 10월~1997년 9월)
- 웨일 고트샬 앤 망게스 선임 변호사 (1989년 4월~1994년 9월)
- 셔먼 앤 스털링 변호사 (1984년 9월~1989년 3월)

발표 내용

시청자들은 콘텐츠의 "황금기"를 살고 있지만, 동시에 신호 도용, 해킹, 그리고 불법 복제가 그 어느 때보다 심각한 시기를 겪고 있습니다. 오늘날 여전히 보안 경고를 발생시키는 많은 보안 취약 웹사이트처럼, ATSC 1.0 및 기타 암호화되지 않고 서명이 없는 방송 서비스는 이러한 위협에 취약합니다. 그러나 새로운 NEXTGEN TV ATSC 3.0 방송 표준은 신호 서명과 관련하여 필수적이며, 암호화와 관련하여 사용자를 신호 도용, 해킹, 그리고 불법 복제로부터 보호하는 보안 조치를 허용합니다. ATSC 3.0 보안 기관(A3SA)은 이러한 보호 조치를 ATSC 3.0 방송 서비스에서 가능하게 하고 운영함으로써, 방송 TV 콘텐츠와 시청자가 이전보다 더욱 안전해지도록 합니다.

Viewers are living in a “golden age” of content, but they are also living in an unparalleled period of signal theft, hacking and piracy. Like the many insecure websites that still exist today and trigger security warnings when accessed, ATSC 1.0 and other unencrypted, unsigned broadcast services are vulnerable to these threats. But the new NEXTGEN TV ATSC 3.0 broadcast standard requires (with respect to signal-signing) and allows (with respect to encryption) the same security measures that protect users of newer, more secure websites from signal theft, hacking and piracy. And the ATSC 3.0 Security Authority (A3SA) enables and operationalizes those protections in ATSC 3.0 broadcast services, thus enabling both broadcast TV content and its viewers to be more secure than ever before.

What Happens Every Day on the Internet: Security

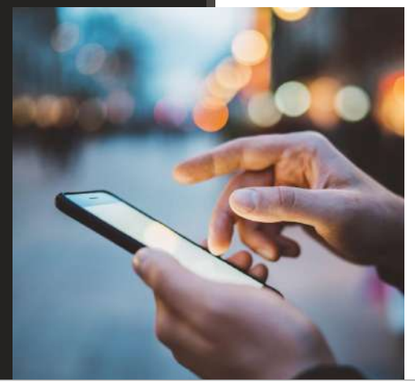
Content security has become fundamental to today's internet video services. It's demanded by content providers, relied upon by--but invisible to--consumers, and well-accepted by device/app providers

Key Technological Mechanisms:

Web Browsers use digital signatures to authenticate websites and DRM encryption to secure communications between web browsers and servers

App stores secure apps and app delivery through digital signatures

Video streaming apps secure content during transmission via DRM encryption, including streamed content that is free to view



The ATSC 3.0 Standard Allows OTA Broadcasters to Offer Content Providers Internet-Style Content Security for the First Time—both Online and Offline

As with Internet streaming services, both content providers and viewers benefit from the improved trustworthiness of the OTA broadcast distribution channel. But A3SA offers an additional benefit Internet streaming services do not: the ability for most receivers to decrypt encrypted content even when offline (called "Unconnected Mode").

Signal Signing
ensures that the signal being received is from a government-licensed broadcaster and the content received has not been tampered with

Broadcast Application Signing
prevents rogue malware from loading and executing on ATSC 3.0 receivers

Content Security
utilizes DRM technology similar to that used by internet content services, including content that is free to view

인터넷에서 매일 일어나는 일: 보안

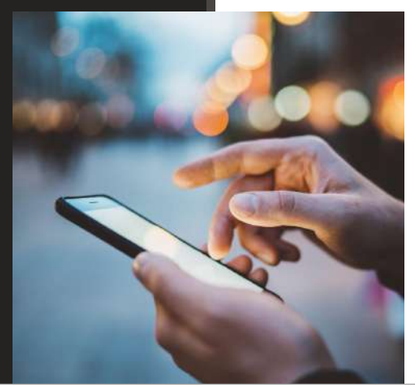
오늘날 인터넷 비디오 서비스에서 콘텐츠 보안은 필수적이다. 콘텐츠 제공자가 요구하며, 소비자는 이에 의존하지만 보이지 않는 존재인 콘텐츠 보안은 장치/앱 제공자에게 널리 수용되고 있다.

핵심 기술 메커니즘 :

웹 브라우저는 디지털 서명을 사용하여 웹사이트를 인증하고 DRM 암호화를 사용하여 웹 브라우저와 서버 간 통신을 보호한다.

앱 스토어는 디지털 서명을 통해 앱을 보호하고 앱을 제공한다.

비디오 스트리밍 앱은 DRM 암호화를 통해 전송 중에 콘텐츠를 보호한다. 여기에는 무료로 볼 수 있는 스트리밍 콘텐츠도 포함된다.



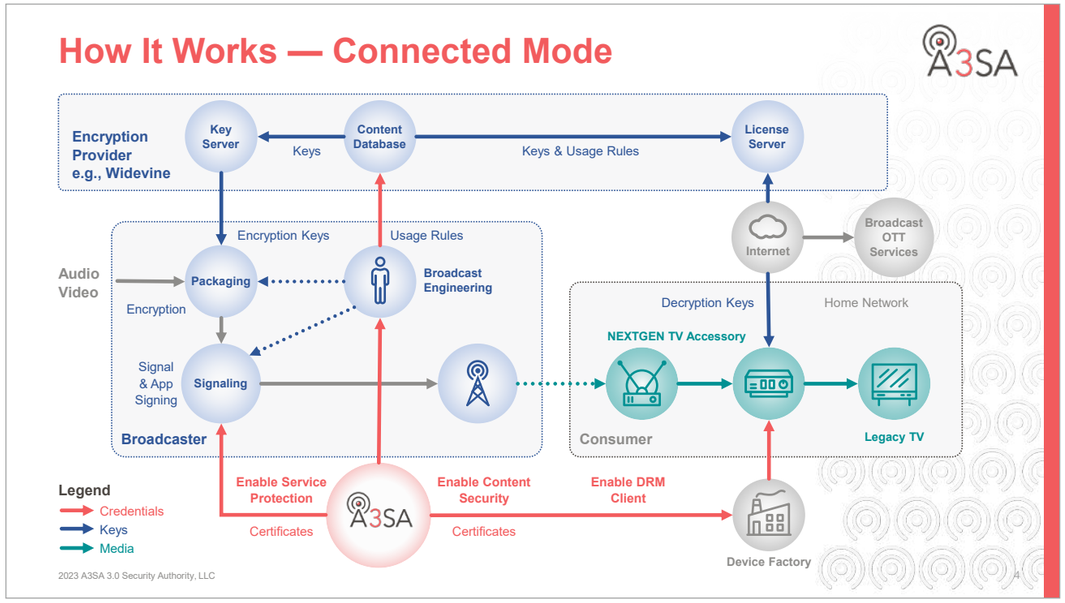
ATSC 3.0 표준은 OTA 방송사가 콘텐츠 제공자에게 인터넷 스타일의 콘텐츠 보안을 온/오프라인으로 모두 제공할 수 있게 한다.

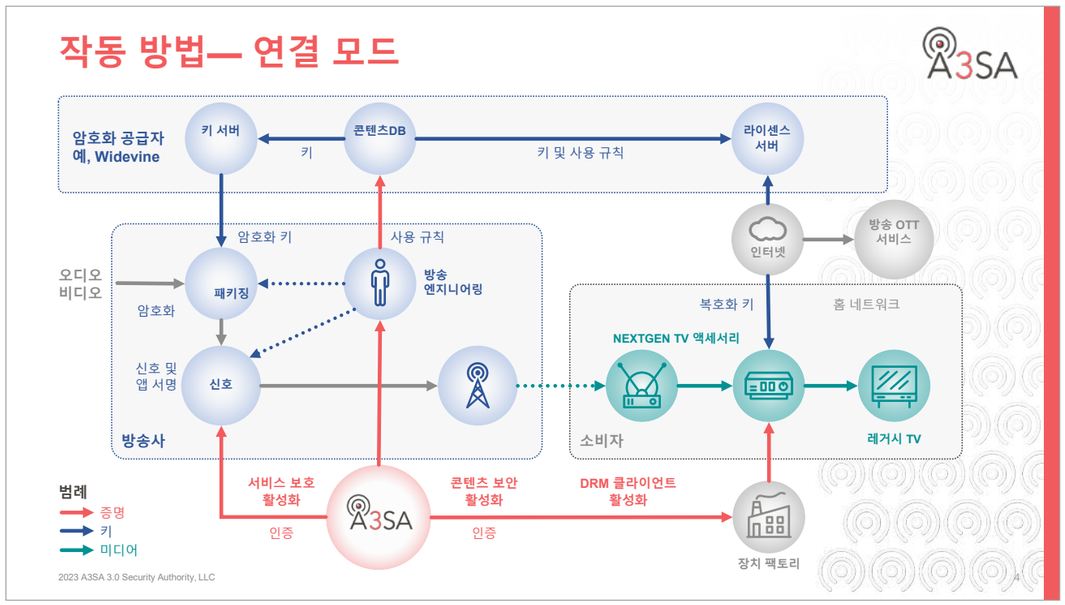
인터넷 스트리밍 서비스와 유사하게, 콘텐츠 제공자와 시청자는 OTA 방송 채널의 신뢰성 향상으로 이익을 얻는다. 또한 A3SA는 인터넷 스트리밍 서비스에서는 제공되지 않는 추가 이점으로, 대부분의 수신기가 오프라인에서도 암호화된 콘텐츠를 해독할 수 있는 기능("연결되지 않은 모드")을 갖추게 된다.

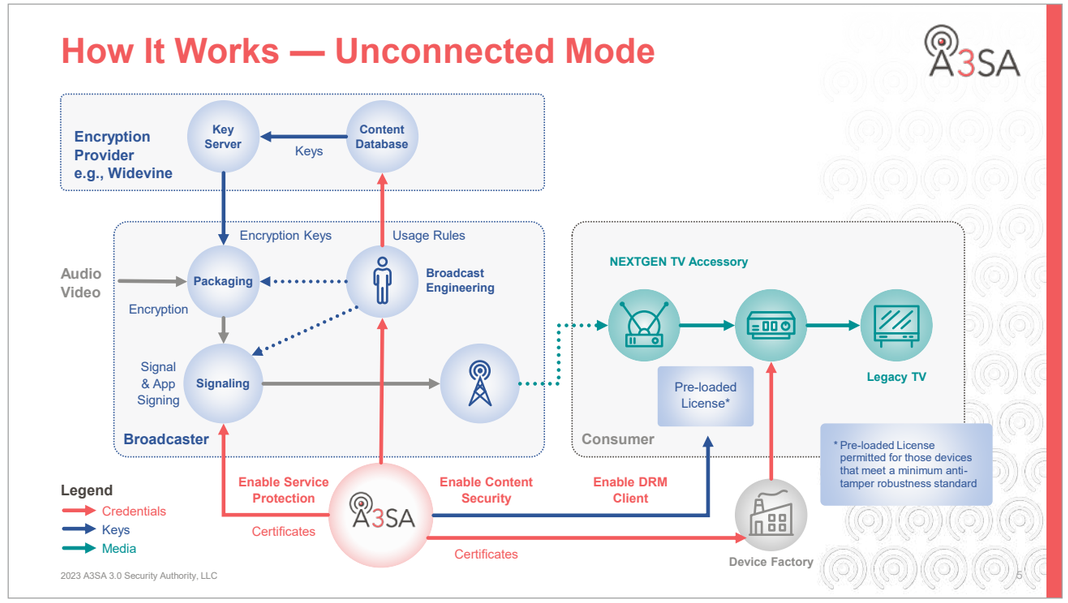
신호 사인
신호가 정부 라이센스를 가진 방송사로부터 왔고, 수신된 콘텐츠가 변경되지 않았음을 확인

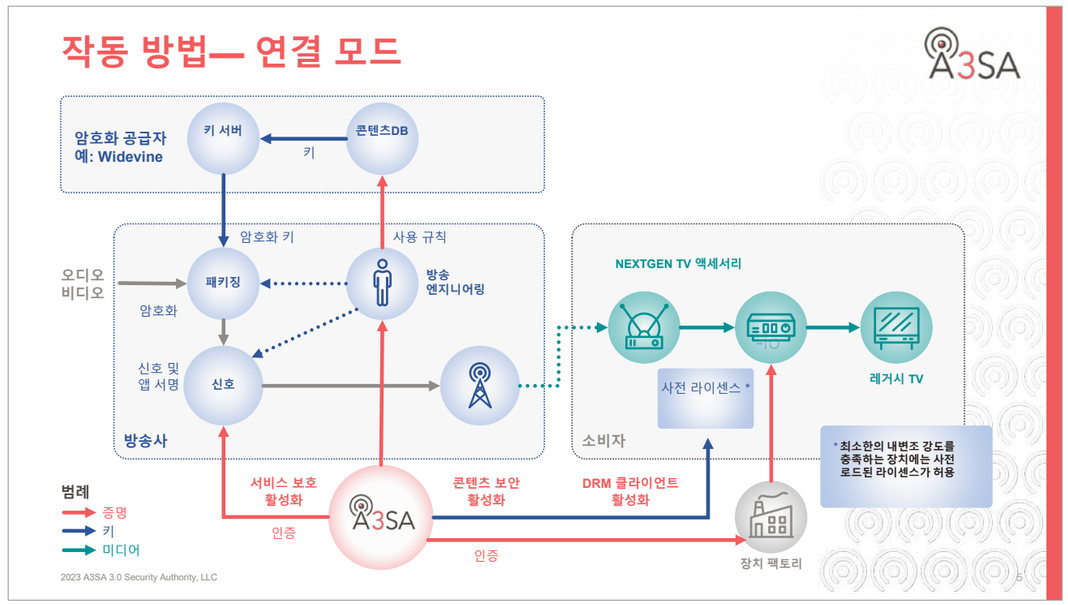
방송 애플리케이션 사인
악성 코드가 ATSC 3.0 수신기에 로딩되고 실행되는 것을 방지

콘텐츠 보안
무료 콘텐츠를 포함하여 인터넷 콘텐츠 서비스에서 사용되는 것과 유사한 DRM 기술 활용


How It Works — Connected Mode
A3SA
Encryption Provider e.g., Widevine
Key Server
Content Database
License Server
Keys
Keys & Usage Rules
Internet
Broadcast OTT Services
Encryption Keys
Usage Rules
Audio Video
Packaging
Broadcast Engineering
Encryption
Signal & App Signing
Signaling
Broadcaster
Decryption Keys
Home Network
NEXTGEN TV Accessory
Consumer
Legacy TV
Legend
Credentials
Keys
Media
Enable Service Protection
Certificates
Enable Content Security
Certificates
Enable DRM Client
Device Factory
2023 A3SA 3.0 Security Authority, LLC
4


작동 방법— 연결 모드
A3SA
암호화 공급자 예, Widevine
키 서버
콘텐츠DB
라이센스 서버
키
키 및 사용 규칙
인터넷
방송 OTT 서비스
암호화 키
사용 규칙
오디오 비디오
패키징
방송 엔지니어링
암호화
신호 및 앱 서명
신호
방송사
복호화 키
홈 네트워크
NEXTGEN TV 액세서리
소비자
레거시 TV
범례
증명
키
미디어
서비스 보호 활성화
인증
콘텐츠 보안 활성화
인증
DRM 클라이언트 활성화
장치 팩토리
2023 A3SA 3.0 Security Authority, LLC
4


How It Works — Unconnected Mode
A3SA
Encryption Provider e.g., Widevine
Key Server
Content Database
Keys
Encryption Keys
Usage Rules
Audio Video
Packaging
Broadcast Engineering
Encryption
Signal & App Signing
Signaling
Broadcaster
NEXTGEN TV Accessory
Pre-loaded License*
Legacy TV
Consumer
* Pre-loaded License permitted for those devices that meet a minimum anti-tamper robustness standard
Legend
Credentials
Keys
Media
Enable Service Protection
Certificates
Enable Content Security
Enable DRM Client
Certificates
Device Factory
2023 A3SA 3.0 Security Authority, LLC
5


작동 방법— 연결 모드
A3SA
암호화 공급자 예: Widevine
키 서버
콘텐츠DB
키
암호화 키
사용 규칙
오디오 비디오
패키징
방송 엔지니어링
암호화
신호 및 앱 서명
신호
방송사
NEXTGEN TV 액세서리
사전 라이센스 *
레거시 TV
소비자
* 최소한의 내변조 강도를 충족하는 장치에는 사전 로드된 라이센스가 허용
범례
증명
키
미디어
서비스 보호 활성화
인증
콘텐츠 보안 활성화
DRM 클라이언트 활성화
인증
장치 팩토리
2023 A3SA 3.0 Security Authority, LLC
5

Benefits of Content Security for Content Providers and Broadcasters

For Content Providers
Deters piracy of individual movies and TV shows

For Broadcasters
Deters unauthorized redistribution of channels and channel content

For Both Content Providers and Broadcasters
Opens new distribution channel for secure distribution of high value content

콘텐츠 제공자 및 방송사를 위한 콘텐츠 보안 장점

콘텐츠 제공자
개별 영화 및 TV 쇼의 불법 복제를 억제

방송사
채널 및 채널 콘텐츠의 무단 재배포를 방지

콘텐츠 제공자와 방송사
고부가가치 콘텐츠의 안전한 유통을 위한 새로운 유통 채널 오픈

Content Security vs. Service Protection in ATSC 3.0

Content Security
--Encrypts content using DRM technology
--Protects against unauthorized redistribution
--Issues and applies licenses and cryptographic keys
--Optional – underlying technology specified in A/360

Service Protection
--Issues & validates Digital Certificates
--Protects against spoofing, hacking, signal intrusion
--Allows receivers to verify that the broadcast signals, content and apps were broadcast by a trusted broadcaster and have not been changed
--Required - Specified in A/360 and A/331 and required for signing broadcast signals & apps

ATSC 3.0의 콘텐츠 보안 vs 서비스 보호

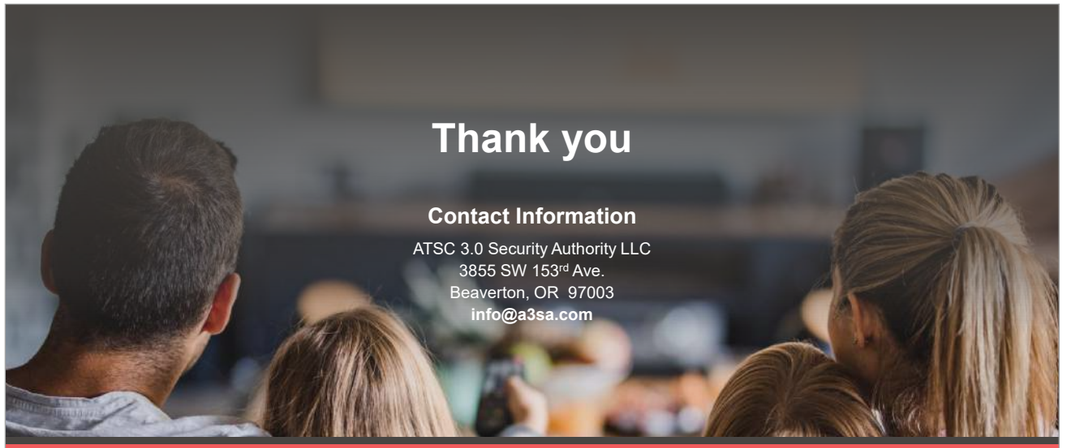
콘텐츠 보안
--DRM 기술을 사용하여 콘텐츠 암호화
--무단 재배포 방지
--라이선스 및 암호화 키 발급 및 적용
--선택 사항
- A/360에 지정된 기본 기술

서비스 보호
--디지털 인증서 발급 및 검증
--스푸핑, 해킹, 신호 침입으로부터 보호
--수신자가 신뢰할 수 있는 방송사에서 방송된 방송 신호, 콘텐츠 및 앱이변경되지 않았는지 확인할 수 있도록 함
--필수 - A/360 및 A/331에 지정되어 있으며 방송 신호 및 앱 서명에 필요함

Broadcast Encoding Rules

Balancing content security with user flexibility, A3SA has approved a set of “encoding rules” for encrypted broadcasts that are simulcast with ATSC 1.0 broadcasts

- Viewers must be allowed to decrypt and record these broadcasts even if they are using a less secure device that requires an internet connection
- Viewers must be allowed to make an unlimited number of copies of these broadcasts
- Such copies cannot have retention limits
- Viewers must be allowed to use “trick play” features such as pause, rewind, fast-forward, and ad-skipping
- Viewers must be allowed to use any authorized digital output (i.e., no selectable output control)
- Viewers must be allowed to use analog outputs to connect to legacy TVs (i.e., no prohibition or required down-resolution)



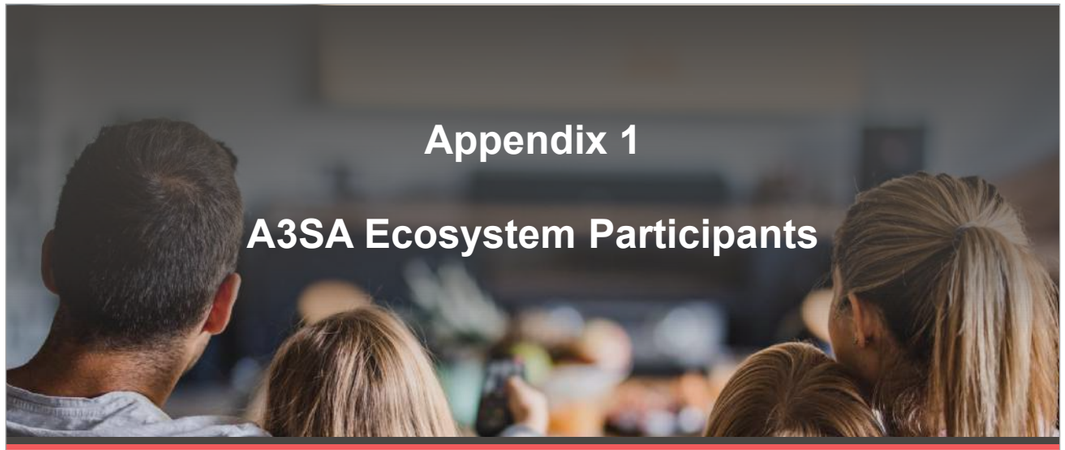
A3SA 3.0 Security Authority, LLC November 06, 2024

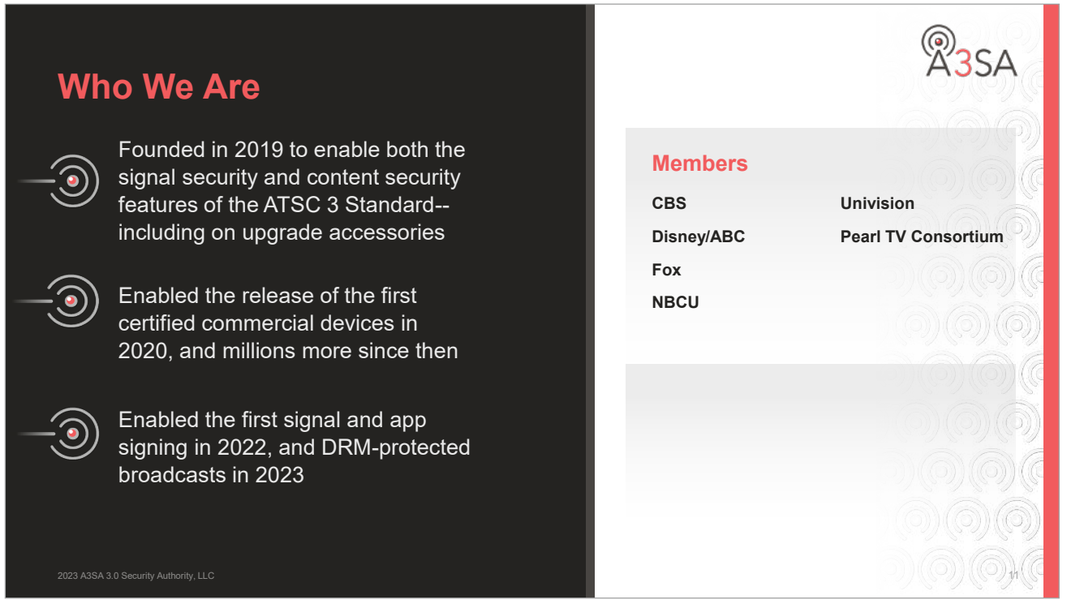
브로드캐스트 인코딩 규칙

A3SA는 콘텐츠 보안과 사용자 편의성을 조화시키기 위해 ATSC 1.0 방송과 동시 송출되는 암호화된 방송에 대한 "인코딩 규칙"을 승인함.

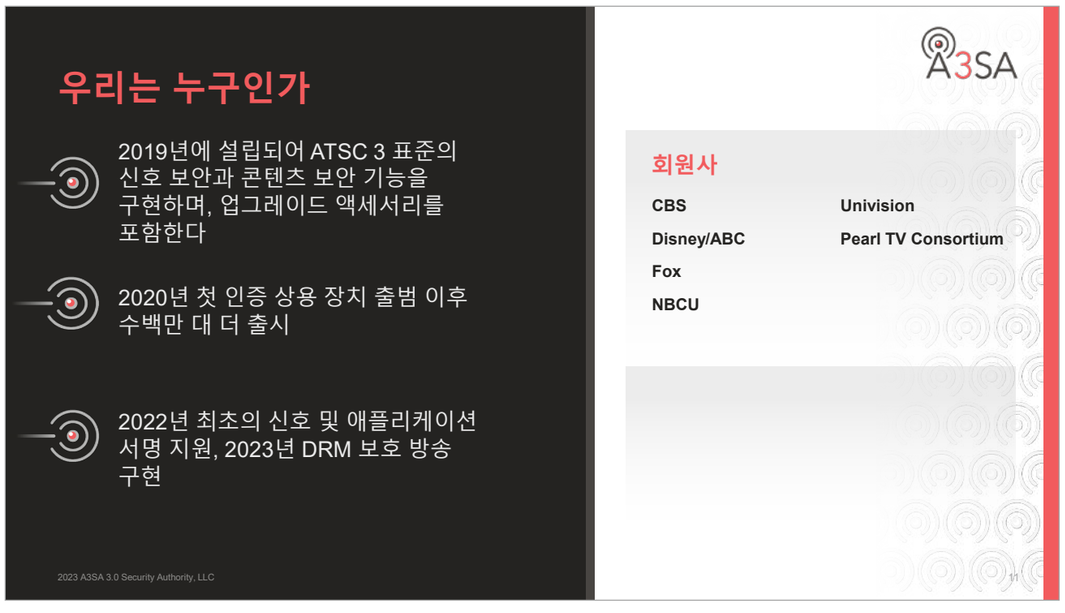
- 시청자는 보안이 낮은 장치를 사용하더라도 인터넷 연결이 필요할 경우 이 방송을 복호화하고 녹화할 수 있어야 한다.
- 시청자는 이러한 방송의 복사본을 무제한으로 만들 수 있어야 한다.
- 복사본은 보존 기간 제한이 없어야 한다.
- 시청자는 일시 정지, 되감기, 빨리 감기, 광고 건너뛰기와 같은 "트릭 플레이" 기능을 자유롭게 사용할 수 있어야 한다.
- 또한, 시청자는 승인된 모든 디지털 출력을 사용할 수 있어야 한다(선택 가능한 출력 제어는 없음),
- 아날로그 출력을 사용하여 구형 TV에 연결할 수 있어야 한다(선택 가능한 출력 제어는 없음),

A3SA 3.0 Security Authority, LLC November 06, 2024


Appendix 1
A3SA Ecosystem Participants
A3SA™
A3SA 3.0 Security Authority, LLC November 06, 2024


Who We Are
Founded in 2019 to enable both the signal security and content security features of the ATSC 3 Standard--including on upgrade accessories
Enabled the release of the first certified commercial devices in 2020, and millions more since then
Enabled the first signal and app signing in 2022, and DRM-protected broadcasts in 2023
2023 A3SA 3.0 Security Authority, LLC
A3SA
Members
CBS
Disney/ABC
Fox
NBCU
Univision
Pearl TV Consortium


부록 1
A3SA 생태계 참여자
A3SA™
A3SA 3.0 Security Authority, LLC November 06, 2024


우리는 누구인가
2019년에 설립되어 ATSC 3 표준의 신호 보안과 콘텐츠 보안 기능을 구현하며, 업그레이드 액세서리를 포함한다
2020년 첫 인증 상용 장치 출범 이후 수백만 대 더 출시
2022년 최초의 신호 및 애플리케이션 서명 지원, 2023년 DRM 보호 방송 구현
2023 A3SA 3.0 Security Authority, LLC
A3SA
회원사
CBS
Disney/ABC
Fox
NBCU
Univision
Pearl TV Consortium

Who We Are

The A3SA's Technical Contributors Working Group (TCWG) provides a forum for existing and future participants in the ATSC 3 ecosystem to contribute to the development of the ecosystem

- Receiver manufacturers
- Broadcasters
- Security vendors
- Professional broadcast equipment manufacturers
- Technical solution providers

TCWG Participants

BitRouter

CBS

DigiCAP Co., Ltd

Disney

DTV Innovations, LLC

Fox

Gray Media Group, Inc.

Inca Networks Incorporated dba WISI America

LG Electronics U.S.A., Inc.

NBCUniversal

Nuvyyo, Inc

Pearl TV LLC

Samsung Electronics Co. Ltd.

Sinclair Broadcast Group, Inc.

Sony Electronics Inc.

Tolka Telecommunications Corporation

우리는 누구인가

A3SA 기술 기여자 작업 그룹(TCWG)은 ATSC 3 생태계의 현재 및 향후 참가자들에게 생태계 발전에 기여할 수 있는 공간 제공

- 수신기 제조업체
- 방송사
- 보안 공급업체
- 전문 방송 장비 제조업체
- 기술 솔루션 제공업체

TCWG 참여사

BitRouter

CBS

DigiCAP Co., Ltd

Disney

DTV Innovations, LLC

Fox

Gray Media Group, Inc.

Inca Networks Incorporated dba WISI America

LG Electronics U.S.A., Inc.

NBCUniversal

Nuvyyo, Inc

Pearl TV LLC

Samsung Electronics Co. Ltd.

Sinclair Broadcast Group, Inc.

Sony Electronics Inc.

Tolka Telecommunications Corporation

Adopter Licensees

Alticast, Inc.

BitRouter

DS Broadcast, Inc.

EiTV USA

Harmonic, Inc.

Hisense USA Corp

Inca Networks Incorporated dba WISI America

iWedia S.A.

LG Electronics USA Inc.

LowaSIS, Inc.

Mediaproxy Pty Ltd.

MediaTek Inc.

Mware Solutions BV

Nuvyyo, Inc.

Panasonic Entertainment & Communications Co., Ltd.

Samsung Electronics Co., Ltd.

Sencore, Inc

Shenzhen TCL New Technology Co., Ltd.

Shenzhen JingYue Times Technology Co., Ltd.

Shenzhen Zenview Intelligence Co., Ltd.

Silicondust USA, Inc.

Shift2Stream, Inc.

Sony Electronics Inc.

Tolka Telecommunications Corporation

Triveni Digital, Inc.

Vela Research LP

Zhuhai Gotech Intelligent Technology Co., Ltd.

Zinwell Corporation

Broadcaster Licensees

ABC, Inc

Allen Media Broadcasting, LLC

CBS Broadcasting Inc.

CMG Media Corporation

Fox Television Holdings, LLC

Graham Media Group, Inc

Gray Media Group, Inc.

Hearst Television, Inc

Meredith Corporation

NBCUniversal Media, LLC

Nexstar Media, Inc.

NPG of California, LLC

Scripps Media, Inc.

Sinclair Broadcast Group, Inc.

Sunbeam TV Corp.

TEGNA, Inc.

Univision Local Media, Inc

WPLG, Inc.

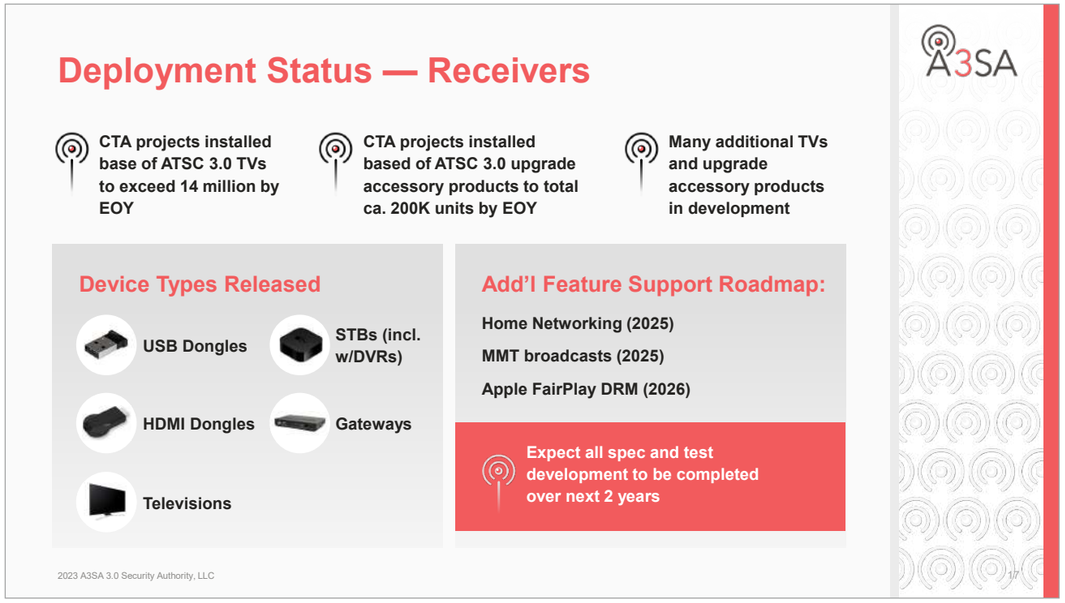
어답터 라이센스

Alticast, Inc.
BitRouter
DS Broadcast, Inc.
EiTV USA
Harmonic, Inc.
Hisense USA Corp
Inca Networks Incorporated dba WISI America
iWedia S.A.
LG Electronics USA Inc.
LowaSIS, Inc.
Mediaproxy Pty Ltd.
MediaTek Inc.
Mware Solutions BV
Nuvyyo, Inc.
Panasonic Entertainment & Communications Co., Ltd.
Samsung Electronics Co., Ltd.
Sencore, Inc
Shenzhen TCL New Technology Co., Ltd.
Shenzhen JingYue Times Technology Co., Ltd.
Shenzhen Zenview Intelligence Co., Ltd.
Silicondust USA, Inc.
Shift2Stream, Inc.
Sony Electronics Inc.
Tolka Telecommunications Corporation
Triveni Digital, Inc.
Vela Research LP
Zhuhai Gotech Intelligent Technology Co., Ltd.
Zinwell Corporation

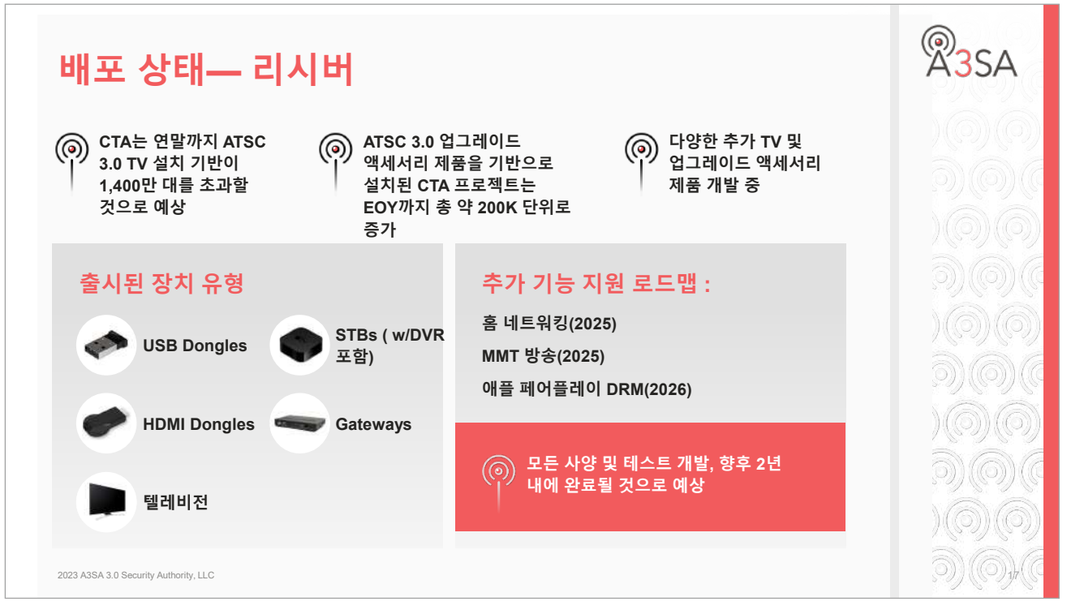
방송사 라이센스

ABC, Inc
Allen Media Broadcasting, LLC
CBS Broadcasting Inc.
CMG Media Corporation
Fox Television Holdings, LLC
Graham Media Group, Inc
Gray Media Group, Inc.
Hearst Television, Inc
Meredith Corporation
NBCUniversal Media, LLC
Nexstar Media, Inc.
NPG of California, LLC
Scripps Media, Inc.
Sinclair Broadcast Group, Inc.
Sunbeam TV Corp.
TEGNA, Inc.
Univision Local Media, Inc
WPLG, Inc.

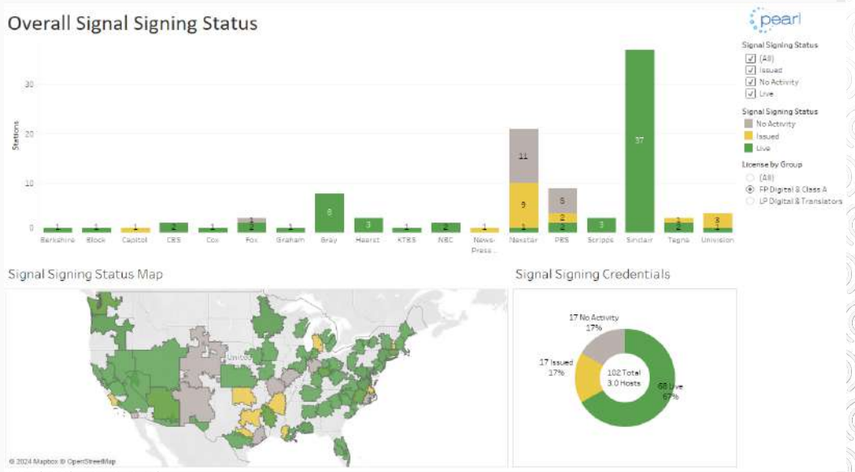

Appendix 2
A3SA Deployment Status
A3SA™
A3SA 3.0 Security Authority, LLC November 06, 2024


Deployment Status — Receivers
CTA projects installed base of ATSC 3.0 TVs to exceed 14 million by EOY
CTA projects installed based of ATSC 3.0 upgrade accessory products to total ca. 200K units by EOY
Many additional TVs and upgrade accessory products in development
Device Types Released
USB Dongles
STBs (incl. w/DVRs)
HDMI Dongles
Gateways
Televisions
Add'l Feature Support Roadmap:
Home Networking (2025)
MMT broadcasts (2025)
Apple FairPlay DRM (2026)
Expect all spec and test development to be completed over next 2 years
2023 A3SA 3.0 Security Authority, LLC
A3SA


부록 2
A3SA 배포 상태
A3SA™
A3SA 3.0 Security Authority, LLC November 06, 2024


배포 상태— 리시버
CTA는 연말까지 ATSC 3.0 TV 설치 기반이 1,400만 대를 초과할 것으로 예상
ATSC 3.0 업그레이드 액세서리 제품을 기반으로 설치된 CTA 프로젝트는 EOY까지 총 약 200K 단위로 증가
다양한 추가 TV 및 업그레이드 액세서리 제품 개발 중
출시된 장치 유형
USB Dongles
STBs (w/DVR 포함)
HDMI Dongles
Gateways
텔레비전
추가 기능 지원 로드맵 :
홈 네트워킹(2025)
MMT 방송(2025)
애플 페어플레이 DRM(2026)
모든 사양 및 테스트 개발, 향후 2년 내에 완료될 것으로 예상
2023 A3SA 3.0 Security Authority, LLC
A3SA

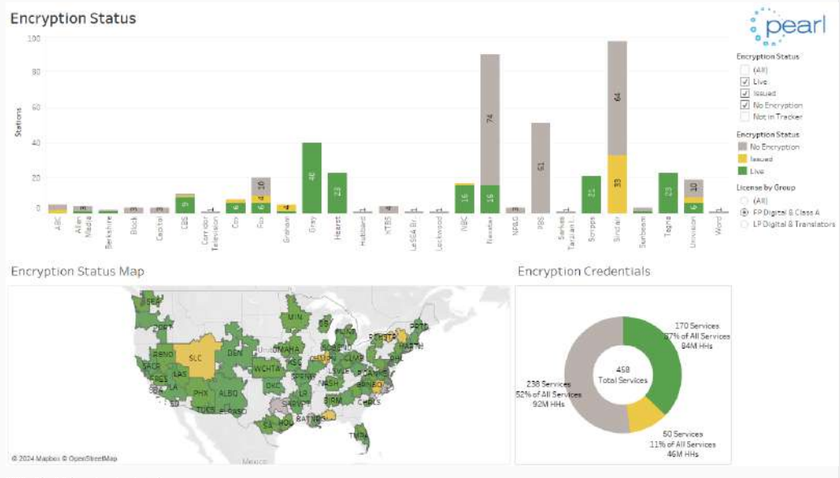
Deployment Status — Signal Signing



2023 A3SA 3.0 Security Authority, LLC

18

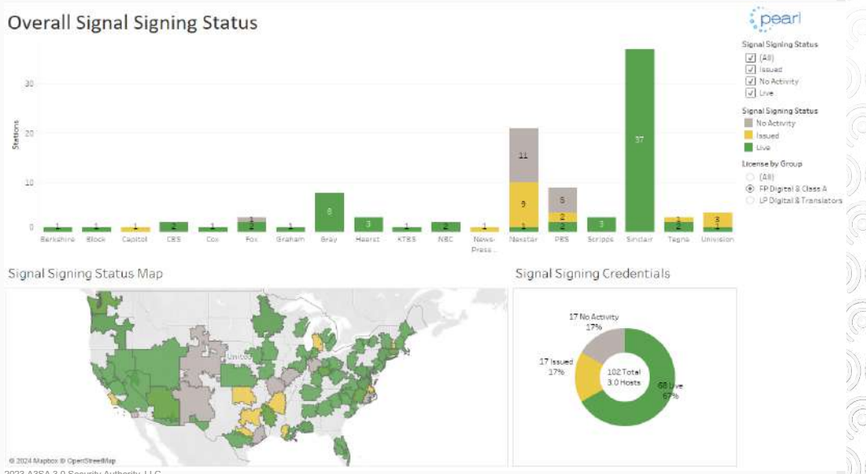
Deployment Status — Encryption



2023 A3SA 3.0 Security Authority, LLC

18

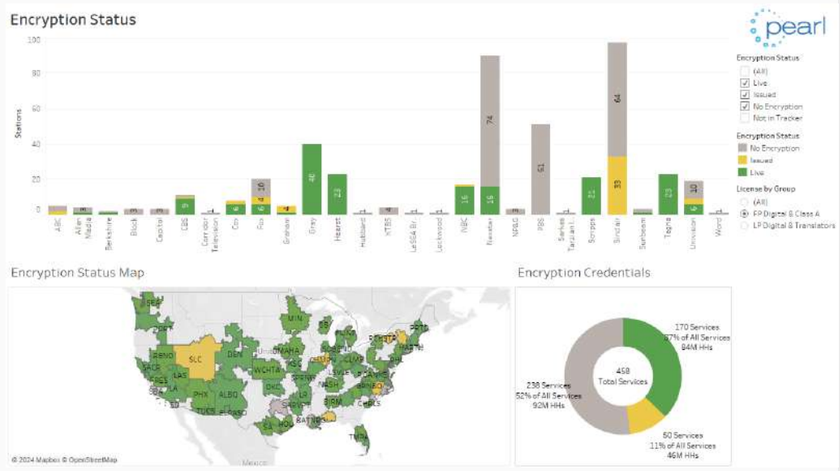
배포 상태 — 신호 서명



2023 A3SA 3.0 Security Authority, LLC

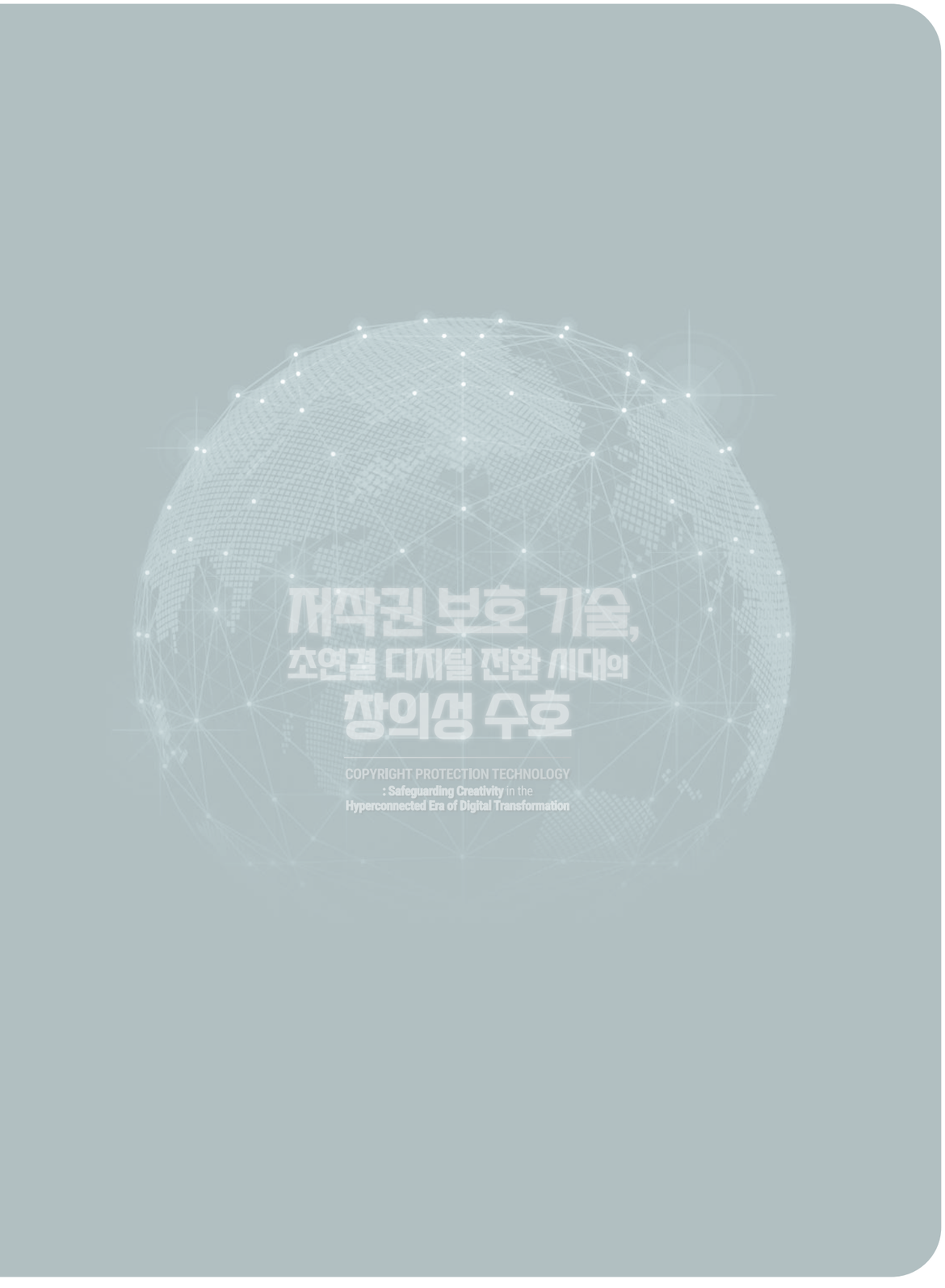
18

배포 상태 — 암호화



2023 A3SA 3.0 Security Authority, LLC

18



튜토리얼 세션

최신 인공지능기반 기술을 활용한 저작권 침해와 보호사례 및 AI 보안과 보호 이슈

우사이먼 성일 | 성균관대학교 교수

튜토리얼 세션

최신 인공지능기반 기술을 활용한 저작권 침해와 보호사례 및 AI 보안과 보호 이슈

우사이먼 성일
성균관대학교 교수

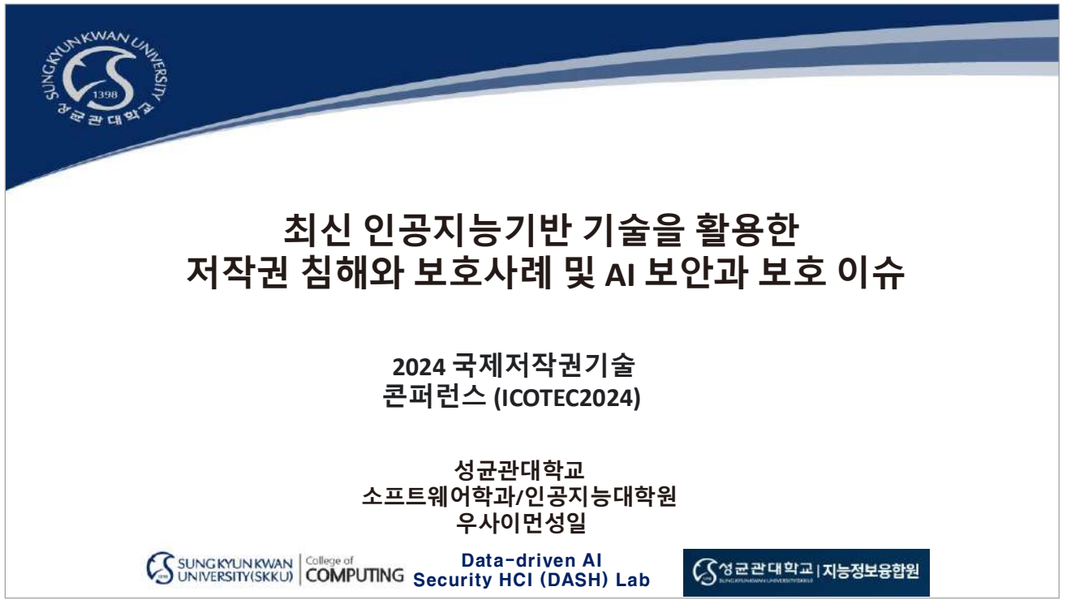
연사 이력

- 성균관대학교 소프트웨어학과/인공지능학과/데이터사이언스융합학과 부교수(2019~현재)
- 성균관대학교 데이터사이언스융합학과 학과장(2024~현재), SKKU Fellowship Professor (2023)
- 과기정통부 악의적 딥페이크 탐지 연구 프로젝트 책임자(2020-현재)
- 개인정보보호위원회 개인정보 기술 포럼 위원(2023-현재)
- 한국정보보호학회이사(2024-현재)
- 중앙선거관리위원회 사이버조사과 자문위원(2024)
- 나사 제트추진 연구소 연구원(2005-2014)

발표 내용

본 연구에서는 최신 인공지능기반 기술을 활용하여 저작권을 침해하는 방법들과 사례에 대해서 논의하고, 최근 문제가 되고 있는 다향한 AI 보안과 보호이슈에 대해서 설명 및 논의합니다.

In this tutorial, we will discuss the new copyright infringement and protection methods using the latest AI-based methods. Also, we will explain and discuss the latest AI Security issues arising from generative AI methods.



Tutorial의 개요

- 본 연구에서는 최신 인공지능기반 기술을 활용하여 저작권을 침해하는 방법들과 사례에 대해서 논의
- 최근 문제가 되고 있는 다양한 AI 보안과 보호이슈에 대해서 설명 및 논의

Content

- 인공지능 기반 저작권 침해 기술
- 인공지능 기반 저작권 보호 기술
- 다른 중요한 사항들

Copyright (Watermark) Attack Methods

Attacks on Text-to-Image Gen Models

- NightShade: **Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models (S&P24)**

Attacks on DNN Model

- **DeepEclipse: How to Break White-Box DNN-Watermarking Schemes (UsenixSec24)**

Nightshade

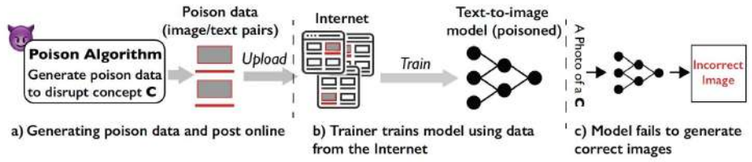


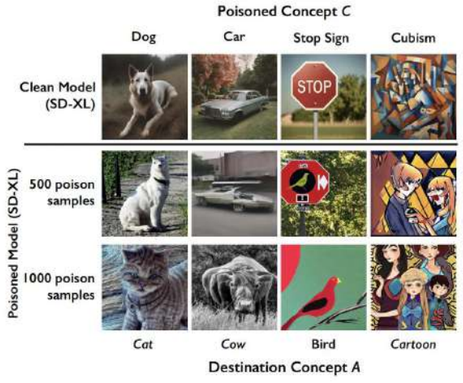
Figure 1. Overview of prompt-specific poison attacks against generic text-to-image generative models. (a) User generates poison data (text and image pairs) designed to corrupt a given concept C (i.e. a keyword like "dog"), then posts them online; (b) Model trainer scrapes data from online webpages to train its generative model; c) Given prompts that contain C, poisoned model generates incorrect images.

Nightshade

Figure 3. Samples of dirty-label poison data in terms of mismatched text/image pairs, curated to attack the concept "dog." Here "cat" was chosen by the attacker as the destination concept $\mathcal{A}$.

Attack Notation. The key to the attack is the curation of the mismatched text/image pairs. To attack a regular concept $\mathcal{C}$ (*e.g.*, "dog"), the attacker performs the following:

- select a "destination" concept $\mathcal{A}$ unrelated to $\mathcal{C}$ as guide;
- build a collection of text prompts $\mathbf{Text}_{\mathcal{C}}$ containing the word $\mathcal{C}$ while ensuring none of them include $\mathcal{A}$;
- build a collection of images $\mathbf{Image}_{\mathcal{A}}$, where each image visually captures the essence of $\mathcal{A}$ but contains no visual elements of $\mathcal{C}$;
- pair a text prompt from $\mathbf{Text}_{\mathcal{C}}$ with an image from $\mathbf{Image}_{\mathcal{A}}$.



Nightshade

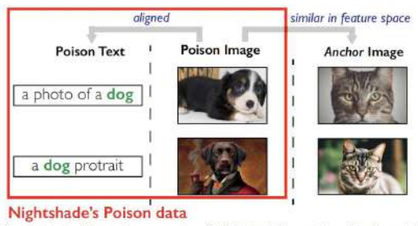


Figure 5. An illustrative example of Nightshade's curation of poison data to attack the concept "dog" using "cat". The anchor images (right) are generated by prompting "a photo of cat" on the clean SD-XL model multiple times. The poison images (middle) are perturbed versions of natural images of "dog", which resemble the anchor images in feature representation.

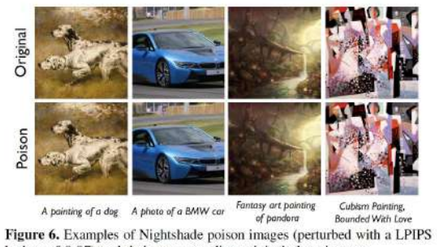


Figure 6. Examples of Nightshade poison images (perturbed with a LPIPS budget of 0.07) and their corresponding original clean images.

Training Scenario	Model Name	Pretrain Dataset (# of pretrain data)	# of Clean Training Data
Train from scratch	LD-CC	-	1 M
Continuous training	SD-V2	LAION (~600M)	100K
	SD-XL	Internal Data (>600M)	100K
	DF	LAION (~600M)	100K

TABLE 2. Text-to-image models and training configurations.

Nightshade

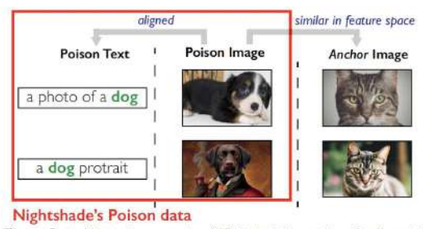


Figure 5. An illustrative example of Nightshade's curation of poison data to attack the concept "dog" using "cat". The anchor images (right) are generated by prompting "a photo of cat" on the clean SD-XL model multiple times. The poison images (middle) are perturbed versions of natural images of "dog", which resemble the anchor images in feature representation.

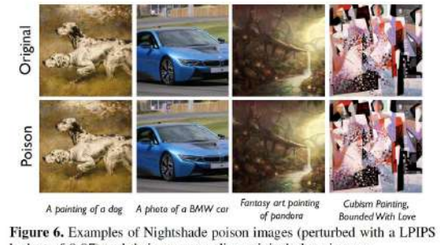


Figure 6. Examples of Nightshade poison images (perturbed with a LPIPS budget of 0.07) and their corresponding original clean images.

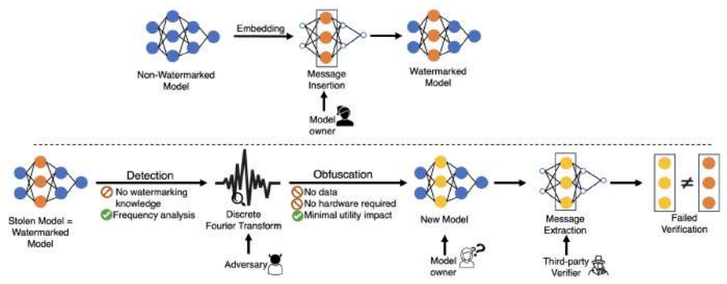
Training Scenario	Model Name	Pretrain Dataset (# of pretrain data)	# of Clean Training Data
Train from scratch	LD-CC	-	1 M
Continuous training	SD-V2	LAION (~600M)	100K
	SD-XL	Internal Data (>600M)	100K
	DF	LAION (~600M)	100K

TABLE 2. Text-to-image models and training configurations.

Nightshade

Figure 15. Image generated from different prompts by a poisoned SD-XL model where concept "fantasy art" is poisoned. Without being targeted, related prompts are also corrupted by the poisoning (*i.e.*, bleed through effect), while unrelated prompts face limited impact. The SD-XL model is poisoned with 200 poison samples.

DeepEclipse



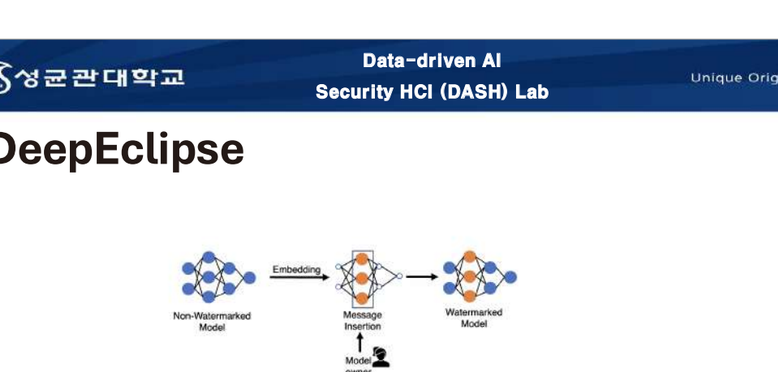


Figure 4: Basic *Convolutional* Layers obfuscation. Each feature maps of the Kernel is expanded with zeros padding.

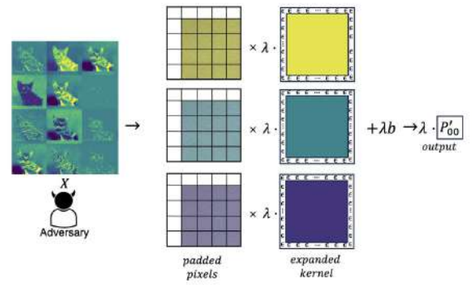
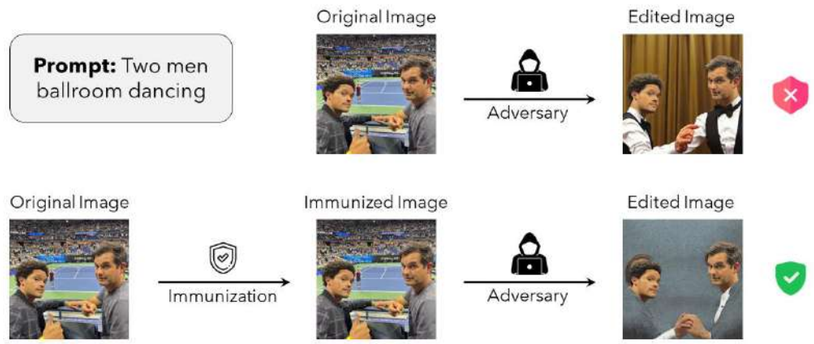


Figure 5: Advanced *Convolutional* Layers obfuscation. Each feature maps of the Kernel is expanded with padding using an ε value, then the whole layer is multiplied by a random constant λ, and the subsequent layer is also multiplied by $\frac{1}{\lambda}$.

Copyright Protection Methods

- Disrupting input when modified: PhotoGuard



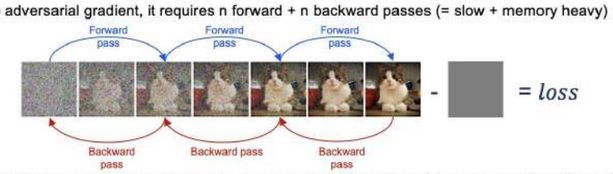
https://github.com/MadryLab/photoguard?tab=readme-ov-file

Copyright Protection Methods

- For someone wants to use Geneative AI Model to modify the original content,
  - → Disrupt the output, when user wants to change
  - → Do not generate user-requested desired outuput

Some issues with PhotoGuard

- There are some issues with Photoguard:
  - To yield the adversarial gradient, it requires n forward + n backward passes (= slow + memory heavy)



  - To make the attack invariant to random seeds, they take the expectation of the adversarial gradients. The synthesis of a single immunized image can take up to 20 minutes.
  - Yet, we confirmed unsatisfactory success rate w.r.t. random seeds (and doesn't work on specific images)

Our work addresses these limitations:
1) 3~4 times faster
2) higher inpainting disruption at all metrics

Disrupting Diffusion-based Inpainters with Semantic Digression

Geonho Son*, **Juhun Lee***, and Simon S. Woo

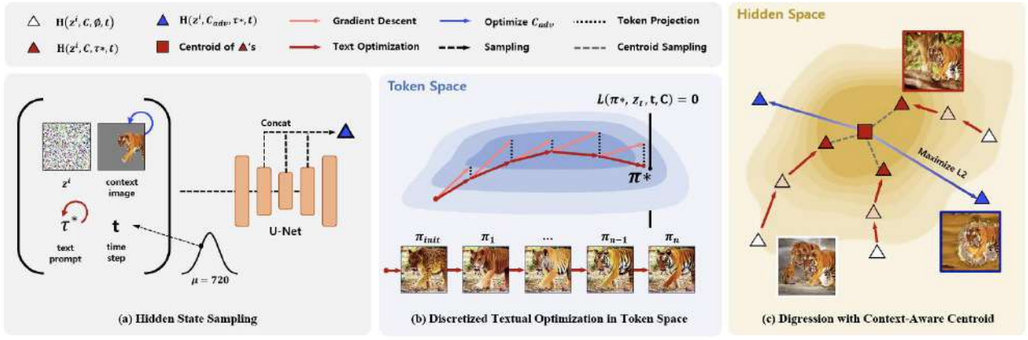
IJCAI 2024

Department of Artificial Intelligence,

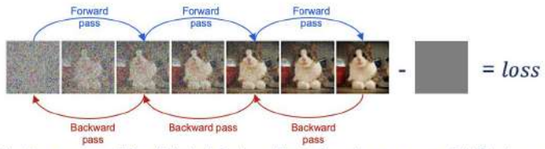
Sungkyunkwan University

- Privacy concerns should cross your mind. Deepfake synthesis are easier than ever.
- Users will want to protect their images from unconsented manipulation.

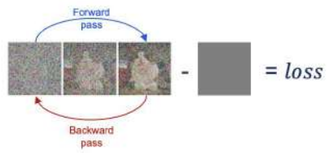
Deepfake Generation w/ SD Inpainters



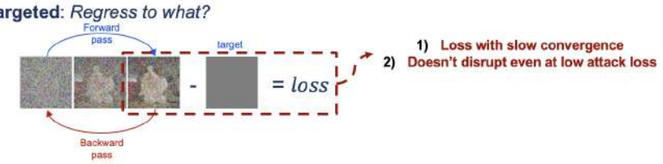
- We ask ourselves if covering the full diffusion process function is needed.



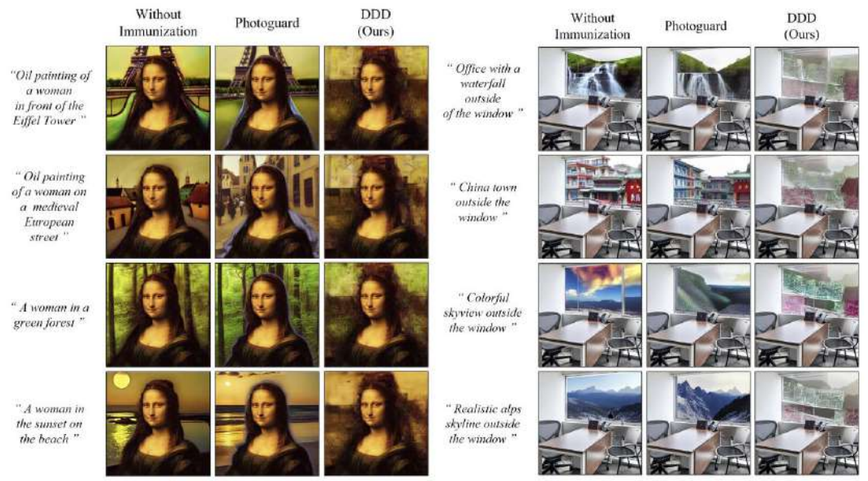
- Alternatively, aware of the "global-to-local" synthesis process of diffusion models, we find it appropriate to attack the early timestep range.

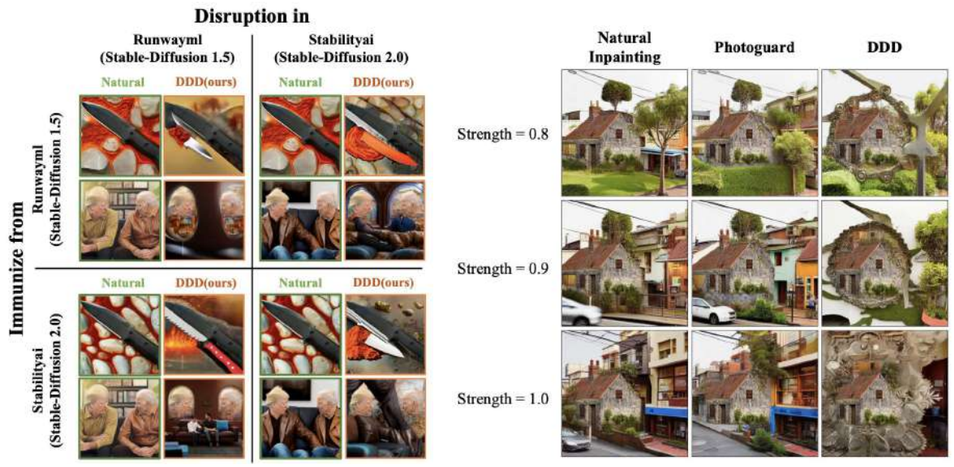


- In the generative space, both the **targeted** and **untargeted** attack setting are cumbersome:
  - **Targeted**: *Regress to what?*



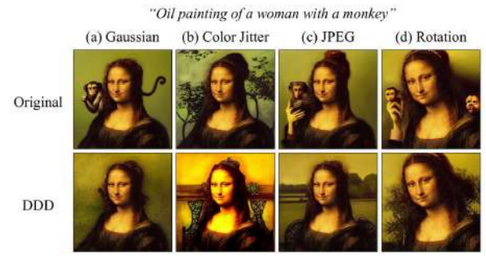
  - **Untargeted**: Sounds fitting to our task but, *digress away from what?*
    - If we can find the optimal representation of the normally generated images apriori, any digression w.r.t. it will be orthogonal to normal generations (a.k.a. abnormal)





Summary

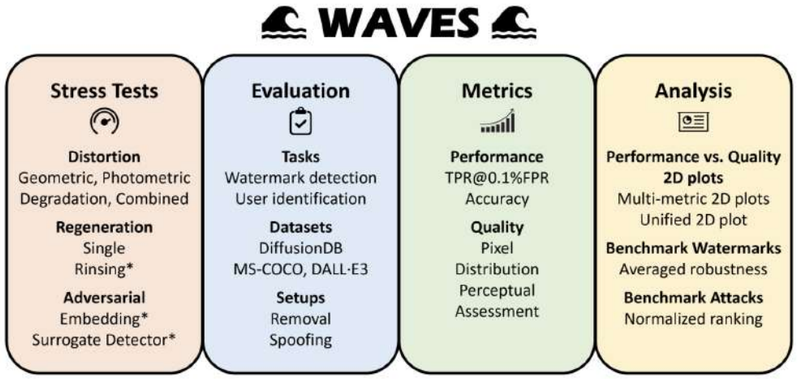
- In terms of disruption rate & strength, more than our method is insignificant
  - The disruption level/rate is far over the ideal bar.
  - If you ever get to research on this topic: optimizing for a stronger disruption is not recommended.
- There are critical issues with this line of research (Photoguard and Ours)
  - The assumption that we know where the malicious user will manipulate is too strong.
  - Robustness w.r.t. image augmentation is still weak.



Watermark Evaluation

- Watermark Evaluation/Benchmarking
- Watermark Robustness

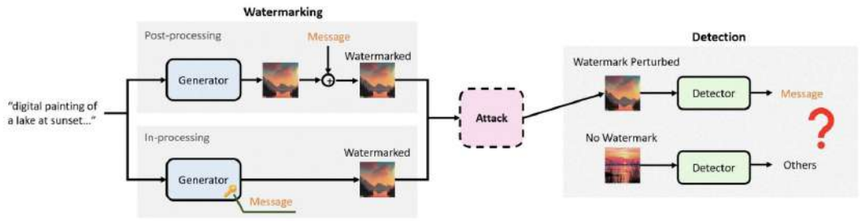
WAVES



An, Bang, et al. "Benchmarking the robustness of image watermarks." *arXiv preprint arXiv:2401.08573* (2024).

WAVES

- Watermark Robustness

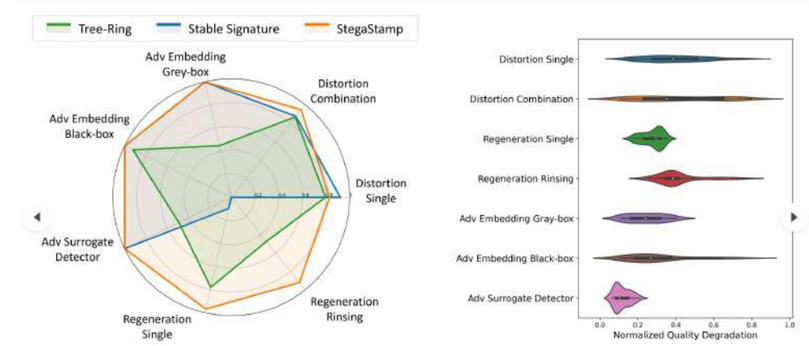


An, Bang, et al. "Benchmarking the robustness of image watermarks." *arXiv preprint arXiv:2401.08573* (2024).

Brief Introduction to Copyright Protection

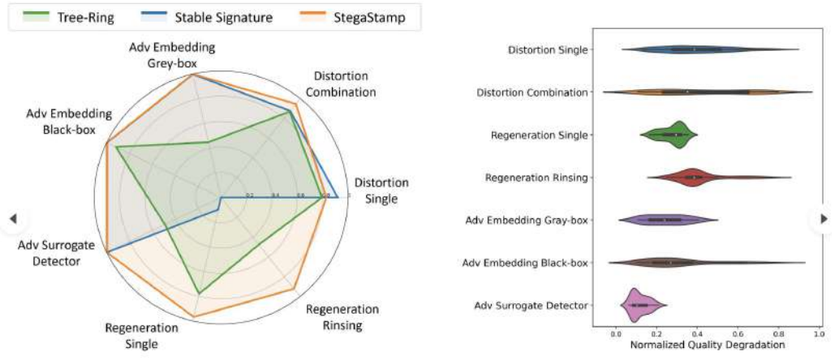
- Insert Watermark

- Benchmark



Watermark detection performance (i.e., TPR@0.1%FPR) of Stable Signature, StegaStamp, and Tree-Ring watermarks after attacks via WAVES. We compute the Average TPR@0.1%FPR across all strength levels and further averaged this metric across different attacks and datasets. Lower Average TPR@0.1%FPR indicates higher vulnerability of the watermark to a certain type of attack. Right figure shows the distribution of **quality degradation** for each type of attack. Lower quality degradation is preferred.

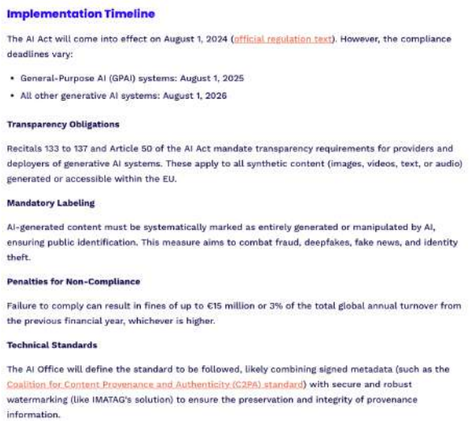
- Benchmark



Watermark detection performance (i.e., TPR@0.1%FPR) of Stable Signature, StegaStamp, and Tree-Ring watermarks after attacks via WAVES. We compute the Average TPR@0.1%FPR across all strength levels and further averaged this metric across different attacks and datasets. Lower Average TPR@0.1%FPR indicates higher vulnerability of the watermark to a certain type of attack. Right figure shows the distribution of **quality degradation** for each type of attack. Lower quality degradation is preferred.

Law. Watermark + GenAI

- The EU's AI Act



Implementation Timeline

The AI Act will come into effect on August 1, 2024 (official regulation text). However, the compliance deadlines vary:

- General-Purpose AI (GPAI) systems: August 1, 2025
- All other generative AI systems: August 1, 2026

Transparency Obligations

Recitals 133 to 137 and Article 50 of the AI Act mandate transparency requirements for providers and deployers of generative AI systems. These apply to all synthetic content (images, videos, text, or audio) generated or accessible within the EU.

Mandatory Labeling

AI-generated content must be systematically marked as entirely generated or manipulated by AI, ensuring public identification. This measure aims to combat fraud, deepfakes, fake news, and identity theft.

Penalties for Non-Compliance

Failure to comply can result in fines of up to €15 million or 3% of the total global annual turnover from the previous financial year, whichever is higher.

Technical Standards

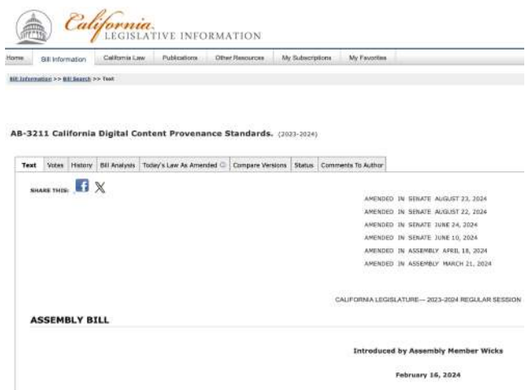
The AI Office will define the standard to be followed, likely combining signed metadata (such as the Coalition for Content Provenance and Authenticity (C2PA) standard) with secure and robust watermarking (like IMATAG's solution) to ensure the preservation and integrity of provenance information.

https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf

Concluding Remarks

- There are critical challenges in watermark methods for Generative AI systems
- Generally, defenses are much harder and more research efforts are required
- Joint efforts from Goverenments, Platforms, and Users are all needed to solve this challengening issues

- **California AI Watermarking Bill**



California LEGISLATIVE INFORMATION

Home | Bill Information | California Law | Publications | Other Resources | My Subscriptions | My Favorites

Bill Information >> Bill Search >> Text

AB-3211 California Digital Content Provenance Standards. (2023-2024)

Text | Votes | History | Bill Analysis | Today's Law As Amended | Compare Versions | Status | Comments To Author

SHARE THIS:

AMENDED IN SENATE AUGUST 23, 2024
AMENDED IN SENATE AUGUST 22, 2024
AMENDED IN SENATE JUNE 24, 2024
AMENDED IN SENATE JUNE 10, 2024
AMENDED IN ASSEMBLY APRIL 18, 2024
AMENDED IN ASSEMBLY MARCH 21, 2024

CALIFORNIA LEGISLATURE— 2023–2024 REGULAR SESSION

ASSEMBLY BILL

Introduced by Assembly Member Wicks

February 16, 2024

Reuters

World | US Election | Business | Markets | Sustainability | Legal | Breakingviews | Technology

Artificial Intelligence | Data Privacy | Intellectual Property | Public Policy

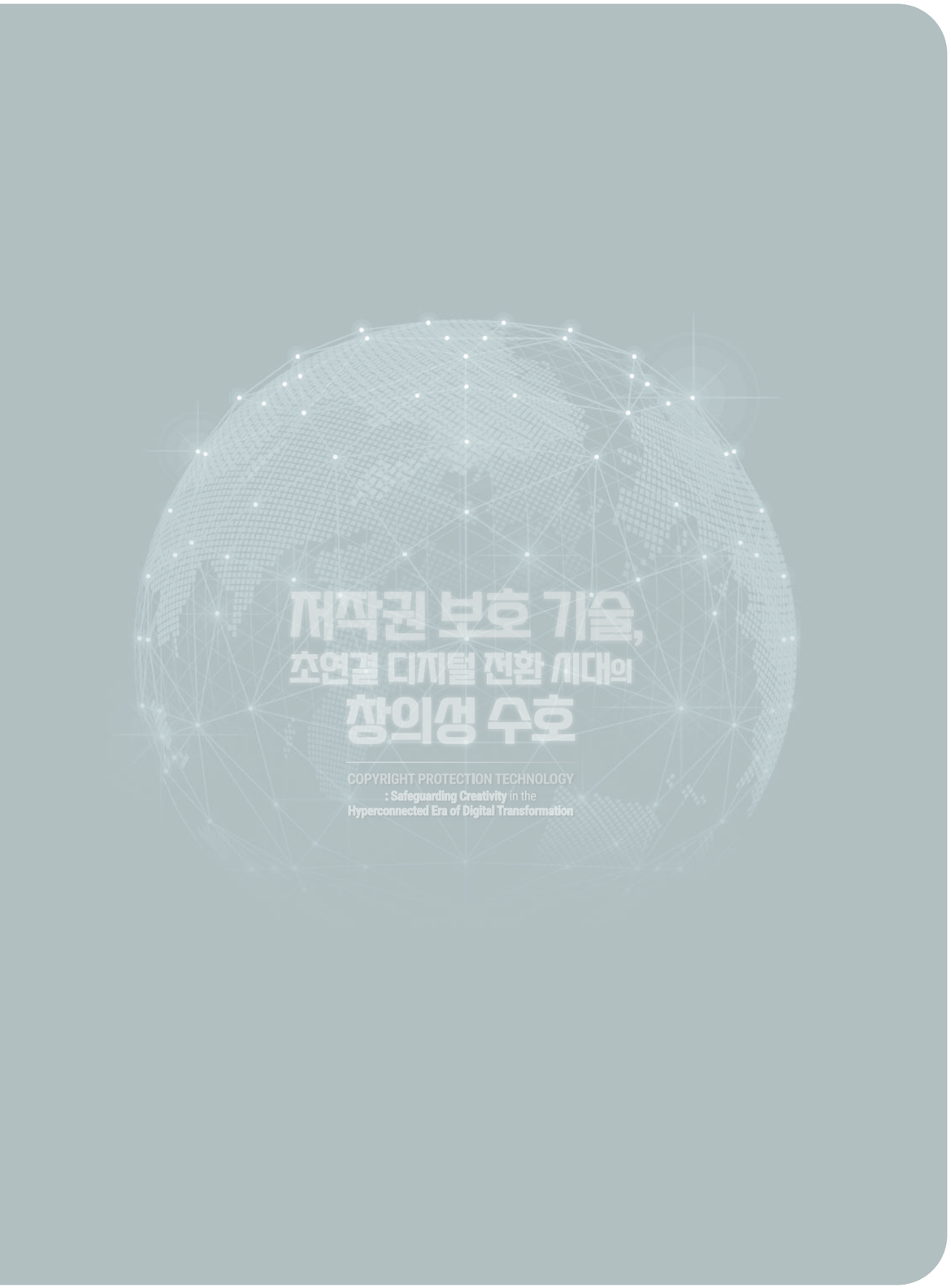
OpenAI supports California AI bill requiring 'watermarking' of synthetic content

By Anna Tong

August 27, 2024 4:30 AM GMT+9 · Updated 2 months ago

OpenAI logo is seen in this illustration taken May 20, 2024. REUTERS/Dado Ruvic/Illustration/File Photo Purchase Licensing Rights

https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB3211



Session 2

콘텐츠 창작의 토대, 저작권 보호 기술

Ⅰ 웹툰 불법유출에 대한 기술적 대응의 중요성

서충현 | 네이버웹툰 실장

Ⅱ 콘텐츠 보호: 트렌드와 과제

에릭 딜 | 소니 픽처스 엔터테인먼트 보안 및 미디어 기술 부사장

Ⅲ OTT 콘텐츠 불법 유출 현황과 이에 대응하는 콘텐츠 보안 기술 소개

김준호 | 잉카엔트웍스 프로덕트 매니저

Ⅳ 콘텐츠 분석 및 워터마킹을 통한 이미지 복제 감지

마테이스 두즈 | 메타 연구원

Session 2 콘텐츠 창작의 토대, 저작권 보호 기술

I 웹툰 불법유출에 대한 기술적 대응의 중요성

서충현
네이버웹툰 실장

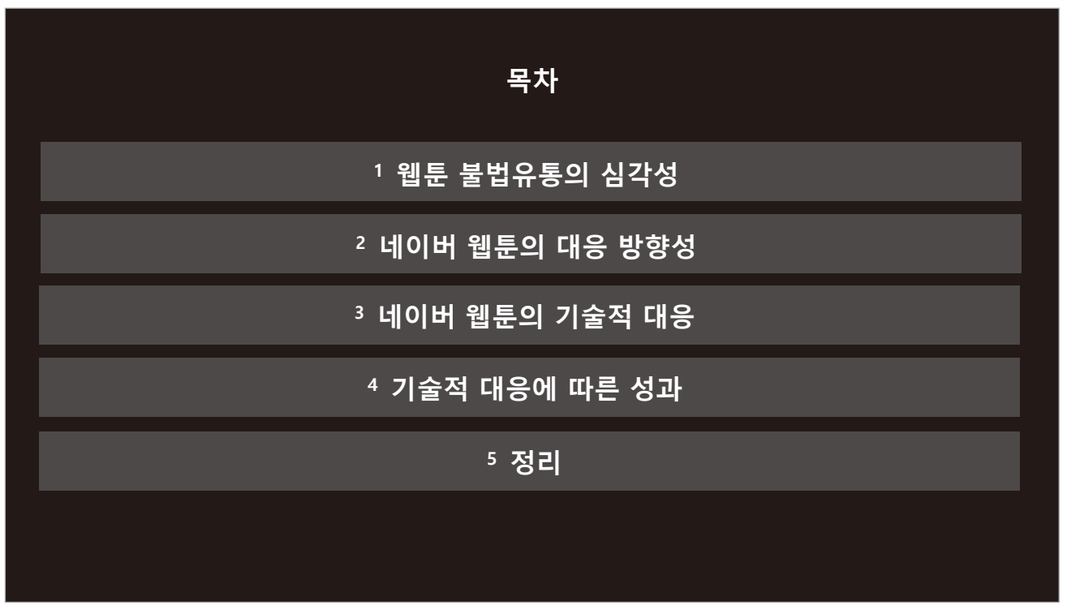
연사 이력

- 네이버웹툰(유) Anti-Piracy 리드(2024~)
- 네이버웹툰(유) W AI 리드(2023~)
- 네이버웹툰(유) AI Protection 리드(2017~2023)
- 네이버(주) 소프트웨어 엔지니어(2011~2017)

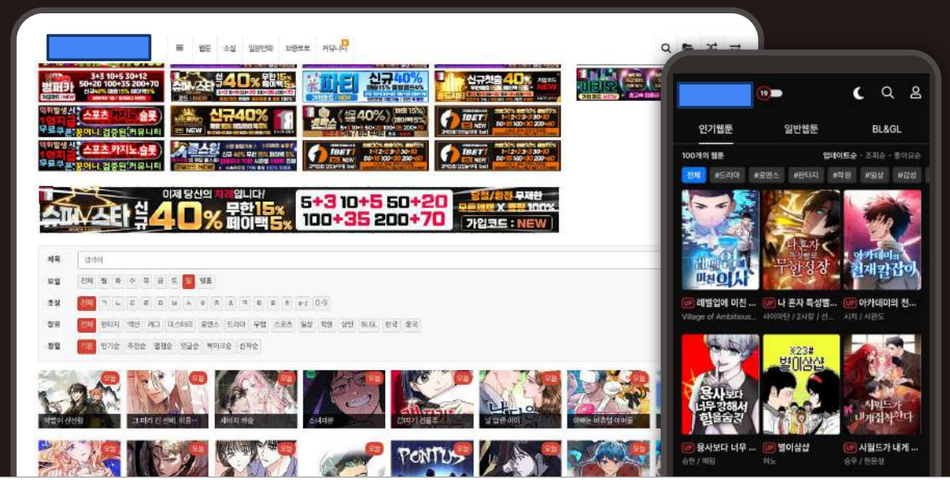
발표 내용

디지털 콘텐츠 시장의 성장과 함께 웹툰 불법 유통 방지는 지식 재산권 보호에 있어 중대한 과제가 되었습니다. 네이버웹툰은 이러한 문제에 대응하기 위해 선제적으로 다양한 기술적 전략을 개발하고 실행해 왔습니다. 본 발표에서는 불법 웹툰 유통 방지를 위한 기술적 대응의 중요성을 강조하고, 플랫폼 차원에서 이를 효과적으로 적용하는 방법과 그 성과에 대해 소개할 예정입니다. 이를 통해 불법 유통 억제뿐만 아니라 콘텐츠 생태계를 보호하는 방안을 공유하고자 합니다.

With the growth of the digital content market, preventing illegal distribution of webcomics has become a significant challenge in protecting intellectual property rights. In response to this issue, WEBTOON Entertainment has proactively developed and implemented various technical strategies. This presentation will highlight the importance of technical measures to prevent the illegal distribution of webcomics and introduce how these measures have been effectively applied at the platform level, along with the results achieved. By sharing these strategies, we aim to contribute to efforts not only to curb illegal distribution but also to safeguard the broader content ecosystem.



1 웹툰 불법유통의 심각성 | 한국



1 웹툰 불법유통의 심각성 | 한국

주요 지표 (월간 순방문자수)

국내 전체 불법 사이트
1,200 만명
수십 개 사이트 운영 중

〈'N' 불법사이트〉
약 470 만명

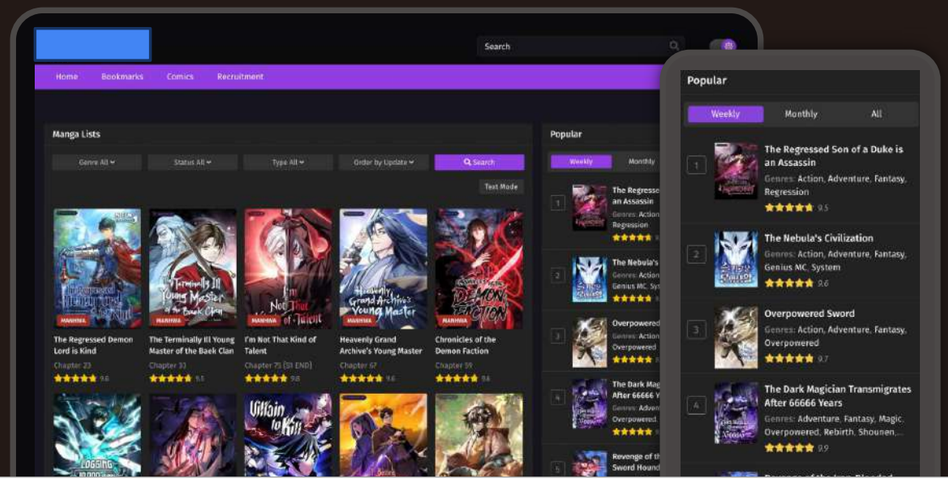
취급 작품

모든 웹툰 플랫폼의 모든 작품

특징

- 〈'N' 불법사이트〉가 불법 공유의 시작
- 오로지 돈을 벌기 위한 범죄행위
  (주로 도박 / 성인사이트 광고 수입)
- 운영자 신원 파악이 매우 어려움
- 정부의 도메인 차단에 잘 대비함

1 웹툰 불법유통의 심각성 | 해외



1 웹툰 불법유통의 심각성 | 해외

주요 지표 (월간 순방문자수)

1차 불법 사이트
4,000 만명
90 개 도메인 기준

〈'A' 불법사이트〉
약 1,100만명
점유율 **28%**

취급 작품

- **주요 웹툰 인기 작품**
- **〈'A' 불법사이트〉: 판타지+액션+무협 등 남성향 작품**

특징

- 〈'A' 불법사이트〉 등 언어별로 소수의 사이트가 최초 번역/공유
- 팬심(팬번역)으로 시작(미출시 이유로), 공식 플랫폼인'척' 운영
- 원본유출-번역-식자-이미지편집-검증 등 분업화
- 커뮤니티(디스코드) 운영, 운영자 신분이 어느정도 노출되어 있음

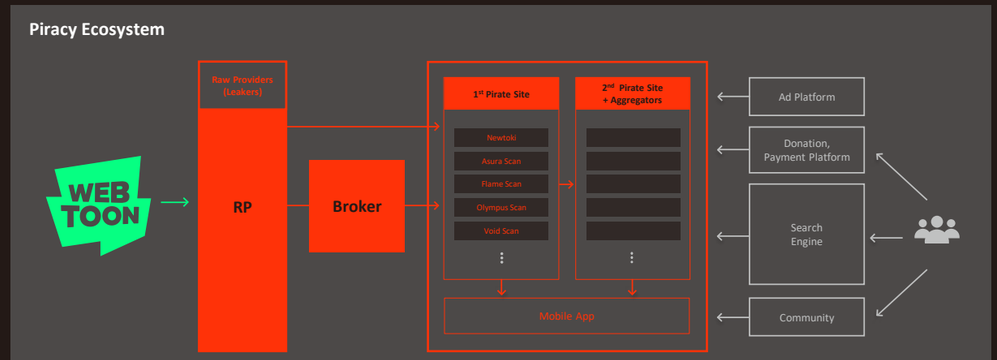
1 웹툰 불법유통의 심각성

2022년 국내 피해규모 추정액

7,215억원

출처 : 2023 웹툰 사업체 실태조사

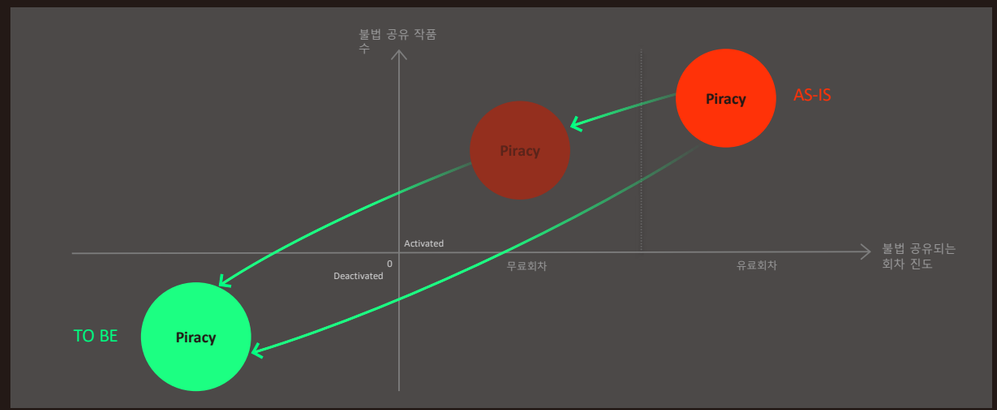
2 네이버 웹툰의 대응 방향성



2 네이버 웹툰의 대응 방향성

- 법적인 조치나 저작권 침해 게시물 신고/삭제 등의 사후 조치만으로는 한계가 뚜렷
- 웹툰 유출 자체를 막는 기술적 대응에 집중
- 기술적 대응의 결과로 **불법 공유되는 시점을 최대한 늦추고 불법 공유되는 작품수를 줄이는 것**
- 위와 같은 목표를 달성하는 것이 창작자와 플랫폼의 이익을 지키는 근본적인 방법

2 네이버 웹툰의 대응 방향성



3 네이버 웹툰의 기술적 대응

- **최근 불법유출 방식**
  - 서비스 취약점을 노린 기술 활용, 자동화/고속화
- **기본적인 보안 조치**
  - 유료 에피소드 열람을 위한 전용 어플리케인션을 사용자에게 제공
  - 이미지 암호화, DRM 적용 등을 통해 이미지 획득 과정을 최대한 복잡하게 구성
  - 어플리케이션 디버깅, 위변조, 해킹에 대비할 수 있는 솔루션 적용(ex. Anti Frida)

WEBTOON AI

TOONRADAR

3 네이버 웹툰의 기술적 대응 | TOONRADAR

사후 차단(ToonMark)	사전 차단(SniffingDog)
웹툰 이미지에 보이지 않는 표식을 삽입하여 최초 유출자를 추적/차단하는 기술 (워터마킹 기술)	유출자 데이터 기반으로 불법유출을 사전에 예측 차단하는 기술

3 네이버 웹툰의 기술적 대응 | TOONRADAR

- **사후 차단 (ToonMark = 워터마킹)**
  - 최초 유출자를 추적, 식별하여 이용차단. 모든 기술적 조치의 시작
  - 2017년 부터 자체적으로 개발한 워터마킹을 도입
  - 2024년 상반기에 Full AI기반의 워터마킹으로 업그레이드
  - 강인성, 비가시성 등을 고려하여 AI가 워터마크 위치를 스스로 판단하여 삽입/추출

3 네이버 웹툰의 기술적 대응 | TOONRADAR

• 사후 차단 (워터마킹) – 다양한 공격에 강인함

원본

공격 예시

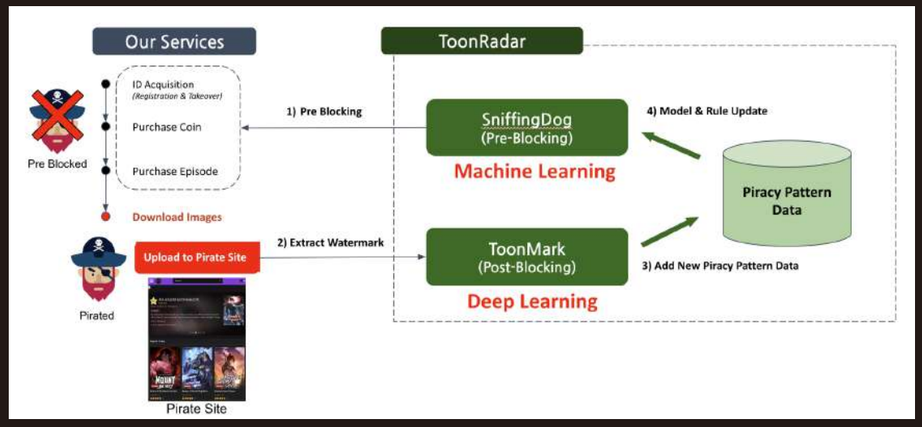
3 네이버 웹툰의 기술적 대응 | TOONRADAR

• 사전 차단 (SniffingDog)

- 워터마킹을 통한 유출자 적발도 여전히 사후 조치임
- 웹툰 유출에 동원되는 계정 규모를 고려하면 사전에 차단하는 것이 필수
- 워터마킹으로 적발한 유출자 데이터를 기반으로 사전에 예측 차단하는 시스템 개발/운영 중
- 룰 및 머신러닝 기반의 하이브리드 예측 모델을 사용

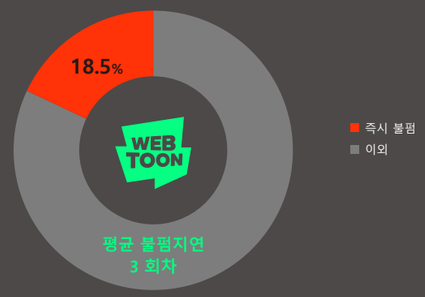
3 네이버 웹툰의 기술적 대응 | TOONRADAR

• ToonRadar System Overview



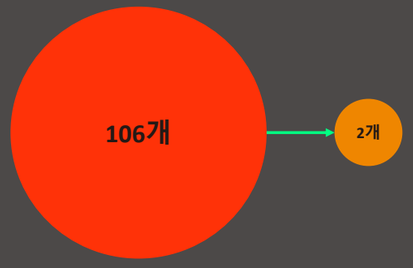
4 기술적 대응에 따른 성과

네이버 웹툰과 불법 사이트 최신 에피소드가 같은 작품 비율 (2024년 10월 기준)



4 기술적 대응에 따른 성과

네이버웹툰을 최초로 유출/공유하는 한국 1차 불법사이트 개수



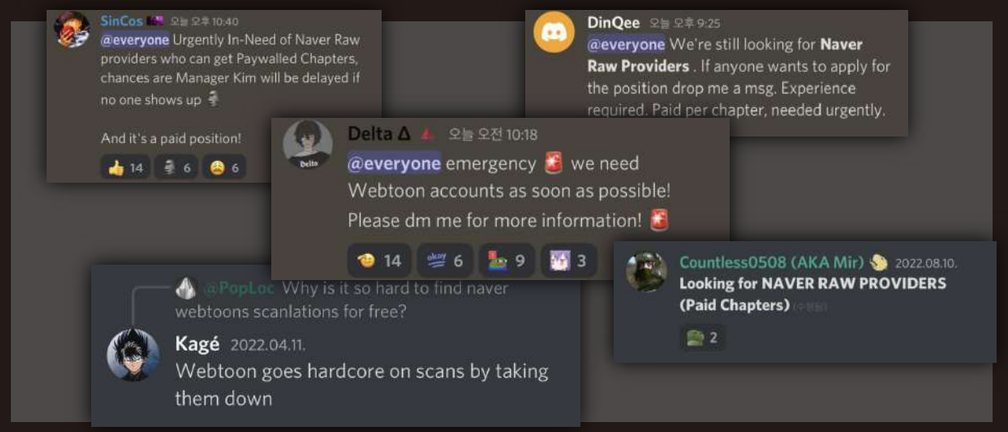
- 2017년~2024년 총 106개 1차 불법사이트 존재
- 현재 단 2개 사이트만 1차 불법사이트로 활동
- 나머지는 사이트가 내려갔거나 2차 불법사이트로 전락

4 기술적 대응에 따른 성과

툰레이더가 보호한 저작물 권리의
경제적 가치 환산 효과

2-3,000 억원

4 기술적 대응에 따른 성과



5 정리

- 플랫폼의 기술적 대응이 중요한 해결책이 될 수 있음
- 기술적 대응과 더불어 저작권 침해 게시물 삭제 및 저작권 인식 전환 캠페인도 중요
- 정부 및 수사기관의 적극적인 대응으로 불법사이트 폐쇄 및 운영자 검거 필요

감사합니다

Session 2 콘텐츠 창작의 토대, 저작권 보호 기술

II 콘텐츠 보호: 트렌드와 과제

에릭 딜
소니 픽처스 엔터테인먼트 보안 및 미디어 기술 부사장

연사 이력

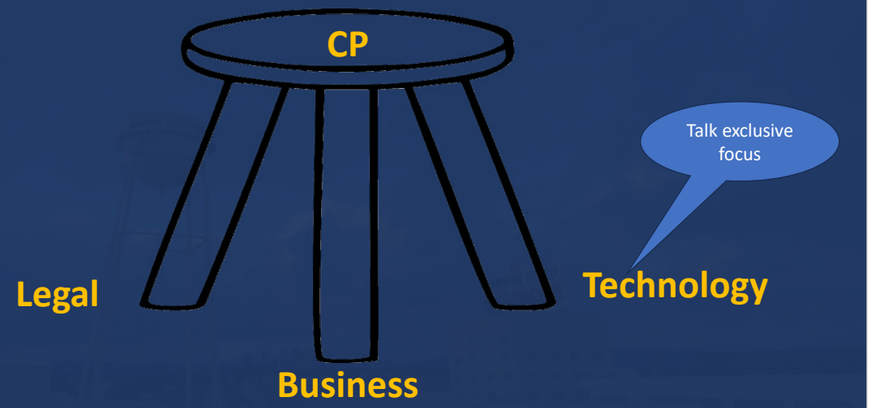
- 소니 법인 수석 엔지니어 (2022~2024)
- 소니 픽처스 엔터테인먼트 보안 및 미디어 부사장 (2014~2024)
- 소니 법인 수석 엔지니어 (2022~2024)
- 테크니컬러 보안 시스템 및 기술 부사장 (2012~2014)
- 테크니컬러 비즈니스 보안 서비스 부사장 (2011~2014)
- 테크니컬러 보안 및 콘텐츠 보호 연구소 부사장 (2010~2012)

발표 내용

이 기술 프레젠테이션은 콘텐츠 보호에서 최신 발전을 탐구합니다. 여기에는 화이트 박스 암호화 같은 콘텐츠 보호에 사용되는 핵심 기술들이 열거되어 있으며, 이들의 최근 진화를 요약합니다. 또한 블록체인, 양자 이후 컴퓨팅(PQC), AI 모델 워터마킹과 같은 새로운 도전 과제들도 탐구합니다.

This technical presentation explores the latest developments in content protection. It enumerates some critical technologies used in content protection, such as white-box cryptography, and summarizes their recent evolution. The presentation also explores new challenges like blockchain, post-quantum computing (PQC), and watermarking AI models.

Scope of the presentation



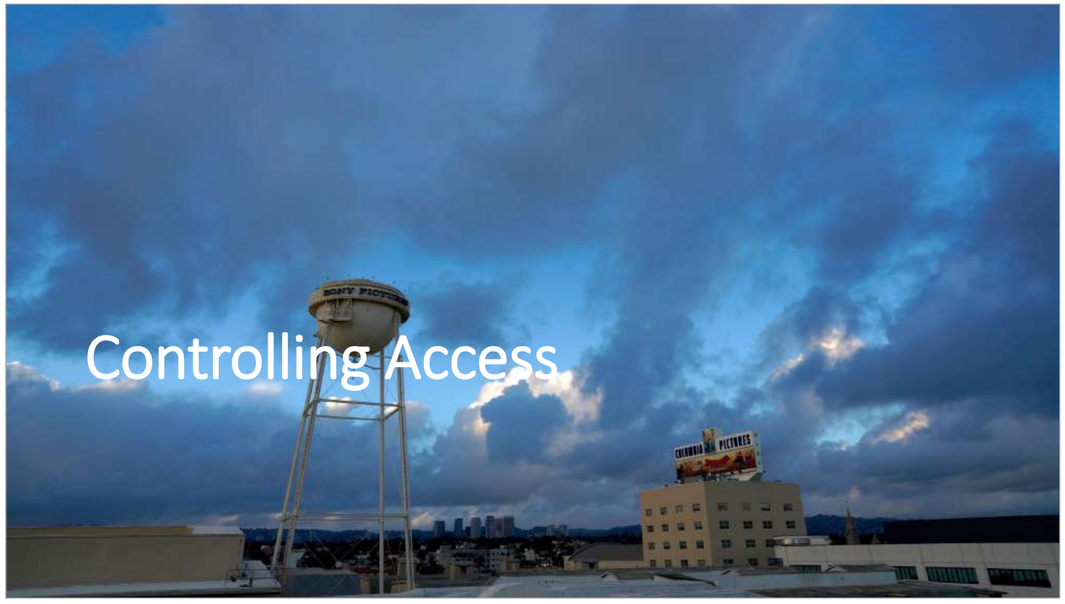
Agenda

Controlling access
Tracing back
Challenges

프레젠테이션 범위

논제

접근 제어
추적
도전 과제



White Box Cryptography

"Secure" software implementation of encryption

Attacker's goal
- Extract the secret key used by the implementation

Threat model
- Attacker fully controls the platform
- Attacker has full access to the binary code

How does it work?
- Lookup tables rather than calculated
- Randomizing the tables' values
- Code obfuscation
- "Black Art"

화이트 박스 암호화

암호화 "보안" 소프트웨어 구현

공격자 목표
- 구현에 사용된 비밀 키 추출

위협 모델
- 공격자가 플랫폼 완전 제어
- 공격자가 바이너리 코드 전체 액세스 권한 보유

어떻게 작동하는가?
- 계산된 값 대신 조회 테이블 사용
- 테이블 값 무작위로 지정
- 코드 난독화
- "블랙 아트"

White Box Cryptography

Regular competitions at CHES

Competition	Target	Result
WhiBox 2017	AES-128	Broken > 28 days
WhiBox 2019	AES-128	One not broken
WhiBox 2021	ECDSA	Broken > 2 days
WhiBox 2024	ECDSA	Broken > 6 days

No commercial solution has been publicly submitted

Hardware Security (1/2)

New kinds of vulnerabilities

- Micro-architecture level
  - Speculative execution, data cache…
- Spectre, Meltdown…

Example:

- SGX is not anymore considered secure

화이트 박스 암호화

CHES 정기 대회

대회	타겟	결과
WhiBox 2017	AES-128	Broken > 28 days
WhiBox 2019	AES-128	One not broken
WhiBox 2021	ECDSA	Broken > 2 days
WhiBox 2024	ECDSA	Broken > 6 days

상업적 솔루션이 공개적으로 제출되지 않음

하드웨어 보안(1/2)

새로운 종류의 취약성

- 아키텍처 수준
  - 추측 실행, 데이터 캐시…
- Spectre, Meltdown…

예시:

- SGX 더 이상 안전하지 않은 것으로 간주

Hardware Security (2/2)

Black Hat 2024

- *Laser Beams & Light Streams: Letting Hackers Go Pew Pew, Building Affordable Light-Based Hardware Security Tooling*. BEAUMONT S., TROWELL
- **500$**
- **Still needs skill and knowledge (a lot)**

Post Quantum?

The Quantum Computer risk

- What is a quantum computer?
- Risk
  - Current public key cryptosystem would be defeated
    - Schor algorithm
  - Symmetric: Reduce complexity O(n) to O(n/2)
    - Grover Algorithm

QC: new type of hard-to-solve problem (e.g., lattice)

- NIST is standardizing four algorithms
  - Key “sharing” (CRYSTAL-KYBER)
  - Digital signature (CRYSTALS-DILITHIUM, Falcon, and SPHINCS+)
- OpenSSH
  - NTRU Prime

하드웨어 보안(2/2)

Black Hat 2024

- *Laser Beams & Light Streams: Letting Hackers Go Pew Pew, Building Affordable Light-Based Hardware Security Tooling*. BEAUMONT S., TROWELL
- **500$**
- **여전히 기술과 지식 많이 필요**

포스트 퀀텀?

양자 컴퓨터 위험

- 양자 컴퓨터란 무엇인가?
- 위험
  - 현재 공개 키 암호 시스템은 무너진다
    - 쇼어 알고리즘
  - 대칭: 복잡도 O(n)을 O(n/2)로 줄임
    - 그로버 알고리즘

QC: 새로운 유형의 풀기 어려운 문제(예: 격자)

- NIST, 네 가지 알고리즘 표준화
  - 키 "공유" (CRYSTAL-KYBER)
  - 디지털 서명(CRYSTALS-DILITHIUM, Falcon, and SPHINCS+)
- 오픈SSH
  - NTRU Prime

Post Quantum & CP standards

Personal opinion

Should we push QC in current or new standards?

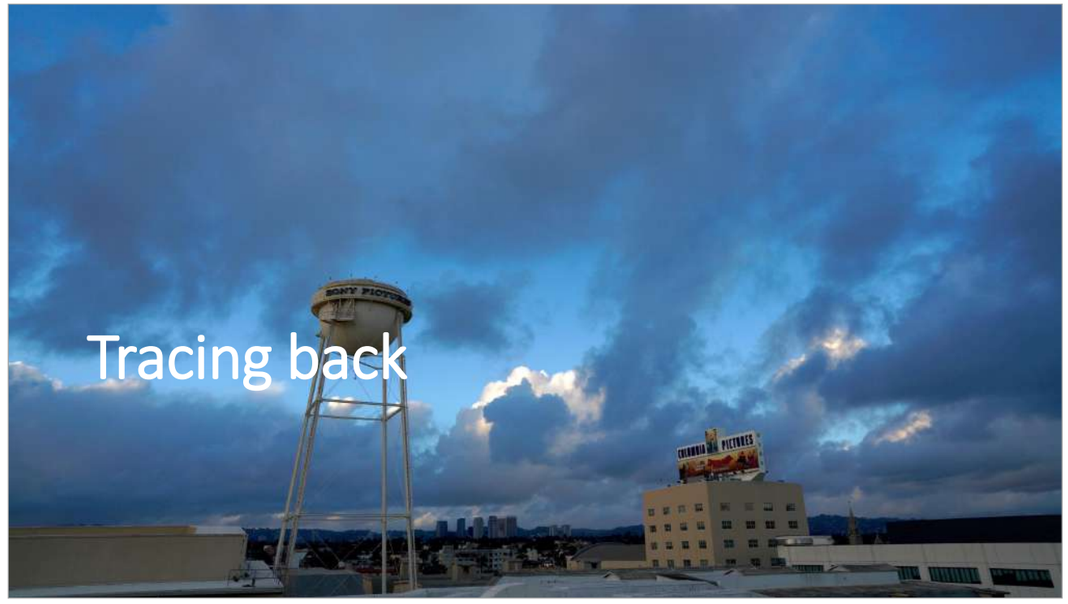
When will Armageddon occur?

- Schor requires millions of logical qubits
- Currently, a few tens of logical qubits
- Cryptographically Relevant Quantum Computer

CP no long-term keys, cost-related

Maturity

- No chip available
- No study on side channel attacks



포스트 퀀텀 및 CP 표준

개인적 의견

현재 또는 새로운 표준에서 QC를 적용해야 하나?

아마겟돈은 언제 일어날까?

- 쇼어는 수백만 개의 논리적 큐비트가 필요
- 현재 수십 개의 논리적 큐비트
- 암호학적으로 관련성 있는 양자

CP 컴퓨터CP에는 장기 키가 없으며 비용과 관련이 있음

완성도

- 사용 가능한 칩 없음
- 사이드 채널 공격에 대한 연구 없음

Machine Learning watermark

Interesting

Generative Adversarial Network (GAN) allows for enhanced robustness

Issue

- No advanced visual model
- Using PSNR as imperceptibility for reward
- How to feedback golden eyes

Future risk

- Using GAN to attack when Oracle attack is available

WM and Homomorphic Encryption

What is homomorphic encryption

$$D\left(f\left(E(m)\right)\right) = f(m)$$

WM is too complex for Full Homomorphic Encryption

Note: Some commercial WM work in encrypted domain

머신 러닝 워터마크

흥미로운 점

생성적 적대 신경망(GAN), 향상된 강인성 제공

이슈

- 고급 비주얼 모델 결여
- PSNR을 비가시성의 보상으로 활용
- 골든 아이즈 피드백을 어떻게 처리할 것인가

향후 리스크

- Oracle 공격이 가능할 경우 GAN으로 공격

WM 및 동형 암호화

동형 암호화란 무엇인가

$$D\left(f\left(E(m)\right)\right) = f(m)$$

WM는 전체 동형 암호화 하기에는 지나치게 복잡

참고: 일부 상업용 워터마크는 암호화된 영역에서 작동함

Challenges

Blockchain

Many startups claim they solved CP

Reduce the problem to:

- Easy lawful licensing of content
- Royalties' distribution

Not tackling unlawful use, such as piracy

Usually, not a proper approach to interface with DRM

도전 과제

블록체인

여러 스타트업들이 CP 문제를 해결했다고 주장

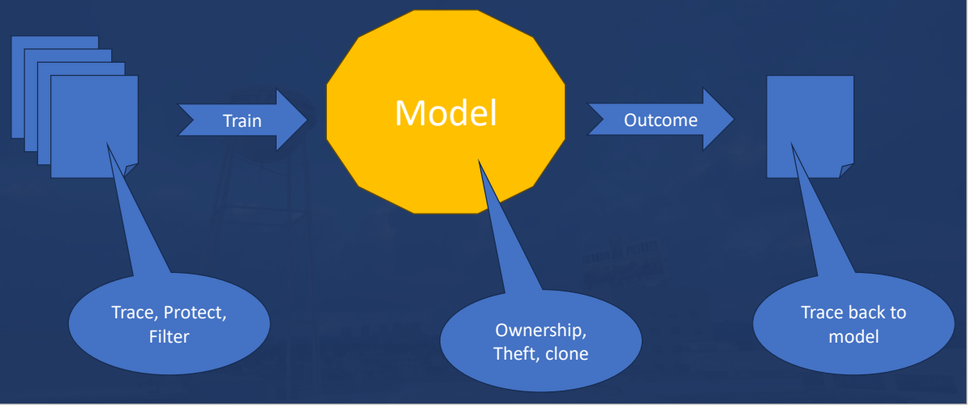
문제 간소화:

- 콘텐츠 합법적 라이선스 용이
- 로열티 배분

불법 사용, 즉 해적 행위를 다루지 않음

대개 DRM과 상호작용하는 적절한 접근 방식이 아님

Content Protection and ML/AI (1/2)



Content Protection and ML/AI (2/2)

Blockchain
- Licensing training data
- How to monitor the usage?

Watermark
- Promising
- Need much research before mature
- Critical for avoiding data pollution?

콘텐츠 보호 및 ML/AI(1/2)

콘텐츠 보호 및 ML/AI(2/2)

블록체인
- 라이선싱 교육 데이터
- 사용량 모니터링 방법은?

워터마크
- 기대되는 점
- 성숙하기 위해서는 많은 연구 필요
- 데이터 오염 방지에 있어 필수적인가?


감사합니다
Thank you
Merci
SONY
PICTURES

Session 2 콘텐츠 창작의 토대, 저작권 보호 기술

III OTT 콘텐츠 불법 유출 현황과 이에 대응하는 콘텐츠 보안 기술 소개

김준호
잉카엔트웍스 프로덕트 매니저

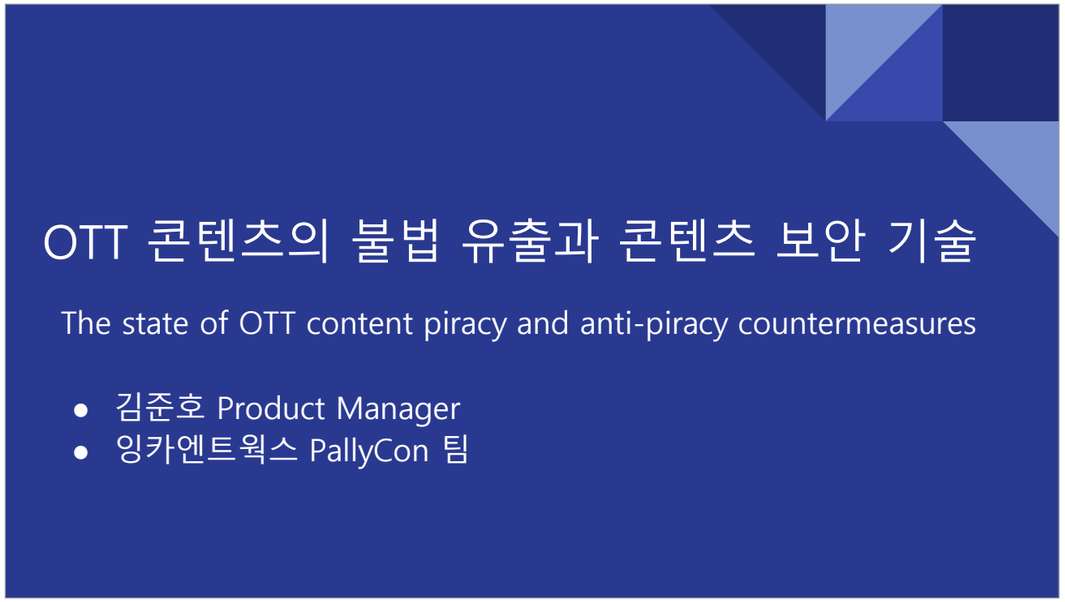
연사 이력

- 잉카엔트웍스 프로덕트 매니저 (2015~2024)
- 잉카엔트웍스 수석 연구원 (2010~2015)
- 유비트로텍 수석 개발자 (2007~2010)
- 에이씨티소프트 개발자 (2002~2007)
- 넷츠커뮤니케이션즈 개발자 (2000~2002)

발표 내용

이번 발표에서는 '제2의 누누티비'와 같은 콘텐츠 불법 유출의 유형과 원인, 그리고 대응 방법을 공유합니다. 클라이언트의 보안 수준에 따른 적절한 DRM(디지털 저작권 관리) 적용 방안을 살펴보고, 소프트웨어 DRM의 취약점과 그에 대한 보완 방안도 제시하겠습니다. 마지막으로, 머신러닝을 활용한 이상치 탐지 모델을 통해 콘텐츠 불법 유출에 효과적으로 대응할 수 있는 방법에 대해 설명할 예정입니다.

In this presentation, I will discuss the types and causes of illegal content leaks, such as "NunuTV 2.0," and the ways to respond. We will explore appropriate DRM (Digital Rights Management) strategies based on the security level of the client's system and examine the vulnerabilities of software-based DRM, along with complementary solutions. Finally, I will explain how we can effectively address illegal content leaks by using anomaly detection models powered by machine learning.



OTT 콘텐츠 불법 유출의 유형

불법 VOD 콘텐츠 유통 구조

제2의 누누티비, 사라지지 않는 불법 스트리밍

스트리밍 콘텐츠 불법 유출 – WEBRIP? WEBDL?

유출 방식	레이블	설명
Cam	CAM-Rip, CAM, HDCAM	영화관 스크린 또는 TV 화면을 촬영해 녹화한 유출본. 상대적으로 화질/음질이 낮음.
Screener	SCR, DVDSCR, BDSCR	시사회 또는 영화제 출품을 위해 배포된 DVD 또는 블루레이 디스크에서 추출된 유출본.
DVD/Blu-ray Rip	DVDRip, BDRip, BluRay	시장에 정식 출시된 DVD/블루레이 디스크에서 추출된 유출본.
Web Rip	WEBRIP, WEB	스트리밍 방식의 콘텐츠를 대상으로 HDMI 출력 영상을 캡쳐/녹화해 만든 유출본. (DRM 암호화 키 유출 X) 재인코딩으로 인해 WEB-DL 방식보다 화질이 낮음.
Web Download	WEBDL, WEB-DL	DRM 암호화된 스트리밍 콘텐츠를 복호화해 원본 영상과 동일하게 만들어낸 유출본. DRM 클라이언트 모듈(Content Decryption Module)의 소프트웨어 취약점을 이용해 암호화 키를 추출.

1080p.WEBRip.1400MB.DD5.1.x264-GalaxyRG

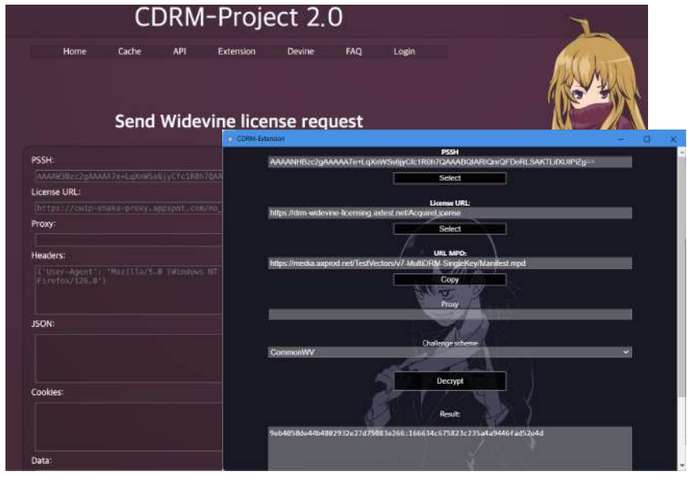
2160p.AMZN.WEB-DL.DDP5.1.HDR.H.265-FLUX[TGx]

WEBRip 1080p DTS DD+ 5.1 Atmos x264-MgB

1080p.WEB.H264-BOOTSINPUSS[TGx]

1080p.10bit.WEBRip.6CH.x265.HEVC-PSA

CDRM-Project와 DRM 콘텐츠 키 유출



정상적인 과정으로 유출 대상 DRM 콘텐츠 재생

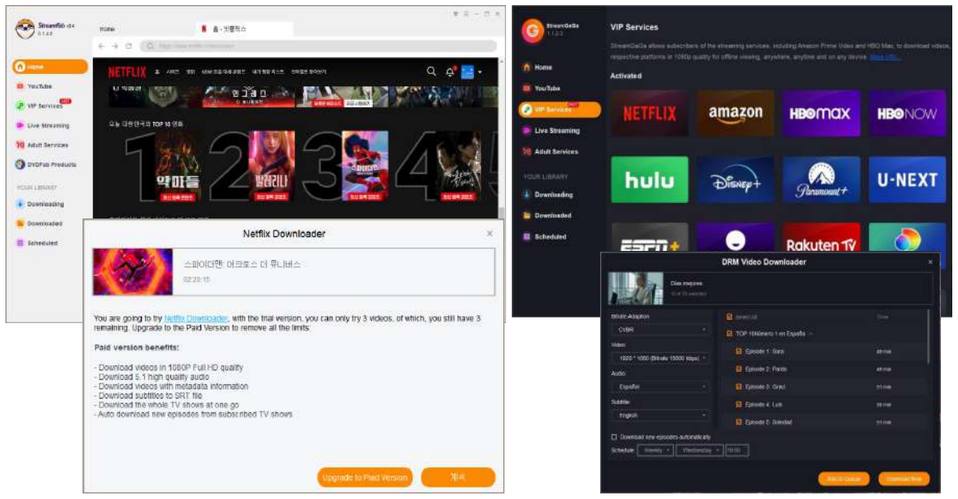
DRM 라이선스 요청 관련 정보 획득

해킹된 CDM을 이용해 DRM 라이선스 요청 및 획득

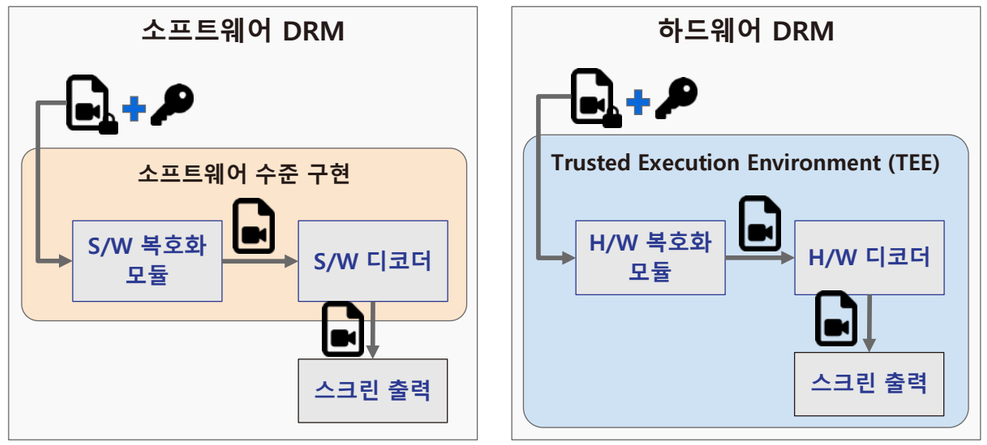
라이선스 내 암호화 키 추출 → DRM 콘텐츠 복호화

클라이언트 보안 수준에 따른 DRM 적용

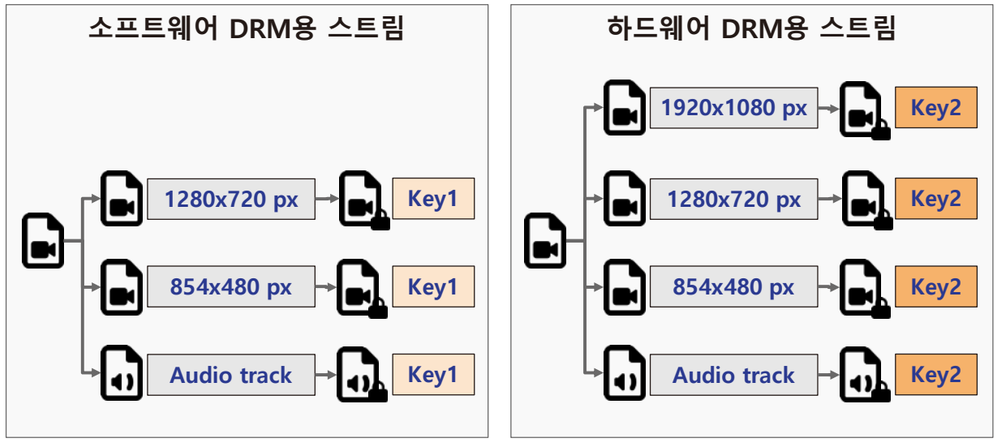
DRM 콘텐츠 다운로더 – StreamFab 및 유사 툴



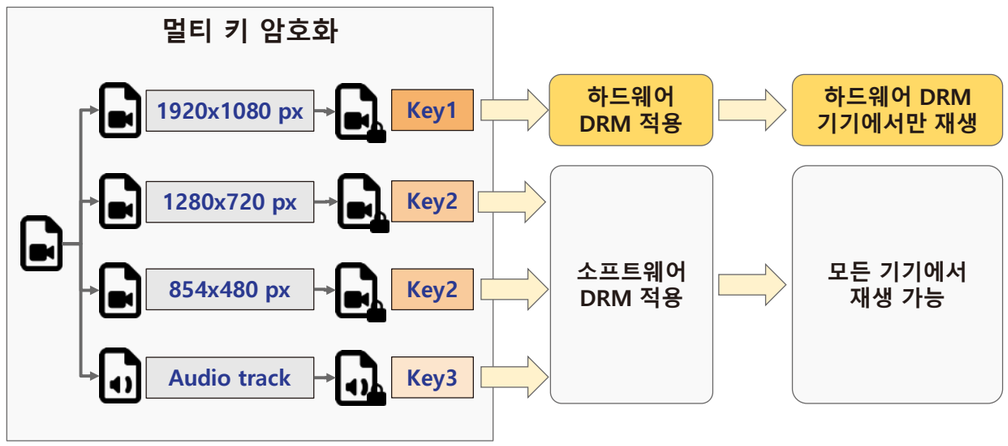
소프트웨어 기반 DRM과 하드웨어 DRM



Adaptive Bitrate 스트림의 단일 키 암호화 예시

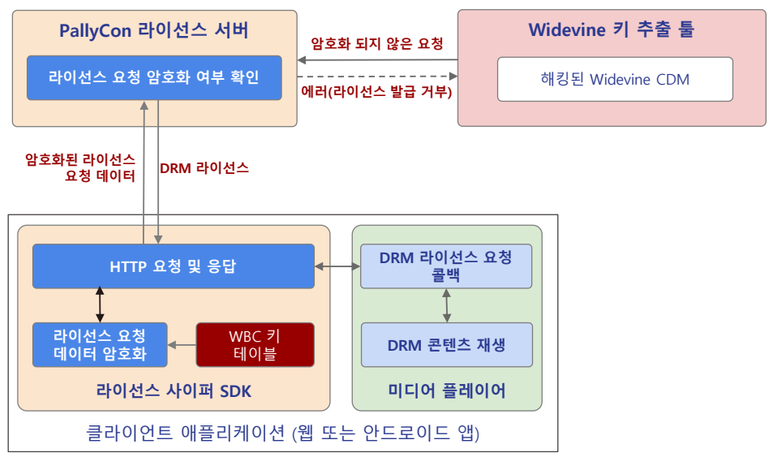


Adaptive Bitrate 스트림의 멀티 키 암호화 예시

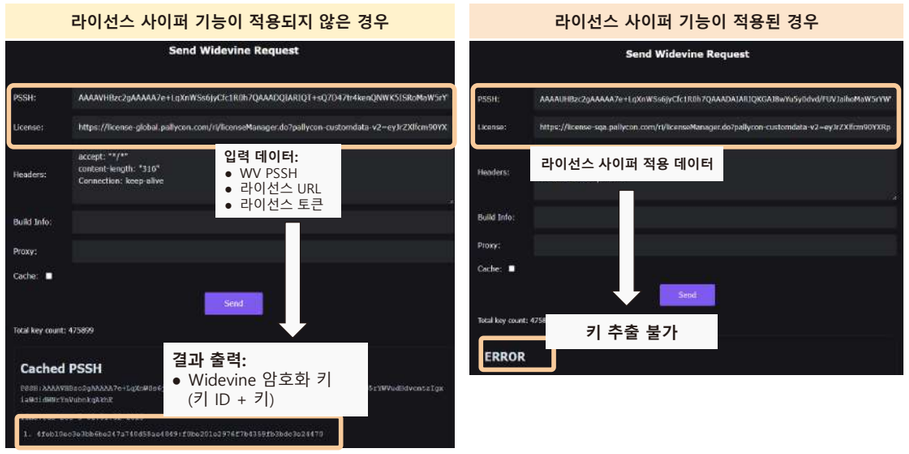


소프트웨어 DRM의 취약성 보완 방안

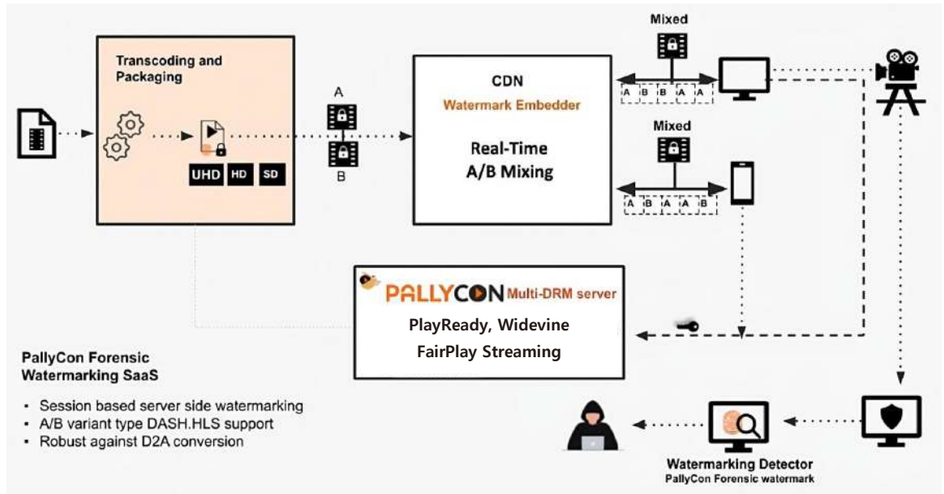
DRM 라이선스 발급 제어를 통한 키 유출 방지



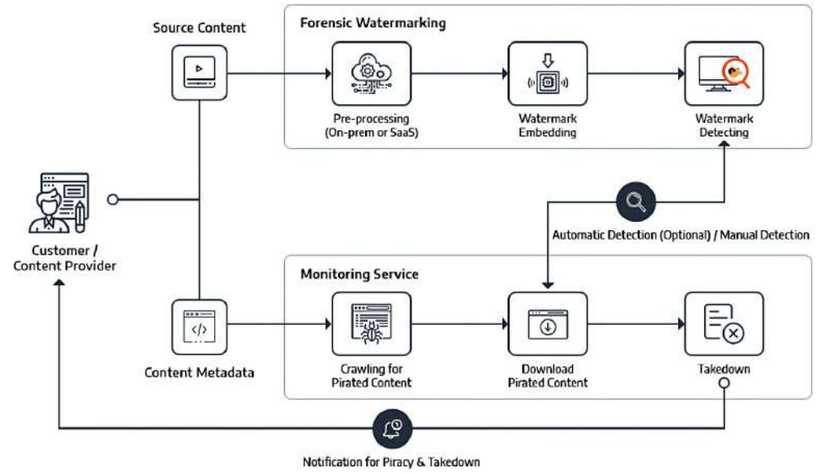
라이선스 사이퍼 동작 확인 – CDRM Project



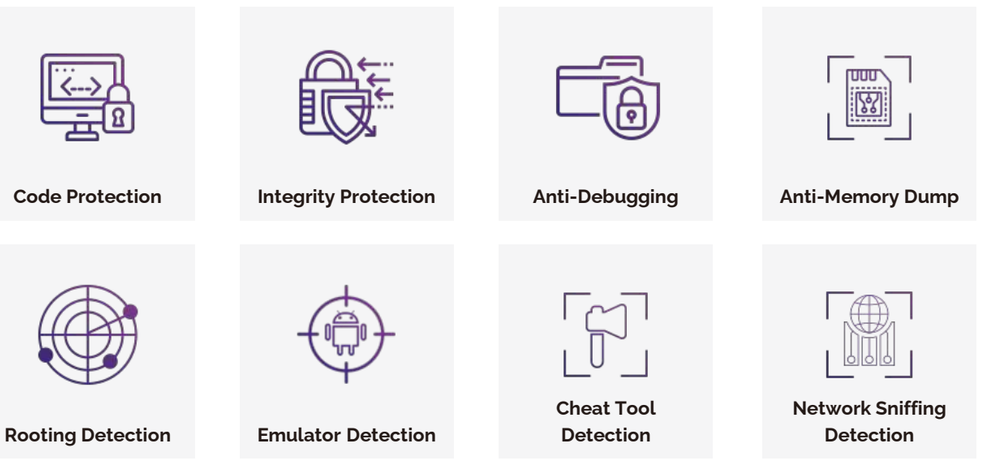
포렌식 워터마킹을 통한 불법 유출자 추적



Anti Piracy 서비스를 통한 불법 유출 대응

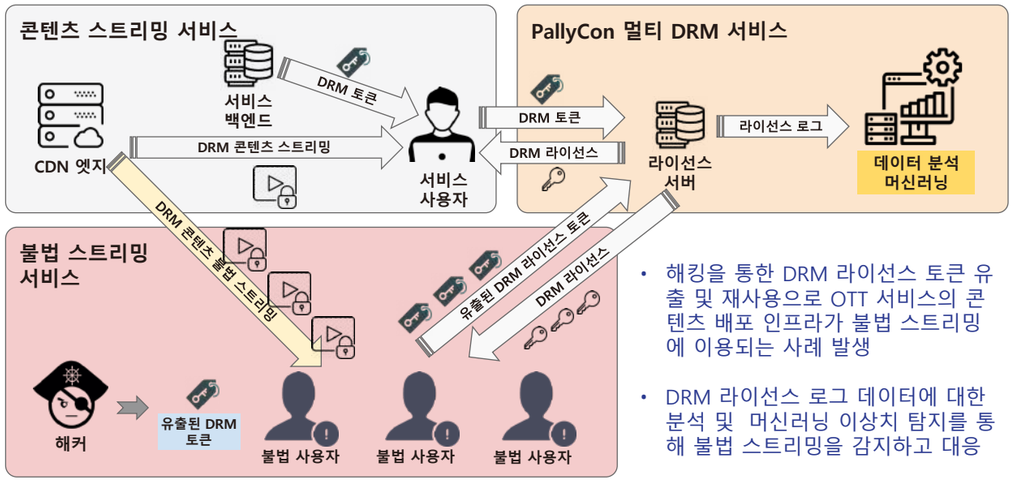


OTT 서비스 애플리케이션 보안



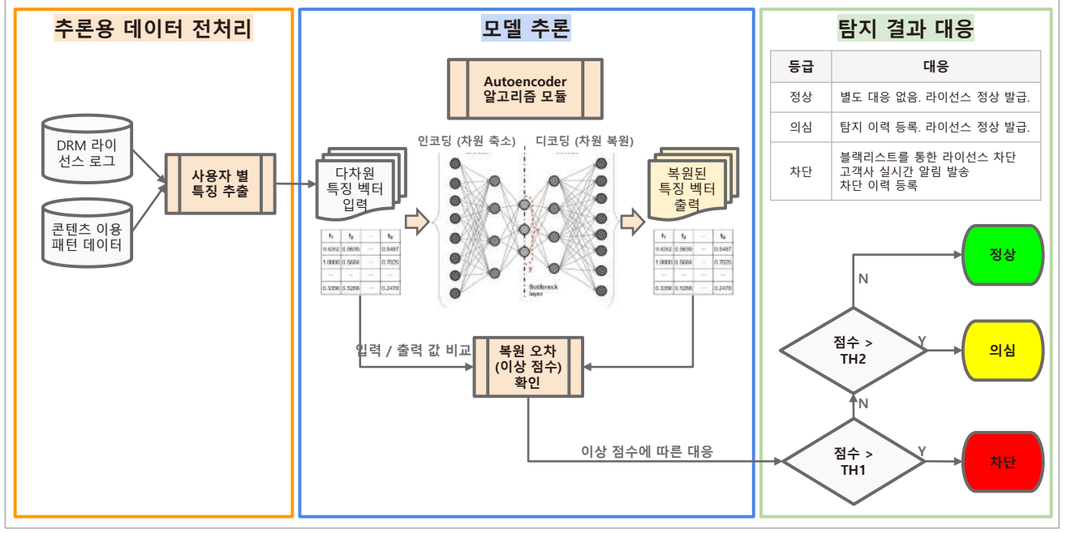
머신러닝 이상치 탐지 모델을 활용한 콘텐츠 불법 유출 대응

AI/ML을 통한 불법 스트리밍(CDN Leeching) 감지

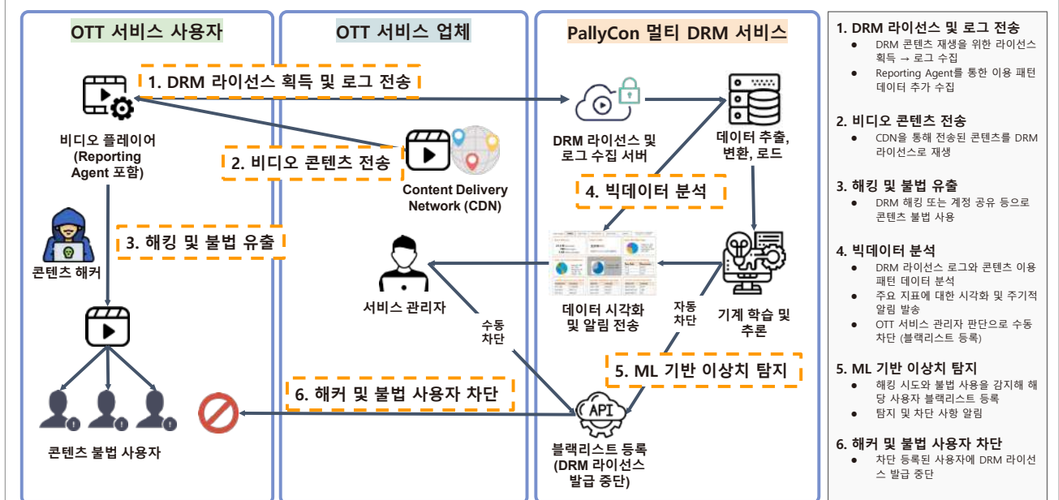


- 해킹을 통한 DRM 라이선스 토큰 유출 및 재사용으로 OTT 서비스의 콘텐츠 배포 인프라가 불법 스트리밍에 이용되는 사례 발생
- DRM 라이선스 로그 데이터에 대한 분석 및 머신러닝 이상치 탐지를 통해 불법 스트리밍을 감지하고 대응

Autoencoder 모델에 의한 이상 탐지 및 대응 프로세스

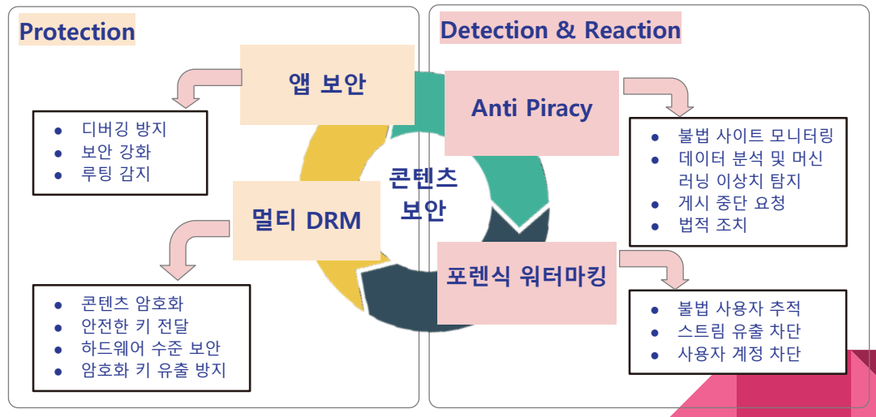


데이터 분석과 머신 러닝 기반의 콘텐츠 불법복제 감지 서비스



결론

멀티미디어 콘텐츠 보안의 패러다임 확장



코로나 팬데믹과 콘텐츠 불법 유출

Protect	Detect	React
		
소프트웨어 DRM 하드웨어 DRM 애플리케이션 보안	DRM 데이터 분석 ML 기반 유출 감지 포렌식 워터마킹	Anti Piracy 불법 사용자 차단

감사합니다

www.pallycon.com

Session 2 콘텐츠 창작의 토대, 저작권 보호 기술

Ⅳ 콘텐츠 분석 및 워터마킹을 통한 이미지 복제 감지

마테이스 두즈
메타 연구원

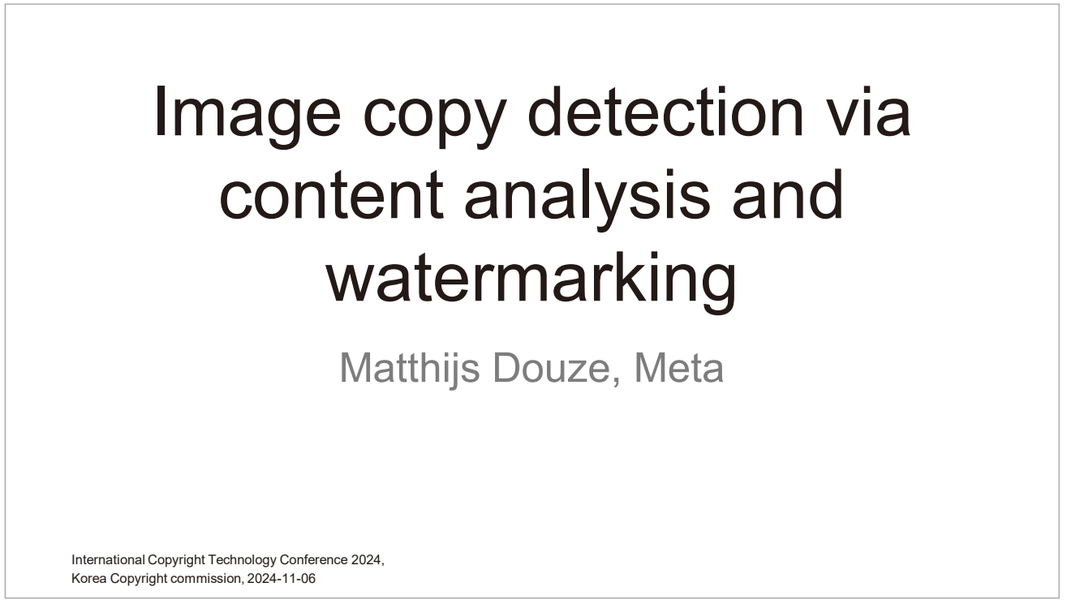
연사 이력

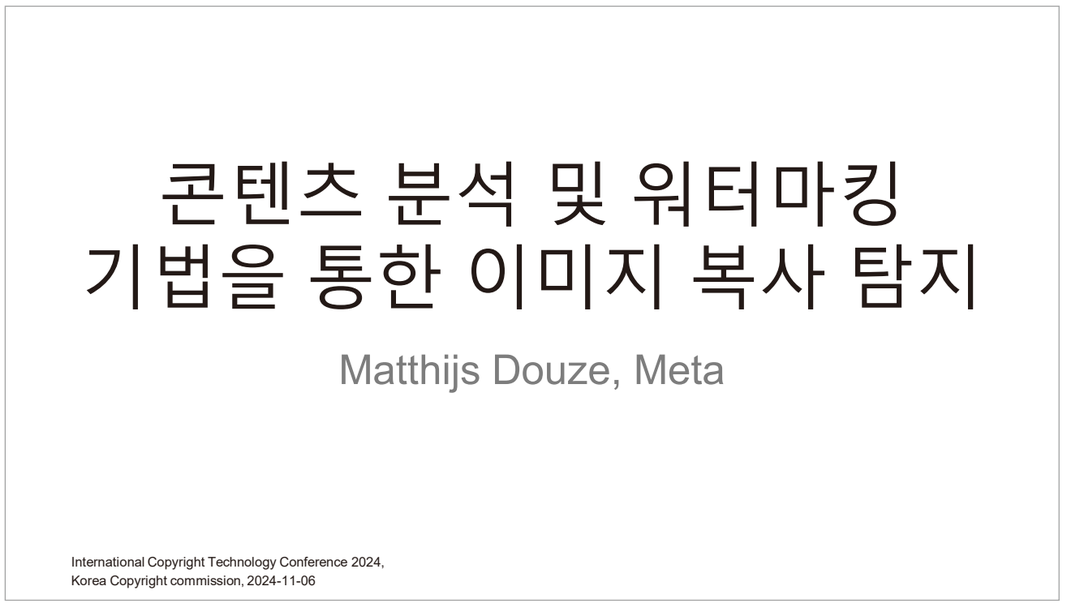
- 2000-2004 툴루즈 대학교 (프랑스) 박사
- 2005-2015 INRIA 그르노블 (프랑스) 연구 엔지니어
- 2016-현재 메타 연구원

발표 내용

소셜 미디어에서 이미지 복제 탐지는 플랫폼에서 저작권이 있는 콘텐츠의 확산을 방지하기 위한 중요한 작업입니다. 이 발표에서는 복제 탐지에 대한 두 가지 접근 방식을 제시합니다. 첫 번째 접근 방식에서는 효율적인 중간 표현 방식인 SSCD를 사용하여 이미지를 저작권이 있는 콘텐츠와 비교합니다. 두 번째 접근 방식에서는 AI가 생성한 이미지의 약간의 변경을 통해 그 출처를 추적하는 방법을 보여줍니다.

In social media, image copy detection is an important task to avoid the dissemination of copyrighted content on the platforms. In this talk we will present two approaches to copy detection. In the first approach, images are compared with copyrighted content using an efficient intermediate representation, SSCD. In the second approach, we show how slight alterations of an image generated by an AI make it possible to trace back where they come from.





About me

Matthijs Douze

- At Facebook since 2015, Paris
- Fundamental AI Research

similarity search

- embeddings
- unsupervised learning
- Image watermarking

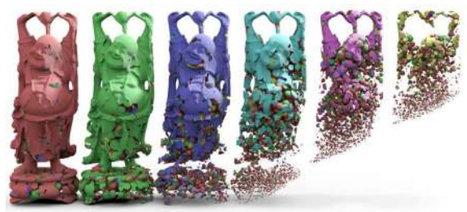
10 years at INRIA (research institution)

- + large-scale 3D reconstruction

인물소개

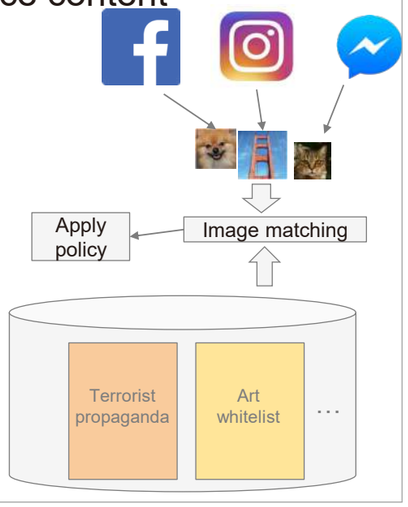
Matthijs Douze

- 2015년부터 파리 페이스북에서 근무
- 기본 AI 연구
- 유사성 검색
- 임베딩
- 비지도 학습
- 이미지 워터마킹
- INRIA(연구 기관)에서 10년 근무
- 대규모 3D 재구성



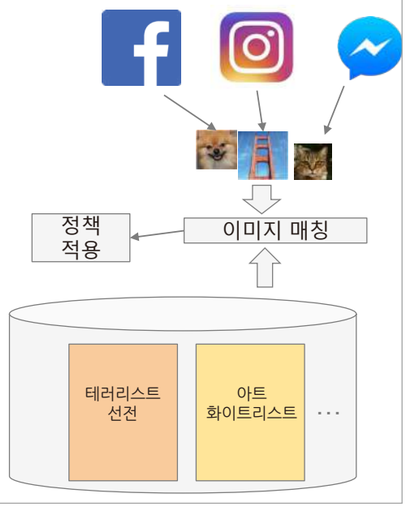
Copy detection infrastructure: reference content

- Database of content (aka. “bank”)
  - Collection of images
  - Associated policy → take down, downvote, escalate to human review, exception
  - Used for human review & automated processing
- Automated processing
  - Incoming images are compared with bank images
  - When match → apply policy
  - Challenge: **image matching**



복사 감지 인프라: 참조 콘텐츠

- 콘텐츠 데이터베이스(일명 “bank")
  - 이미지 컬렉션
  - 연관 정책 → 삭제, 비추천, 인간 검토로 전환, 예외
  - 인간 검토 및 자동 처리에 사용
- 자동 처리
  - 인커밍 이미지와 데이터베이스 이미지 비교
  - 일치할 경우 → 정책 적용
  - 도전 과제: 이미지 매칭



Copy detection infrastructure: watermarking

- Previous approach is “passive”
  - Does not affect image content
- **Active** approach
  - Imperceptible modification of the image
  - Carries information that can be recovered
- Applications
  - Controlled diffusion
  - Images posted and re-posted on different platforms
  - AI generated image content

A Self-Supervised Descriptor for Image Copy Detection

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal and Matthijs Douze, CVPR'22

Meta AI

복사 감지 인프라: 워터마킹

- 이전 접근 방식은 "수동적"
  - 이미지 콘텐츠에 영향을 미치지 않음
- **능동적** 접근 방식
  - 이미지의 눈에 띄지 않는 수정
  - 복구 가능한 정보 전달
- 응용 프로그램
  - 제어된 확산
  - 다른 플랫폼에 게시 및 재게시된 이미지
  - AI가 생성한 이미지 콘텐츠

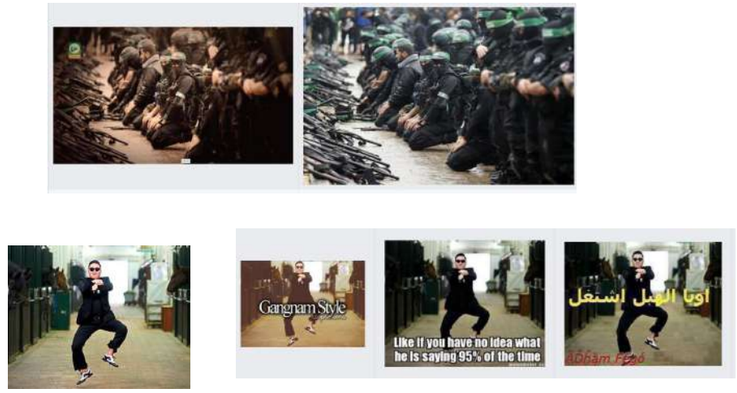
이미지 복사 탐지를 위한 자기 지도 디스크립터

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal and Matthijs Douze, CVPR'22

Meta AI

“Easy” recognition case

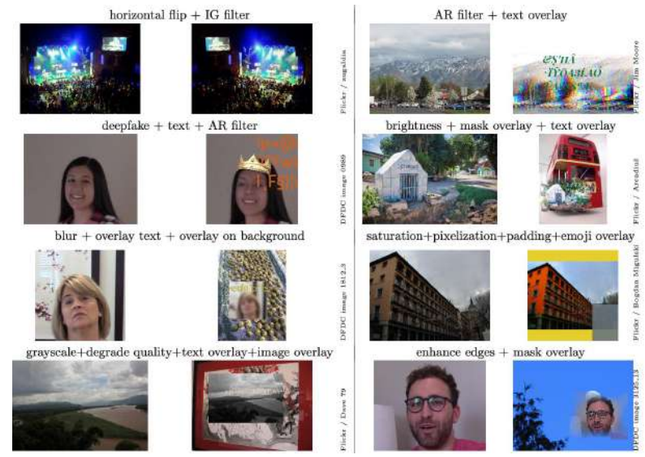
Mild crops, overlays and photometric changes



Motivation: dataset for the Image Similarity Challenge (DISC2021)

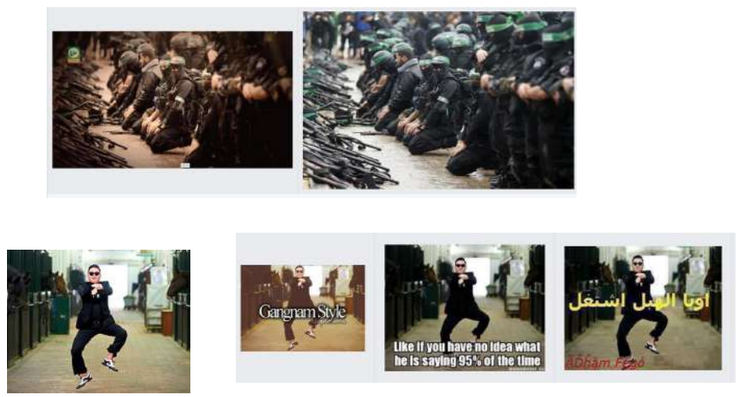
Strong image transformations

1M images to find



"쉬운" 인식 사례

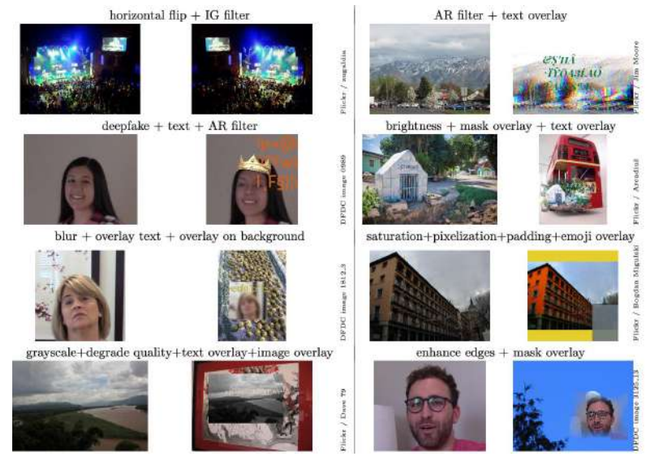
마일드 크롭(부드럽게 자르기), 오버레이(겹쳐 놓기) 및 광학 변화



동기: 이미지 유사성 과제 데이터 세트(DISC2021)

강력한 이미지 변환

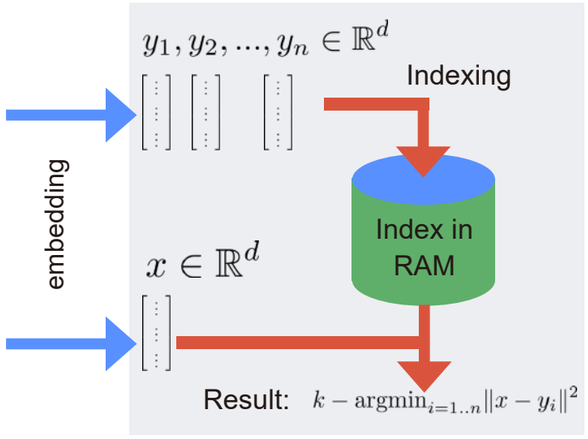
찾아야 할 1M 이미지



Approach: image embeddings

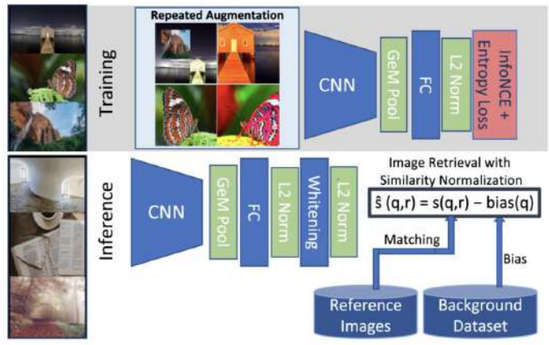
Collection:

Query:



Challenge: the embedding extraction

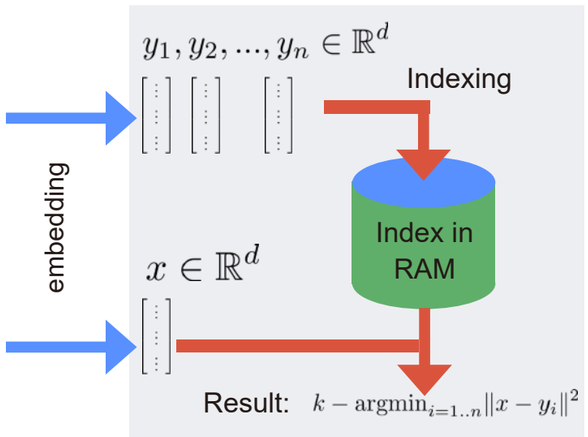
Embedding extraction



접근 방식: 이미지 임베딩

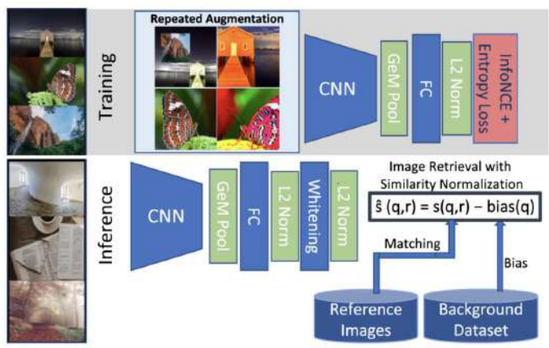
Collection:

Query:



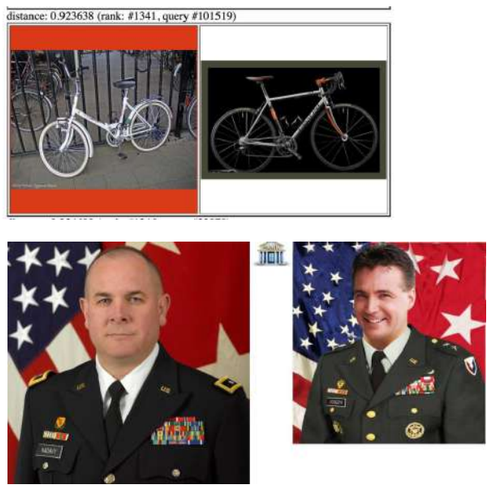
Challenge: the embedding extraction

임베딩 추출



Challenges: abusive generalization

- Let's use an off-the shelf resnet-50 or CLIP [10]
- classification models are
  - too invariant
  - Not discriminant
- When they don't "know" CNNs "fire" on textures

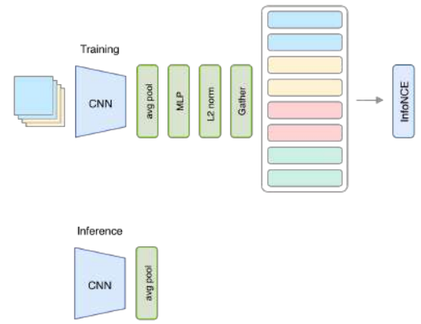


Background: SimCLR

- Contrastive learning objective:
  Learns by training on matching image copies
- Embedding MLP for matching copies is discarded for inference
- Contrastive InfoNCE loss

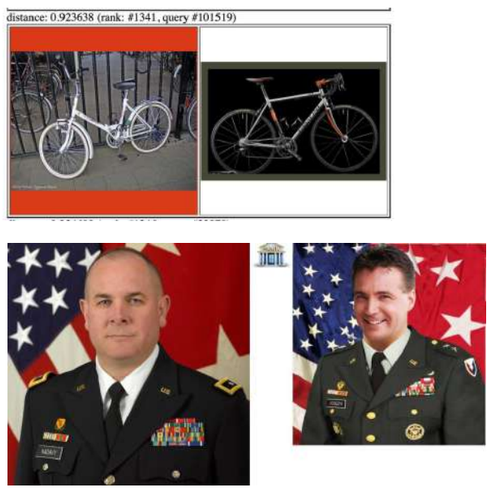
$$\ell_{i,j} = -\log \frac{\exp(s_{i,j})}{\sum_{k \neq i} \exp(s_{i,k})}$$

$$\mathcal{L}_{\text{InfoNCE}} = \frac{1}{|P|} \sum_{i,j \in P} \ell_{i,j}.$$



과제: 부적절한 일반화

- 기성의 ResNet-50이나 CLIP을 사용합시다. [10]
- 분류 모델들이
  - 너무 분별적임
  - 구별력이 없음
- 모델이 "모른다" 판단 시 CNN이 텍스처에서 반응

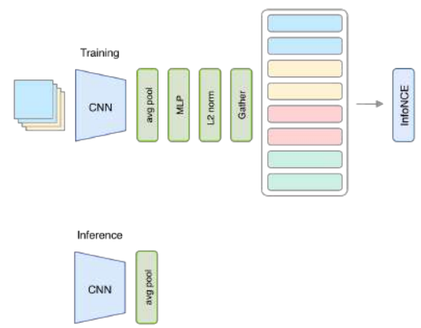


배경: SimCLR

- 대조 학습 목표:
  일치하는 이미지 복사본을 학습을 통해 배움
- 복사본을 일치시키기 위한 임베딩 MLP는 추론을 위해 버려짐
- 대조적 InfoNCE 손실

$$\ell_{i,j} = -\log \frac{\exp(s_{i,j})}{\sum_{k \neq i} \exp(s_{i,k})}$$

$$\mathcal{L}_{\text{InfoNCE}} = \frac{1}{|P|} \sum_{i,j \in P} \ell_{i,j}.$$



Key motivations

- Optimize self-supervised contrastive learning for copy detection
- Create a calibrated descriptor, such that distance has a similar meaning throughout the space
  - Use threshold: distance below threshold ⇒ match

SimCLR for copy detection

SimCLR for copy detection adaptations:

- generalized mean (GeM) pooling
- strengthening the blur augmentation
- using a lower InfoNCE softmax temperature
- using a simple linear projection to 512d

We call this $SimCLR_{CD}$.

name	method	dimensions	µAP	µAPSN
SimCLR	trunk features	2048	13.1	33.9
	+ GeM pooling	2048	21.5	45.3
SimCLR	projection	128	9.4	17.3
	+ GeM pooling	128	11.1	18.8
	+ strong blur	128	14.1	26.0
	+ low temp	128	26.0	41.5
	+ 512d	512	27.5	43.5
$SimCLR_{CD}$	+ linear proj	512	33.0	51.6

주요 동기

- 복사 탐지를 위한 자기 지도 대조적 학습 최적화
- 거리의 의미가 전체 공간에서 유사하도록 조정된 설명자를 생성
  - 적용: 임계값 이하의 거리 ⇒ 일치

복사 탐지를 위한 SimCLR

복사 탐지를 위한 SimCLR 조정 사항:

- 일반화된 평균(GeM) 풀링
- 블러 증강 강화
- 낮은 InfoNCE 소프트맥스 온도 사용
- 간단한 선형 프로젝션을 통한 512차원 변환

이를 $SimCLR_{CD}$.라 칭함

이름	방법	차원	µAP	µAPSN
SimCLR	trunk features	2048	13.1	33.9
	+ GeM pooling	2048	21.5	45.3
SimCLR	projection	128	9.4	17.3
	+ GeM pooling	128	11.1	18.8
	+ strong blur	128	14.1	26.0
	+ low temp	128	26.0	41.5
	+ 512d	512	27.5	43.5
$SimCLR_{CD}$	+ linear proj	512	33.0	51.6

Part 2: Calibrated descriptor distance

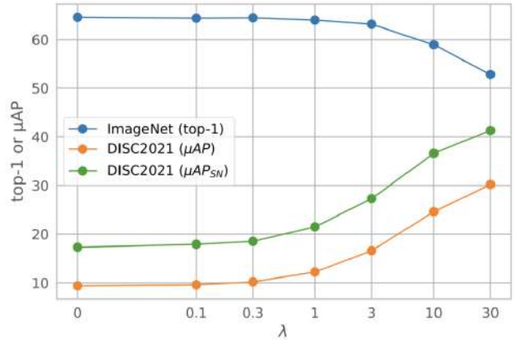
Descriptor spaces vary in density.

The meaning of descriptor distance varies based on local density.

A calibrated descriptor would provide a uniform notion of distance.

SimCLR + differential entropy

SimCLR with varying differential entropy regularization strengths λ (and no other changes)



$$\mathcal{L}_{\text{KoLeo}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\min_{j\neq\hat{P}_i}\|z_i - z_j\|\right)$$

$$\mathcal{L}_{\text{basic}} = \mathcal{L}_{\text{InfoNCE}} + \lambda\mathcal{L}_{\text{KoLeo}}$$

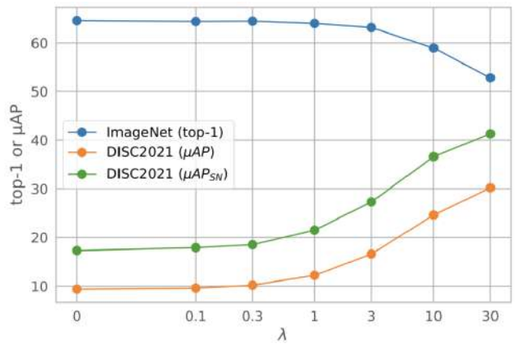
Part 2: 보정된 설명자 거리

설명자 공간은 밀도에 따라 다르게 나타난다.
설명자 거리의 의미는 지역 밀도에 따라 변동한다.
보정된 설명자는 일관된 거리 개념을 제공한다.

SimCLR + 이산 엔트로피

다양한 이산 엔트로피 정규화 강도 λ를 갖는 SimCLR

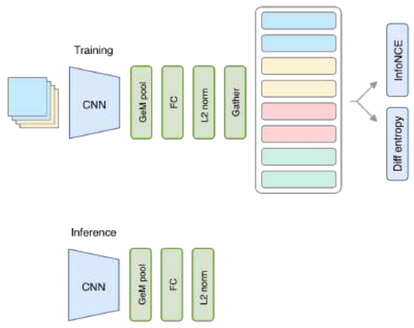
(및 기타 변경 사항 없음)



$$\mathcal{L}_{\text{KoLeo}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\min_{j\neq\hat{P}_i}\|z_i - z_j\|\right)$$

$$\mathcal{L}_{\text{basic}} = \mathcal{L}_{\text{InfoNCE}} + \lambda\mathcal{L}_{\text{KoLeo}}$$

SSCD: SimCLR$_{CD}$ + differential entropy



SSCD combines SimCLR$_{CD}$ optimizations with differential entropy regularization

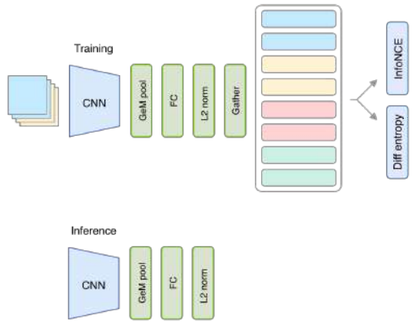
model	μAP	μAP_{SN}	recall@1	MRR
SimCLR$_{CD}$	33.0	51.6	58.6	60.5
$\lambda = 1$	33.1	51.9	58.7	60.9
$\lambda = 3$	38.0	56.1	62.9	65.1
$\lambda = 10$	45.3	61.5	67.7	69.5
$\lambda = 30$	50.4	64.5	69.8	71.4

Additional experiments

- Additional augmentations
  - Rotations, Emoji, Text
  - MixUp and CutMix to model collages
- Datasets
  - Training on DISC dataset (reduce domain shift)
  - Evaluate on Copydays dataset
- Larger trunk model

method	trained on	transforms	dims	μAP	μAP_{SN}
Multigrain [7]	ImageNet*		2048	20.5	41.7
DINO [9] †	ImageNet		1500	32.2	53.8
SimCLR [10] trunk	ImageNet	SimCLR	2048	13.1	33.9
SimCLR [10] proj	ImageNet	SimCLR	128	9.4	17.3
SimCLR$_{CD}$ trunk	ImageNet	strong blur	2048	39.8	56.8
SSCD	ImageNet	strong blur	512	50.4	64.5
SSCD	ImageNet	advanced	512	55.5	71.0
SSCD	ImageNet	adv.+mixup	512	56.8	72.2
SSCD	DISC	strong blur	512	54.8	63.6
SSCD	DISC	advanced	512	60.4	71.1
SSCD	DISC	adv.+mixup	512	61.5	72.5
SSCD$_{large}$ †	DISC	adv.+mixup	1024	**63.7**	**75.3**

SSCD: SimCLR$_{CD}$ + 이산 엔트로피



SSCD는 SimCLRCD 최적화와 이산 엔트로피 정규화를 결합한다

model	μAP	μAP_{SN}	recall@1	MRR
SimCLR$_{CD}$	33.0	51.6	58.6	60.5
$\lambda = 1$	33.1	51.9	58.7	60.9
$\lambda = 3$	38.0	56.1	62.9	65.1
$\lambda = 10$	45.3	61.5	67.7	69.5
$\lambda = 30$	50.4	64.5	69.8	71.4

추가 실험

- 추가 증강
  - 회전, 이모지, 텍스트
  - 콜라주를 모델링하기 위한 MixUp 및 CutMix
- 데이터 세트
  - 데이터 세트에 대한 훈련 (도메인 이동 감소)
  - Copydays 데이터 세트에 대한 평가
- 더 큰 트렁크 모델

method	trained on	transforms	dims	μAP	μAP_{SN}
Multigrain [7]	ImageNet*		2048	20.5	41.7
DINO [9] †	ImageNet		1500	32.2	53.8
SimCLR [10] trunk	ImageNet	SimCLR	2048	13.1	33.9
SimCLR [10] proj	ImageNet	SimCLR	128	9.4	17.3
SimCLR$_{CD}$ trunk	ImageNet	strong blur	2048	39.8	56.8
SSCD	ImageNet	strong blur	512	50.4	64.5
SSCD	ImageNet	advanced	512	55.5	71.0
SSCD	ImageNet	adv.+mixup	512	56.8	72.2
SSCD	DISC	strong blur	512	54.8	63.6
SSCD	DISC	advanced	512	60.4	71.1
SSCD	DISC	adv.+mixup	512	61.5	72.5
SSCD$_{large}$ †	DISC	adv.+mixup	1024	**63.7**	**75.3**

Example matches

DISC2021 examples where SSCD's first result is correct, and SimCLR's is not.

SSCD	SimCLR	queries
✓	✓	38.9 %
✓	✗	39.0 %
✗	✓	0.3 %
✗	✗	21.8 %

In conclusion

- Contrastive learning is a promising approach for copy detection, but requires problem-specific tuning
- Copy detection benefits from unusually strong differential entropy regularization
- We argue that uniform distributions are uniquely compatible with copy detection

예시 일치

SSCD의 첫 번째 결과는 정확 사례,
SimCLR의 결과는 정확하지 않은 DISC2021 사례이다

SSCD	SimCLR	queries
✓	✓	38.9 %
✓	✗	39.0 %
✗	✓	0.3 %
✗	✗	21.8 %

결론

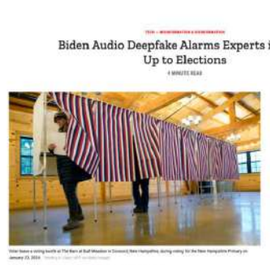
- 대조 학습은 복사 탐지에 있어 유망한 접근법이지만, 특정 문제에 맞게 조정이 필요하다.
- 복사 탐지는 특히 강력한 미분 엔트로피 정규화의 이점을 누린다.
- 균일 분포는 복사 탐지와 독특하게 호환된다고 본다.

Image Watermarking

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon.
"The Stable Signature: Rooting Watermarks in Latent Diffusion Models." ICCV, 2023.

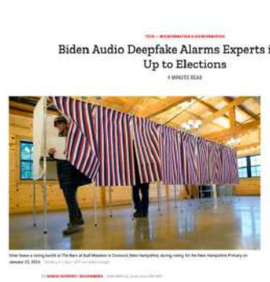
GenAI - Some Emerging Risks

Election manipulation, Scam, Fraud

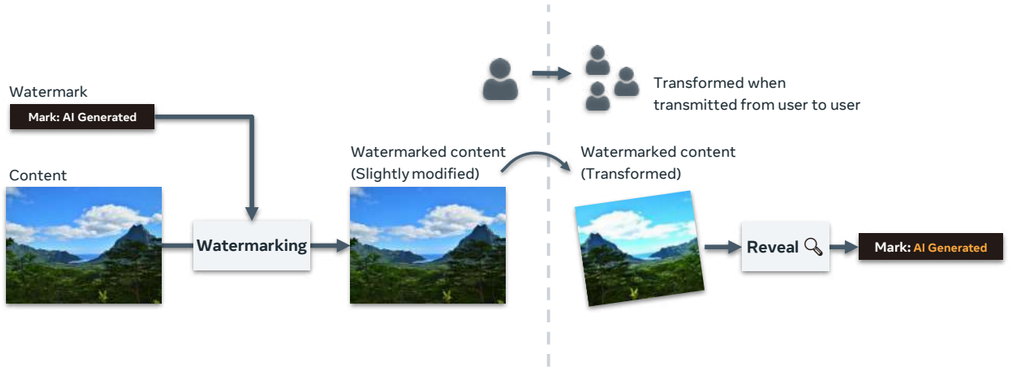


GenAI - 일부 신규 위험

선거 조작, 사기, 부정행위



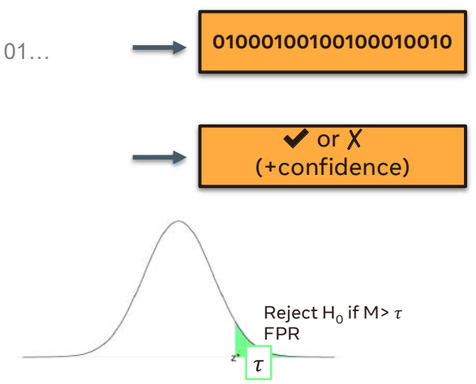
Watermarking



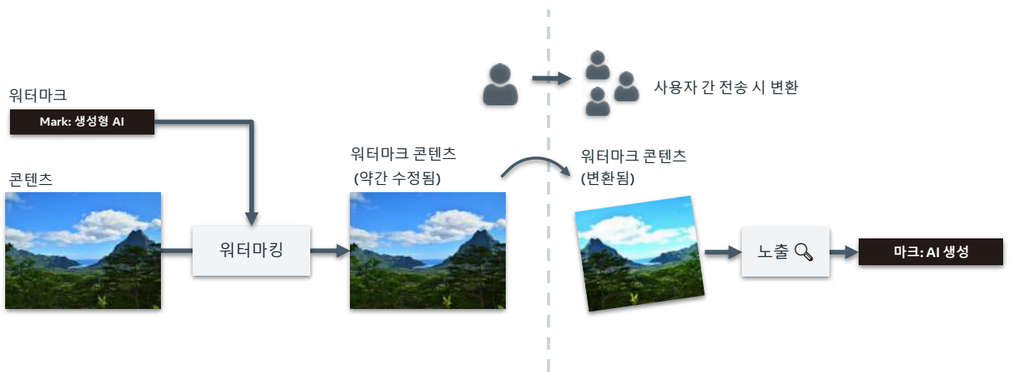
Traditionally used for IP protection.
Can it be used to **enhance detection of GenAI content**? How?

Different Flavors of Watermarking

- **Multi-bit** watermarking:
  Hide and decode a *k*-bits binary vector: 01000101…
- **Zero-bit** watermarking
  Modify the content, Watermarked: ✔ or ✗ ?
- From multi-bit to zero-bit:
  - Compute the number of matching bits *M*
  - Test: $M(m,m') > \tau$



워터마킹

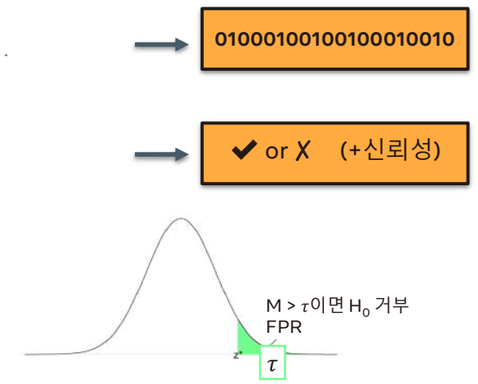


전통적으로 IP 보호에 사용되어 왔다.

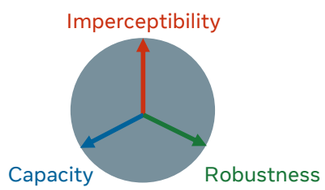
이것이 GenAI 콘텐츠 탐지 개선에도 활용될 수 있을까? 어떻게 가능할까?

워터마킹의 다양한 유형

- **다중 비트** 워터마킹:
  k 비트 이진 벡터 숨기기 및 디코딩: 01000101…
- **제로 비트** 워터마킹
  콘텐츠, 워터마크 수정: ✔ 또는 ✗ ?
- **다중 비트** 워터마킹:
  - 일치하는 비트 수 M 계산
  - 테스트: $M(m,m') > \tau$



Three Criteria for Watermarking



1. Imperceptibility
   Distortion must be low

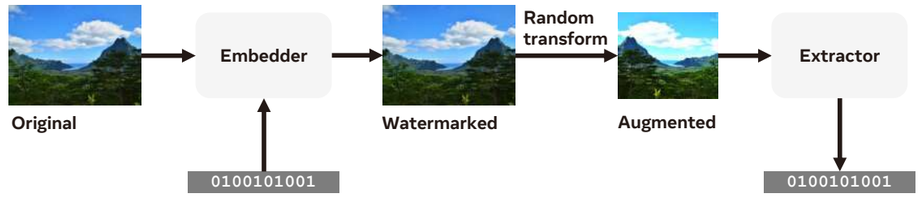
2. Capacity
   The message to hide can be long enough
   ↔ can be detected with high confidence

3. Robustness
   The message must be recovered even when the content is edited

28

Watermarking with Deep Neural Networks

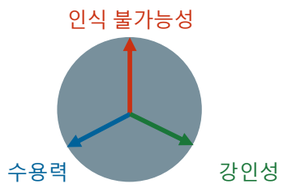
Jointly trains 2 **deep neural networks** to embed/extract watermarks:



[Zhu, Jiren, Russell Kaplan, Justin Johnson, et Li Fei-Fei. « HiDDeN: Hiding Data with Deep Networks ». In ECCV, 2018.]
[Ahmadi, Mahdi, Alireza Norouzi, Nader Karimi, Shadrokh Samavi, and Ali Emami. "ReDMark: Framework for residual diffusion watermarking based on deep networks." Expert Systems with Applications (2020).]

29

워터마킹의 세 가지 기준



1. 인식 불가능성
   왜곡이 낮아야 한다.

2. 수용력
   숨길 메시지가 충분히 길어야 한다.
   ↔ 높은 신뢰도로 탐지가 가능해야 한다.

3. 강인성
   내용이 수정되더라도 메시지를 복원이 가능해야 한다.

28

심층 신경망을 이용한 워터마킹

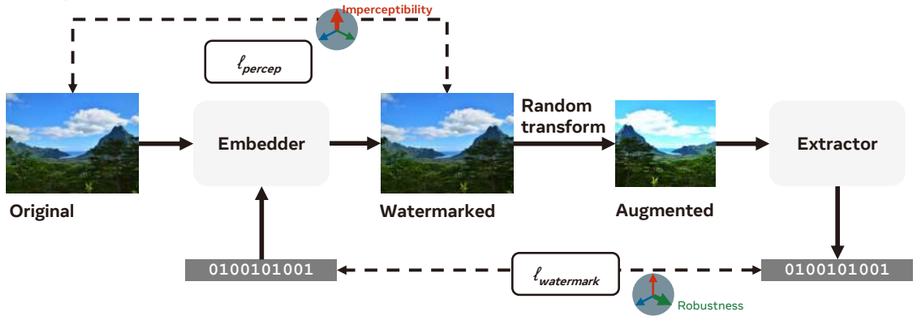
2개의 심층 신경망을 함께 훈련하여 워터마크를 삽입하고 추출:

[Zhu, Jiren, Russell Kaplan, Justin Johnson, et Li Fei-Fei. « HiDDeN: Hiding Data with Deep Networks ». In ECCV, 2018.]
[Ahmadi, Mahdi, Alireza Norouzi, Nader Karimi, Shadrokh Samavi, and Ali Emami. "ReDMark: 딥 네트워크 기반의 잔여 확산 워터마킹 프레임워크." Expert Systems with Applications (2020).]

29

Watermarking with Deep Neural Networks

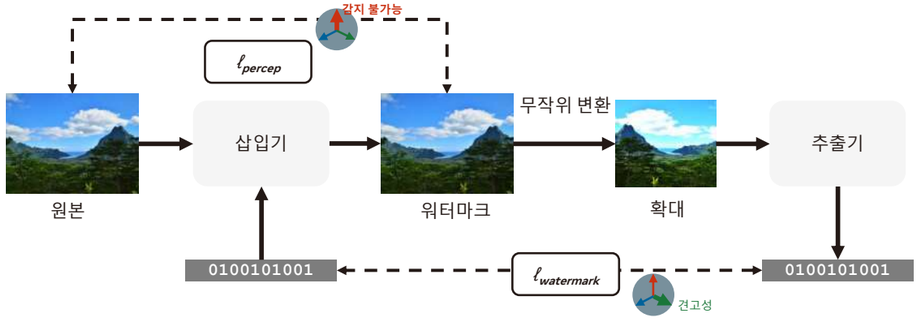
Jointly trains 2 **deep neural networks** to embed/extract watermarks:



[Zhu, Jiren, Russell Kaplan, Justin Johnson, et Li Fei-Fei. « HiDDeN: Hiding Data with Deep Networks ». In ECCV, 2018.]
[Ahmadi, Mahdi, Alireza Norouzi, Nader Karimi, Shadrokh Samavi, and Ali Emami. "ReDMark: Framework for residual diffusion watermarking based on deep networks." Expert Systems with Applications (2020).]

심층 신경망을 이용한 워터마킹

2개의 심층 신경망을 함께 훈련하여 워터마크를 삽입하고 추출:

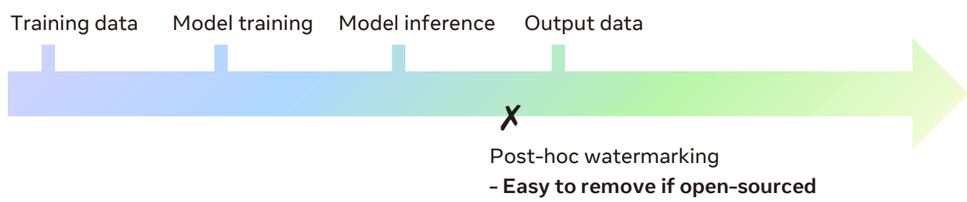


[Zhu, Jiren, Russell Kaplan, Justin Johnson, et Li Fei-Fei. « HiDDeN: 딥 네트워크로 데이터 숨기기». In ECCV, 2018.]
[Ahmadi, Mahdi, Alireza Norouzi, Nader Karimi, Shadrokh Samavi, and Ali Emami. "ReDMark: 딥 네트워크 기반의 잔여 확산 워터마킹 프레임워크." Expert Systems with Applications (2020).]

Example of Stable Diffusion

Stable Diffusion

- Text-to-image model
- Fully open-sourced (training code, inference, model, etc.) in late 2022
- Post-hoc watermarking present in code release (images are generated, then watermarked)

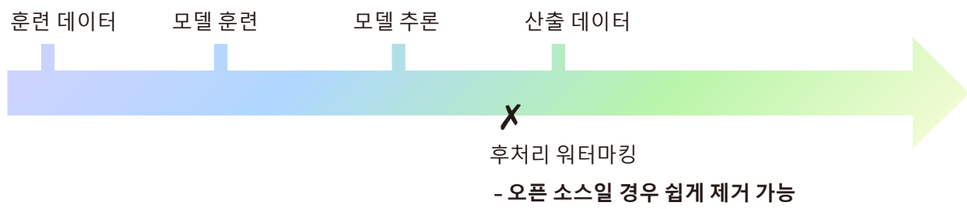


[Rombach et al. , High-Resolution Image Synthesis with Latent Diffusion Models. 2022.]

안정적 확산 예시

안정적 확산

- 텍스트-이미지모델
- 2022년 말에 완전하게 오픈 소스화됨
- 후처리 워터마킹이 코드 배포에 포함됨 (이미지가 생성된 후 워터마크가 추가됨)



[Rombach et al. , High-Resolution Image Synthesis with Latent Diffusion Models. 2022.]

Example of Stable Diffusion

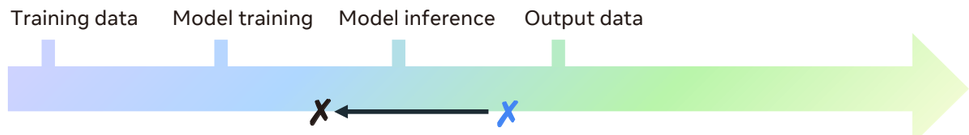
https://github.com/Stability-AI/stablediffusion/blob/main/scripts/txt2img.py#L363

```
img = generator.generate(text_prompt)
# img = wm_encoder.put_watermark(img)
img.save(img_path)
```

32

Stable Signature

Post-hoc watermarking:
+ **More flexible**
- **Easy to remove** if open-sourced



Stable Signature
Rooting watermarks in generative model?

+ **Harder to remove**
+ Works in **open-source** scenarios
+ **Fast** (no need to post-process image)

33

안정적 확산 예시

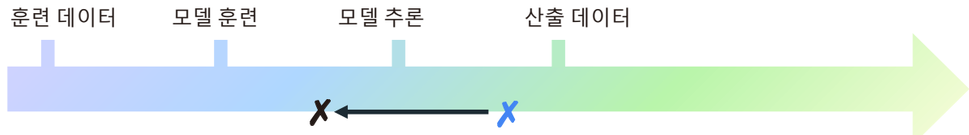
https://github.com/Stability-AI/stablediffusion/blob/main/scripts/txt2img.py#L363

```
img = generator.generate(text_prompt)
# img = wm_인코더.put_watermark(img)
img.save(img_path)
```

32

안정적 서명

후처리 워터마킹:
+ **유연성 증가**
- **오픈 소스인 경우 제거하기 쉬움**



안정적 서명
생성 모델에 워터마크를 통합할 수 있을까요?

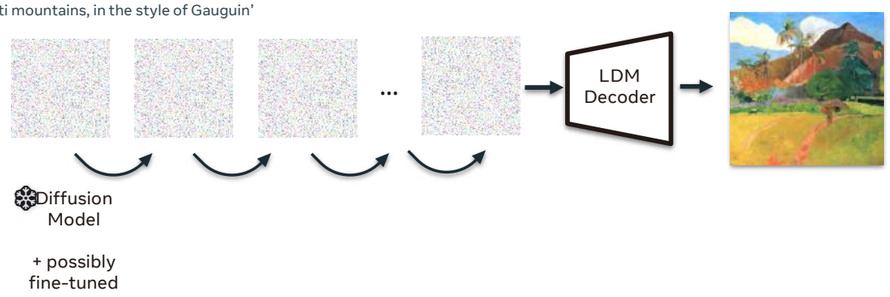
+ **제거하기 더 어려움**
+ **오픈 소스 환경**에서도 작동
+ **빠름**(이미지를 사후 처리할 필요 없음)

33

Stable Diffusion

SD = Big LDM

'Tahiti mountains, in the style of Gauguin'



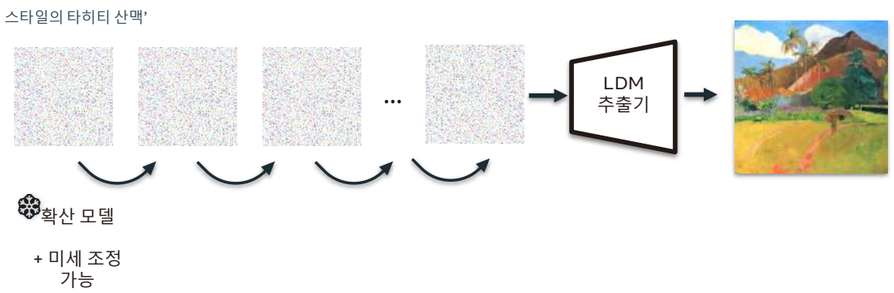
[Rombach et al. , High-Resolution Image Synthesis with Latent Diffusion Models. 2022.]

34

안정적 확산

SD = Big LDM

'고갱 스타일의 타히티 산맥'



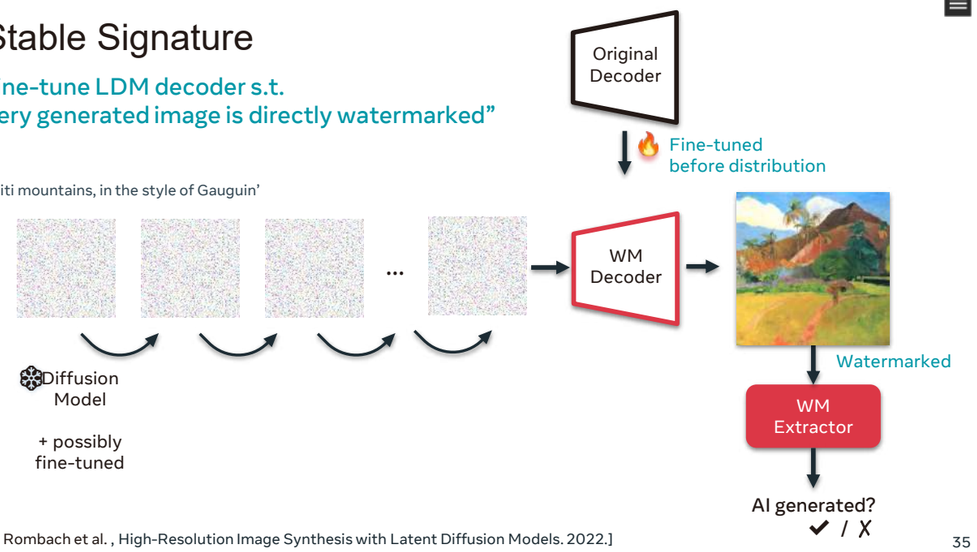
[Rombach et al. , High-Resolution Image Synthesis with Latent Diffusion Models. 2022.]

34

Stable Signature

"Fine-tune LDM decoder s.t.
every generated image is directly watermarked"

'Tahiti mountains, in the style of Gauguin'



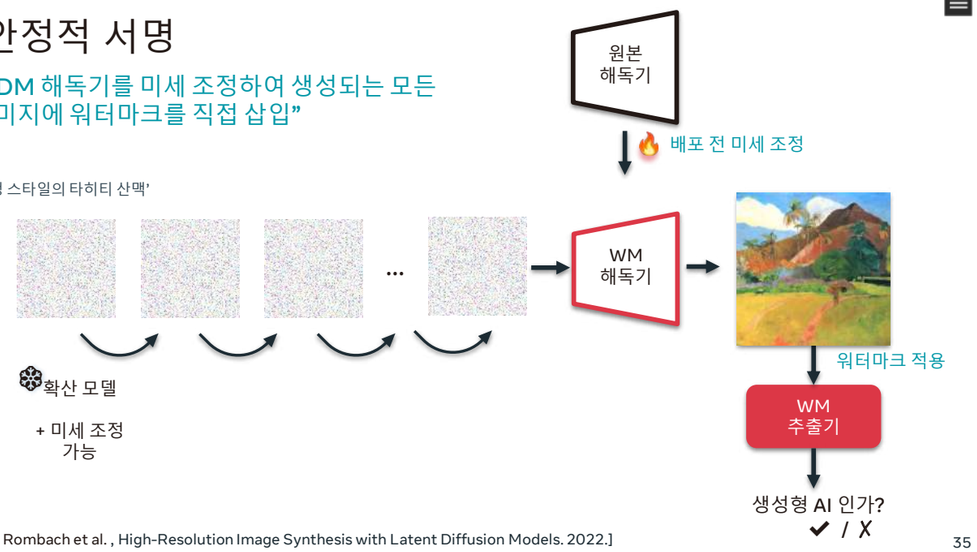
[Rombach et al. , High-Resolution Image Synthesis with Latent Diffusion Models. 2022.]

35

안정적 서명

"LDM 해독기를 미세 조정하여 생성되는 모든
이미지에 워터마크를 직접 삽입"

'고갱 스타일의 타히티 산맥'



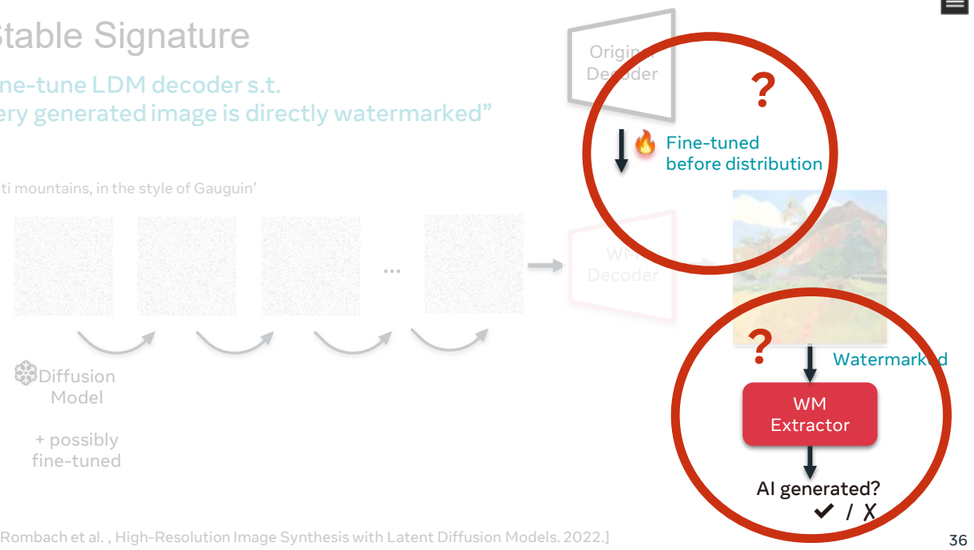
[Rombach et al. , High-Resolution Image Synthesis with Latent Diffusion Models. 2022.]

35

Stable Signature

"Fine-tune LDM decoder s.t.
every generated image is directly watermarked"

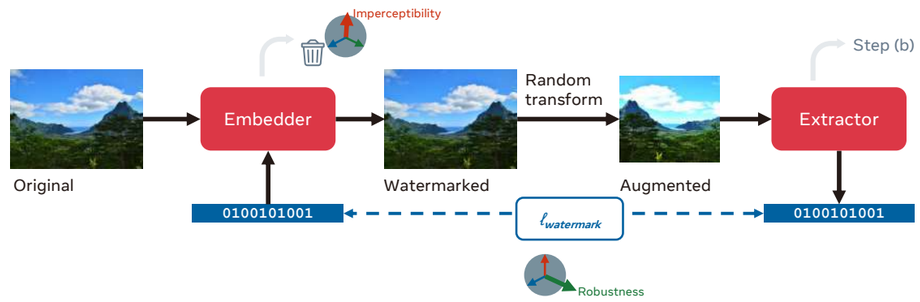
'Tahiti mountains, in the style of Gauguin'



[Rombach et al. , High-Resolution Image Synthesis with Latent Diffusion Models. 2022.]

Stable Signature: Method in 2 Steps

(a) Pre-train watermark encoder/extractor



- 48-bits
- 100k images from COCO, resolution 256x256
- 300 epochs (1 day/ 8 GPUs)

안정적 서명

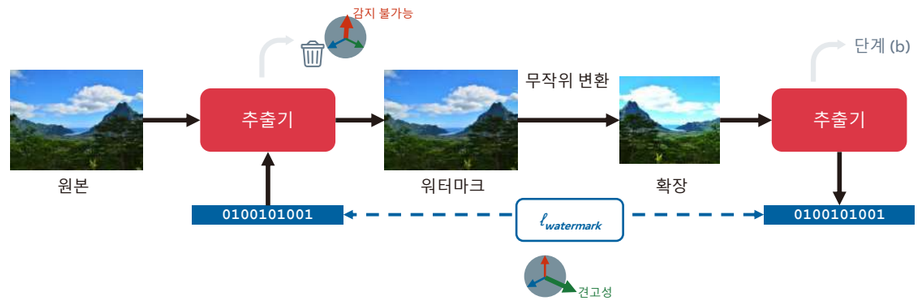
"LDM 해독기를 미세 조정하여 생성되는 모든
이미지에 워터마크를 직접 삽입"

'고갱 스타일의 타히티 산맥'

[Rombach et al. , High-Resolution Image Synthesis with Latent Diffusion Models. 2022.]

안정적 서명: 2단계 방법

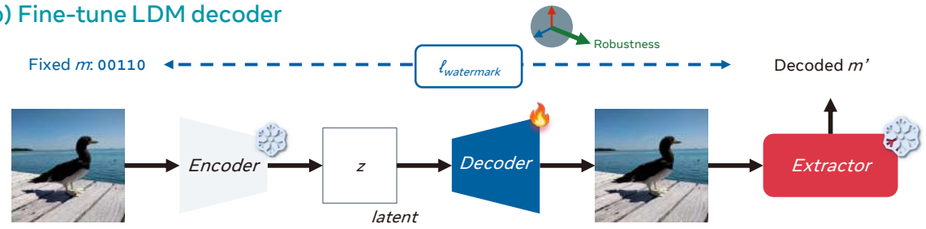
(a) 워터마크 삽입기/추출기 사전 훈련



- 48비트
- COCO에서 가져온 100k 이미지, 해상도 256x256
- 300 에포크(1일/8개 GPU))

Stable Signature: Method in 2 Steps

(b) Fine-tune LDM decoder

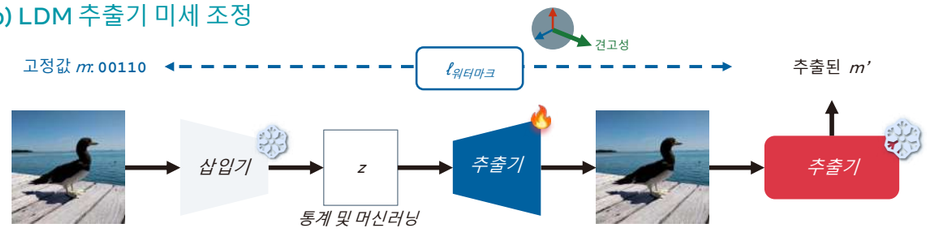


- 100 steps only!
- 400 images, 1min / 1 GPU

38

안정적 서명: 2단계 방법

(b) LDM 추출기 미세 조정

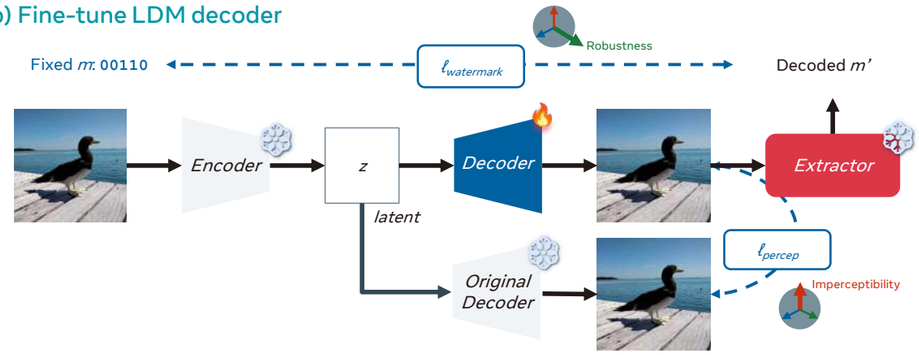


- 100단계만!
- 400개 이미지 당 1분/1 GPU

38

Stable Signature: Method in 2 Steps

(b) Fine-tune LDM decoder

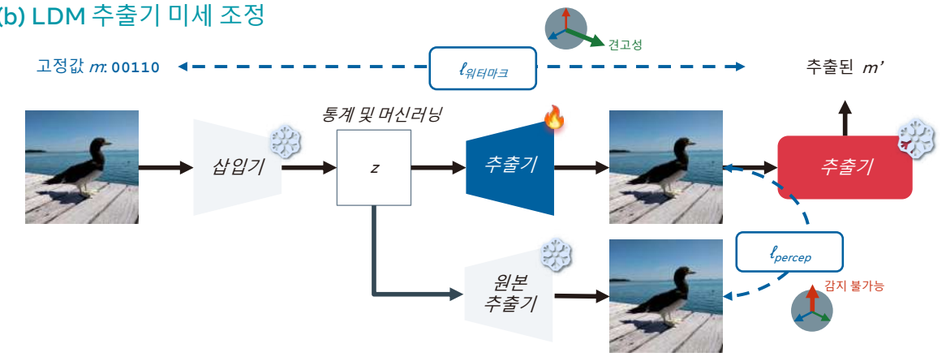


- 100 steps only!
- 400 images, 1min / 1 GPU

39

안정적 서명: 2단계 방법

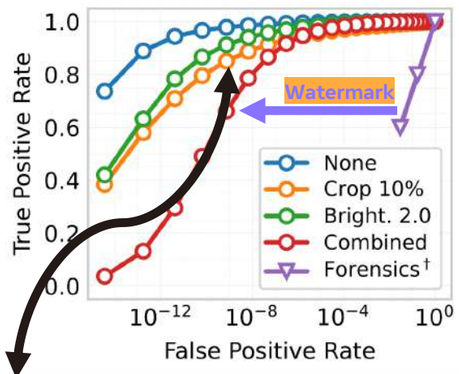
(b) LDM 추출기 미세 조정

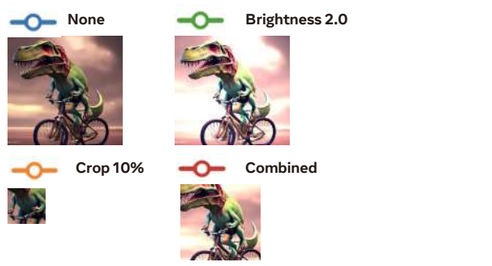


- 100단계만!
- 400개 이미지 당 1분/1 GPU

39

Detection Results





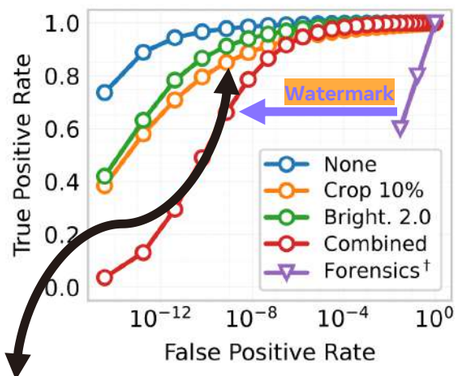
Forensics = passive detection with classifier

[Corvi, Riccardo, 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]

Detect >80% of AI generated images - **flagging 1 natural image every 1B times**

41

감지 결과



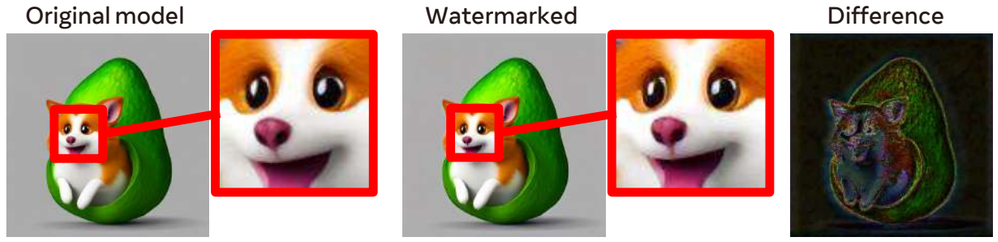
포렌식 = 분류기를 사용한 수동 감지

[Corvi, Riccardo, et al. "확산 모델에서 생성된 합성 이미지 감지에 관하여." ICASSP 2023-2023 IEEE 음향, 음성 및 신호 처리 국제 컨퍼런스(ICASSP). IEEE, 2023]

AI가 생성한 이미지의 80% 이상 감지- **1B번 마다 자연스러운 이미지 1개를 플래그로 표시**

41

Generation Quality



Quantitative results:

10k generated
512x512 images

			PSNR / SSIM ↑	FID ↓
Tasks	Text-to-Image	LDM [68]	30.0 / 0.89	19.6 (−0.3)
	Image Edition	DiffEdit [13]	31.2 / 0.92	15.0 (−0.3)
	Inpainting - Full	Glide [52]	31.1 / 0.91	16.8 (+0.6)
	- Mask only		37.8 / 0.98	9.0 (+0.1)
	Super-Resolution	LDM [68]	34.0 / 0.94	11.6 (+0.0)

42

생성 품질

양적 결과:

10k 생성
512x512 이미지

			PSNR / SSIM ↑	FID ↓
Tasks	Text-to-Image	LDM [68]	30.0 / 0.89	19.6 (−0.3)
	Image Edition	DiffEdit [13]	31.2 / 0.92	15.0 (−0.3)
	Inpainting - Full	Glide [52]	31.1 / 0.91	16.8 (+0.6)
	- Mask only		37.8 / 0.98	9.0 (+0.1)
	Super-Resolution	LDM [68]	34.0 / 0.94	11.6 (+0.0)

42

Conclusion

- Two approaches
  - Passive / active
  - Different use cases
- Tradeoffs
  - True positives vs. false positives
- Thanks!

43

결론

- 두 가지 접근 방식
  - 수동적/능동적
  - 다른 사용 사례
- 트레이드오프
  - 정탐 vs. 오탐
- 감사합니다!

43