DEVIEW 2021

진짜 같은 가짜를 생성하는 기술과 그 가짜를 탐지하는 기술, Deepfake Creation & Detection

정용현 삼성SDS AI 연구센터
김도연 삼성SDS 보안사업부

CONTENTS

1. 딥페이크 정의와 생성 기술

1.1 딥페이크란?

Deepfake = Deep learning + Fake

- 딥러닝과 가짜의 합성어
- Deep Neural Network를 활용하여 생성된 자료
- 이미지, 영상, 음성 등
- Reddit community 2017 등장

1.2 딥페이크 생성 기술

딥페이크는 Generative Model로 생성

Glow: Generative Flow with Invertible 1×1 Convolutions

Diederik P. Kingma*, Prafulla Dhariwal*
OpenAI, San Francisco

Flow based

NVAE: A Deep Hierarchical Variational Autoencoder

Arash Vahdat, Jan Kautz
NVIDIA
{avahdat, jkautz}@nvidia.com

Variational AutoEncoder

Denoising Diffusion Probabilistic Models

Jonathan Ho
UC Berkeley
jonathanho@berkeley.edu

Ajay Jain
UC Berkeley
ajayj@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@cs.berkeley.edu

Score based

Analyzing and Improving the Image Quality of StyleGAN

Tero Karras
NVIDIA

Samuli Laine
NVIDIA

Miika Aittala
NVIDIA

Janne Hellsten
NVIDIA

Jaakko Lehtinen
NVIDIA and Aalto University

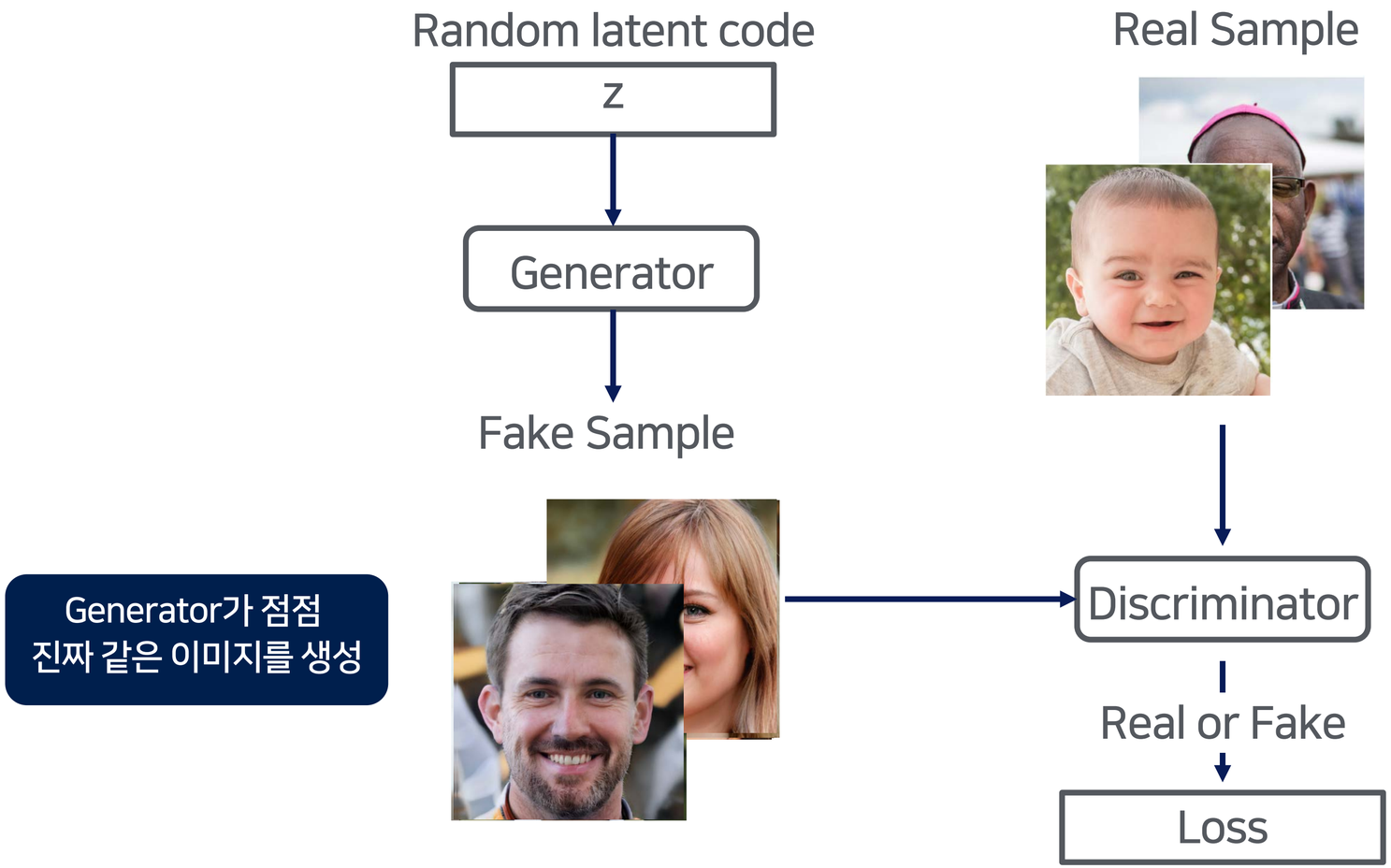
Timo Aila
NVIDIA

GAN

1.3 Generative Adversarial Networks

DEVIEW 2021

Generator와 Discriminator의 경쟁적 학습



※ Image: A Style-Based Generator Architecture for Generative Adversarial Networks (CVPR 2019)

1.3 Generative Adversarial Networks

DEVIEW 2021

Deepfake 생성을 위한 GAN의 3가지 장점

Benefits

High Fidelity Generation

- 사람이 진짜, 가짜 여부를 판단하기 어려운 수준
- 1024x1024의 초 고화질

Benefits

Controllable Generation

- 생성 이미지의 특정 특징을 제어
- 이미지 익명화 및 변조에 활용 가능

Benefits

Fast Samping

- 손쉽고 빠른 이미지 생성

1.3 Generative Adversarial Networks

DEVIEW 2021

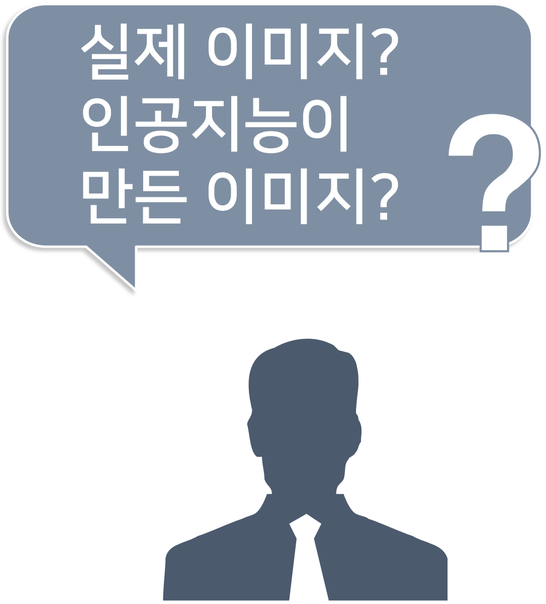
High Fidelity Generation

※ Image: Generative Adversarial Networks (NIPS 2014)
Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks (ICLR 2016)
Coupled Generative Adversarial Networks (NIPS 2016)
Progressive Growing of GANs for Improved Quality, Stability, and Variation (ICLR 2018)
A Style-Based Generator Architecture for Generative Adversarial Networks (CVPR 2020)

1.3 Generative Adversarial Networks

DEVIEW 2021

사람의 판단력: 첫 시도 시 60%, 연습 후 75%



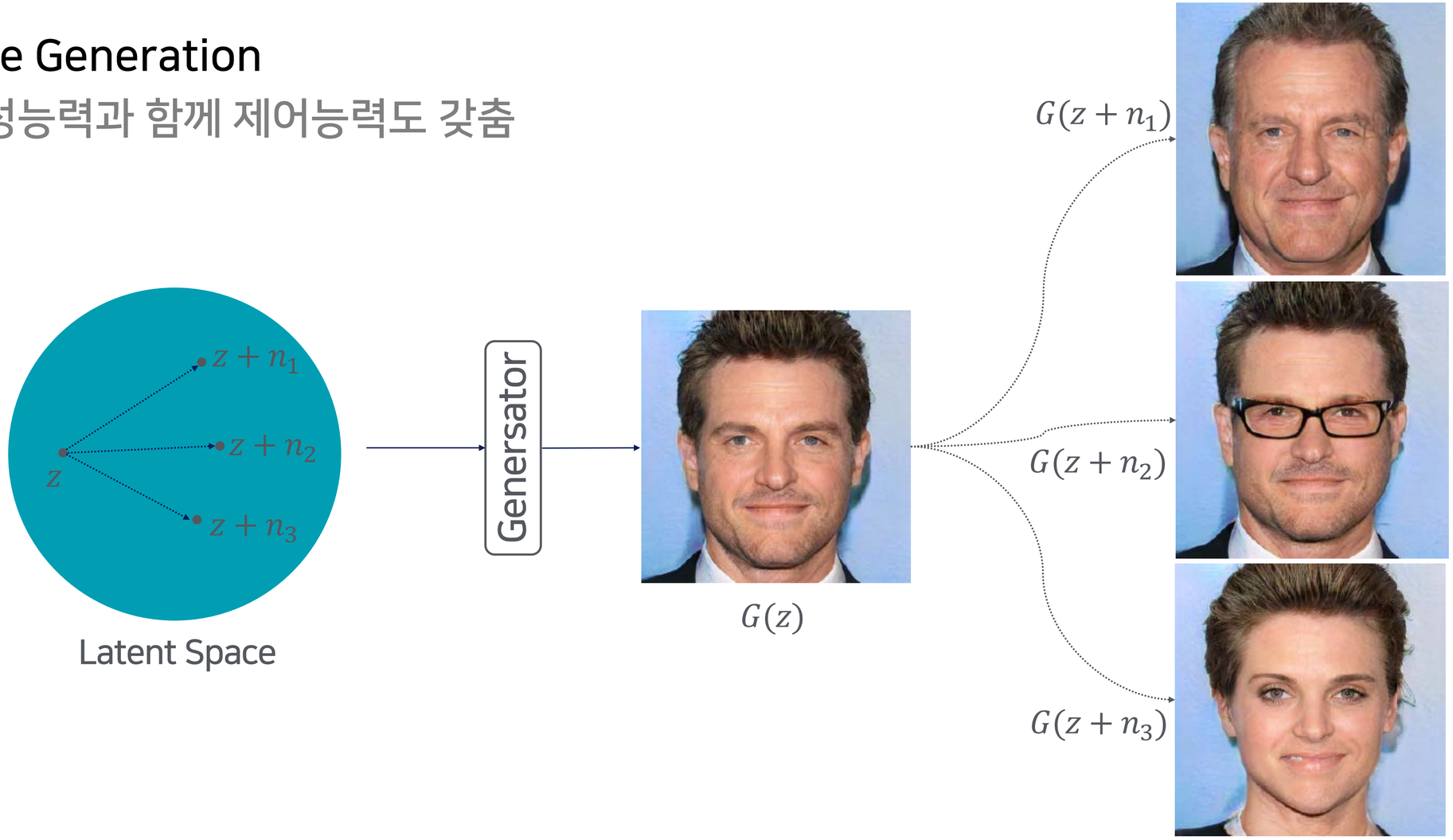
※Image : https://www.thispersondoesnetexist.com

1.3 Generative Adversarial Networks

DEVIEW 2021

Controllable Generation

- GAN은 생성능력과 함께 제어능력도 갖춤



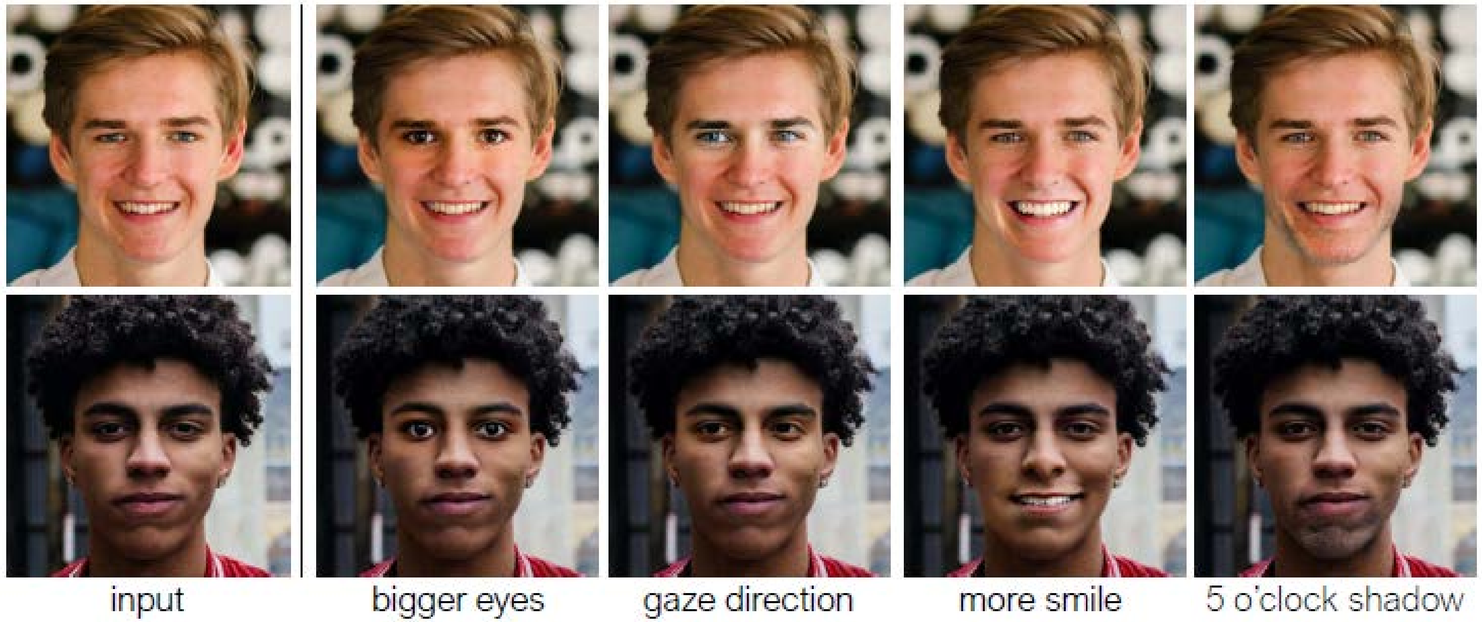
※ Image: Interpreting the Latent Space of GANs for Semantic Face Editing (CVPR 2020)

1.3 Generative Adversarial Networks

DEVIEW 2021

Controllable Generation

- 이미지의 부분적인 특징만 자연스럽게 변조



※ Image Swapping Autoencoder for Deep Image Manipulation (NeurIPS 2020)

1.3 Generative Adversarial Networks

DEVIEW 2021

Face Swapping

※ Video: Advancing high fidelity identity swapping for forgery detection (CVPR 2020)

2.이미지 딥페이크 활용

2.1 딥페이크의 긍정적 활용

사례: AI 김주하 앵커

※ Video: MBN News 2020.11.6 (https://www.youtube.com/watch?v=k8X_Em-NQn0)

2.1 딥페이크의 긍정적 활용

딥페이크 가상 인물의 장점과 활용

- AI 앵커는 24시간 언제든 뉴스 브리핑이 가능.
- 섭외 비용 절감.
- AI 뱅커 및 가상 인물을 통한 영상 제작 등 활용.
- 나를 닮은 가상 캐릭터 생성 (메타버스).

관련 기사 예

- 가상인간 OO 벌써 10억 벌었다, '신명품' 가방까지 접수.
  머니투데이 2021.09.27
- OO은행, 금융권 최초 'AI 뱅커' 도입한다. 2021.09.28
  라이선스 뉴스
- 팔로어만 9만8000명… 패션 아이콘된 '가상모델'
  동아일보 21.10.01

2.2 개인정보 보호를 위한 생성 기술

DEVIEW 2021

유럽 및 우리나라 등 개인정보 보호법 시행

유럽연합의 개인정보 보호법 (GDPR)

- 2016년 5월 제정
- 2018년 5월 25일 시행
- 법 위반 기업은 세계 연 매출 4% 또는 2천만 유로 중 큰 금액을 과징금으로 부과

우리나라 개인정보 보호법

- 2018년 ICT 발전에 따른 보호와 활용을 위한 개정안 발의
- 법 위반 기업은 매출의 최대 3% 과징금과 함께 2년 이하 징역 또는 2천만원 이하의 벌금형 선고
- 2020년 8월 데이터 3법 개정

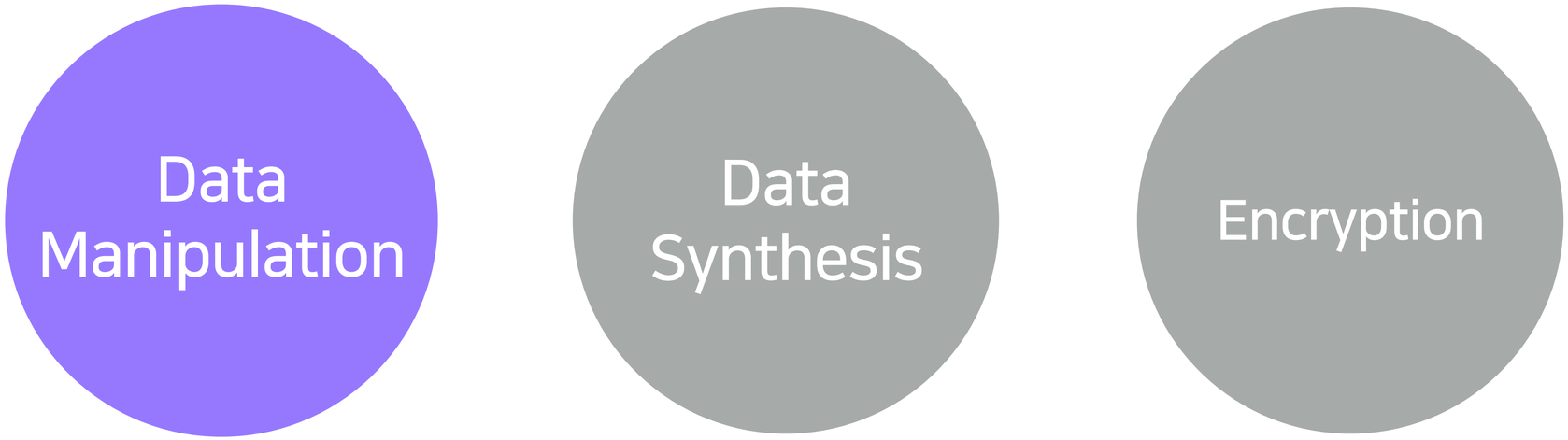
2011년 9월 20일 개인정보보호법 시행과 제정 10주년을 맞아 '제1회 개인정보 보호의 날' 기념 행사 개최

가명 처리한 데이터를 활용

2.2 개인정보 보호를 위한 생성 기술

데이터 가명 처리 방법들

- Data Manipulation: 개인정보 속성 조작 (예, De-identification)
- Data Synthesis : 데이터 분포를 학습하여 가짜 데이터 생성 (예, 차등정보보호)
- Encryption: 데이터 암호화 (예, 동형 암호)



2.3 얼굴 이미지 비식별화

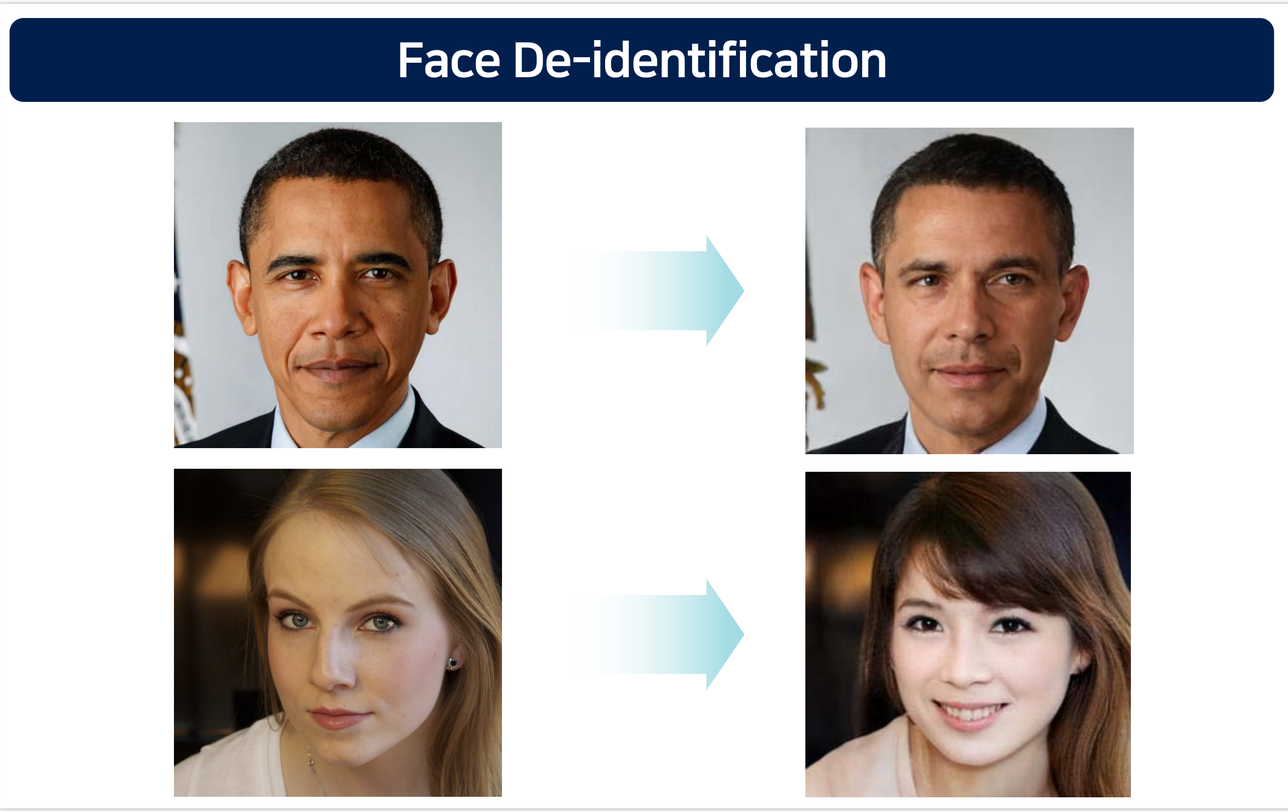
DEVIEW 2021

얼굴 이미지 비식별화(또는 익명화)

- 얼굴 비식별화(De-identification)이란 신원(Identity)과 관련된 속성을 조작하여, 사람의 얼굴의 신원 정보를 제거하는 반면, 포즈, 표정, 그림자, 조명과 같은 신원과 관련 없는 속성은 유지합니다.

주요특징

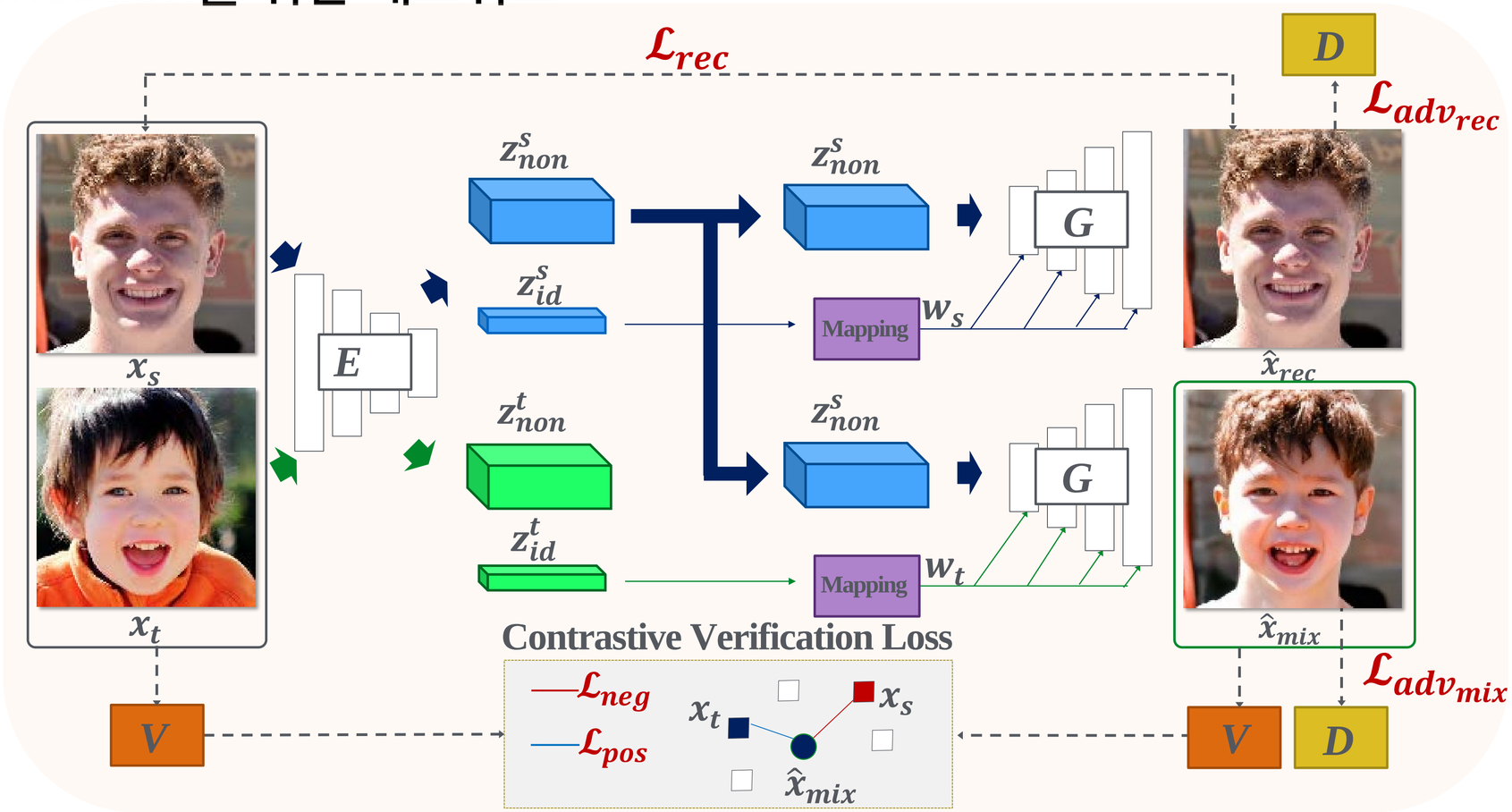
- 개인 정보를 보호
- 얼굴 이미지의 속성 유지를 통한 데이터 분석
- 이미지 내 외부 특성을 유지 (배경, 의상)



※ Paper: FICGAN: Facial Identity Controllable GAN for De-identification (https://arxiv.org/abs/2110.00740)

2.3 얼굴 이미지 비식별화

Face De-identification을 위한 네트워크

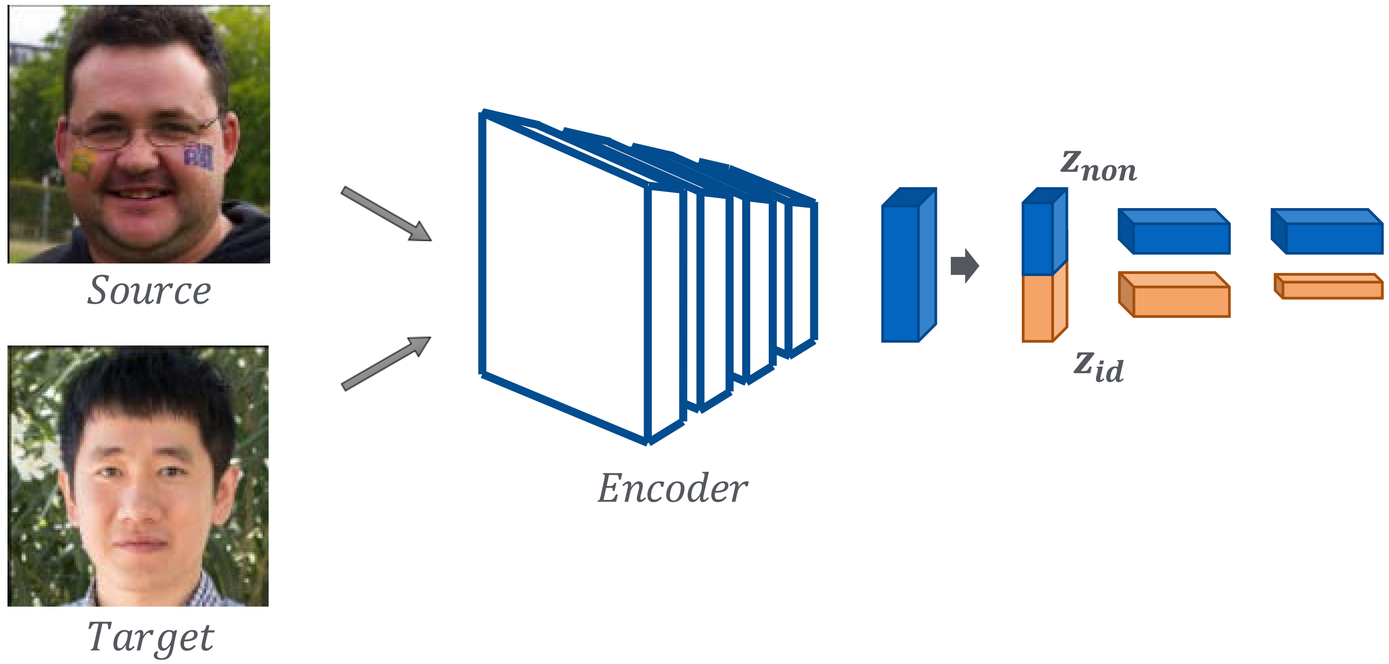


※ Paper: FICGAN: Facial Identity Controllable GAN for De-identification (https://arxiv.org/abs/2110.00740)

2.3 얼굴 이미지 비식별화

얼굴 이미지 비식별화

- Encoder Network를 통한 Identity 와 non-identity의 Disentanglement



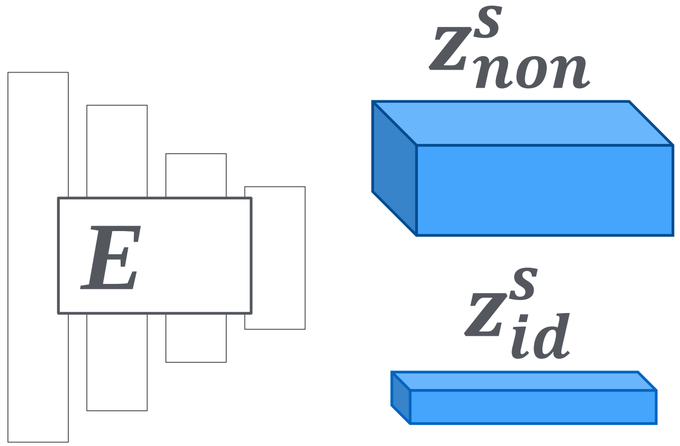
※ Paper: FICGAN: Facial Identity Controllable GAN for De-identification (https://arxiv.org/abs/2110.00740)

2.3 얼굴 이미지 비식별화

얼굴 이미지 비식별화

- Stage1: Source 이미지의 Reconstruction

Source



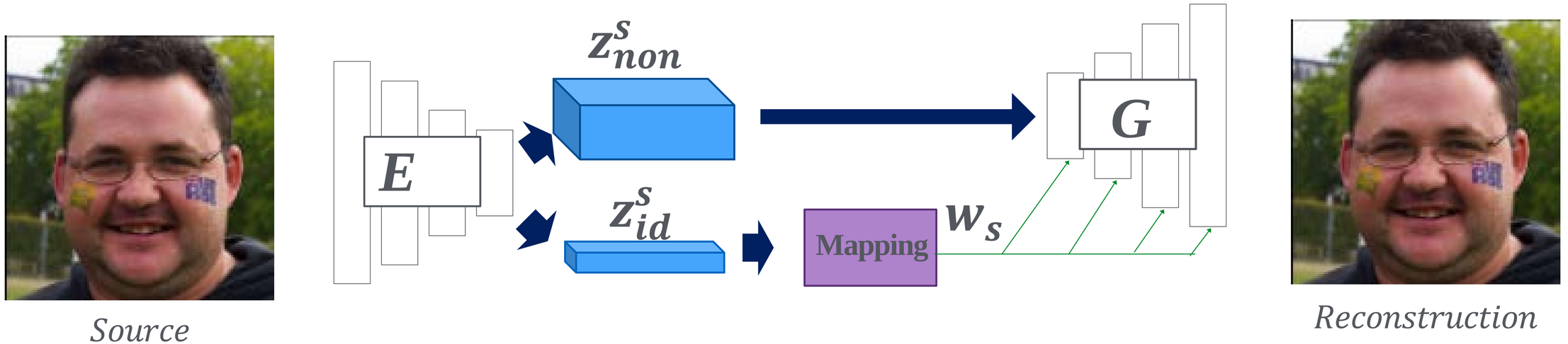
Reconstruction

※ Paper: FICGAN: Facial Identity Controllable GAN for De-identification (https://arxiv.org/abs/2110.00740)

2.3 얼굴 이미지 비식별화

얼굴 이미지 비식별화

- Stage1: Source 이미지의 Reconstruction



Source

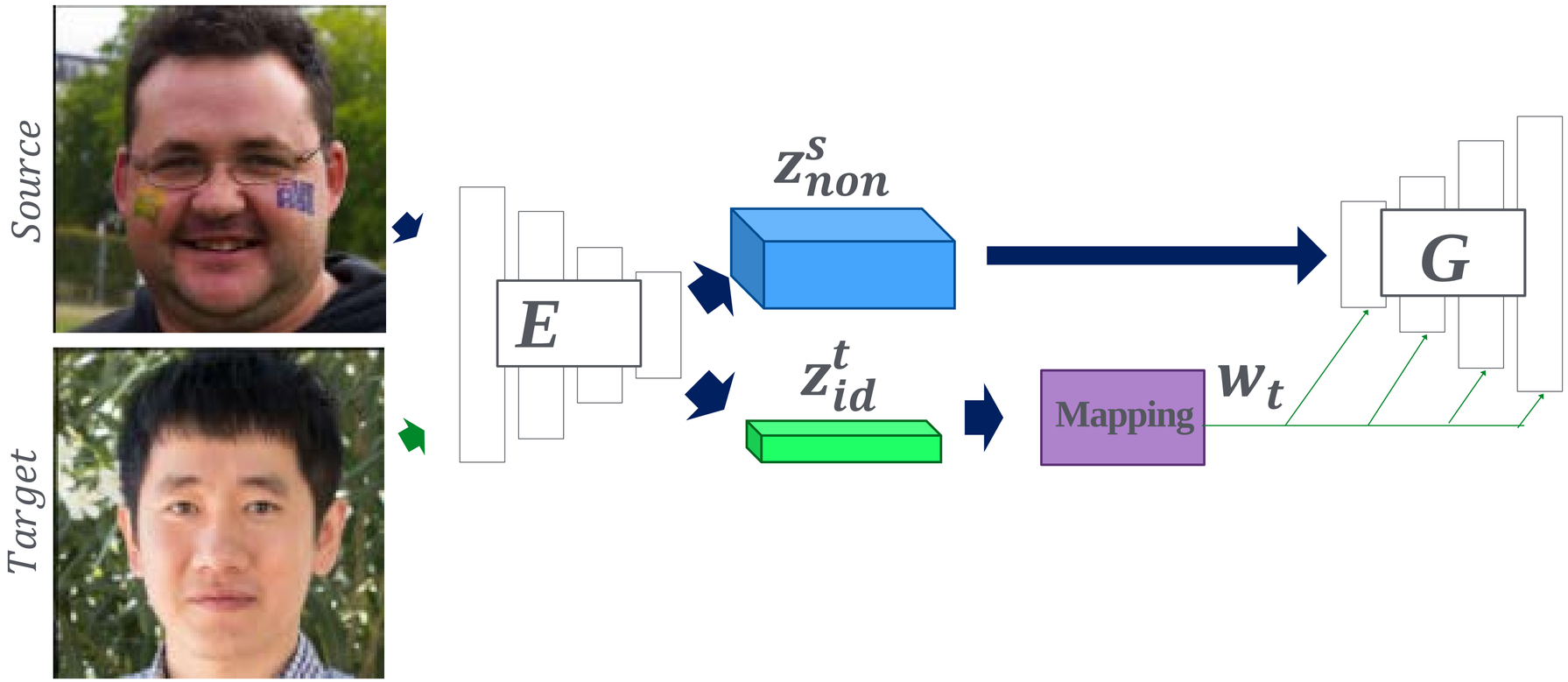
Reconstruction

※ Paper: FICGAN: Facial Identity Controllable GAN for De-identification (https://arxiv.org/abs/2110.00740)

2.3 얼굴 이미지 비식별화

얼굴 이미지 비식별화

- Stage2: Source, Target 이미지의 identity 믹싱



Mixing

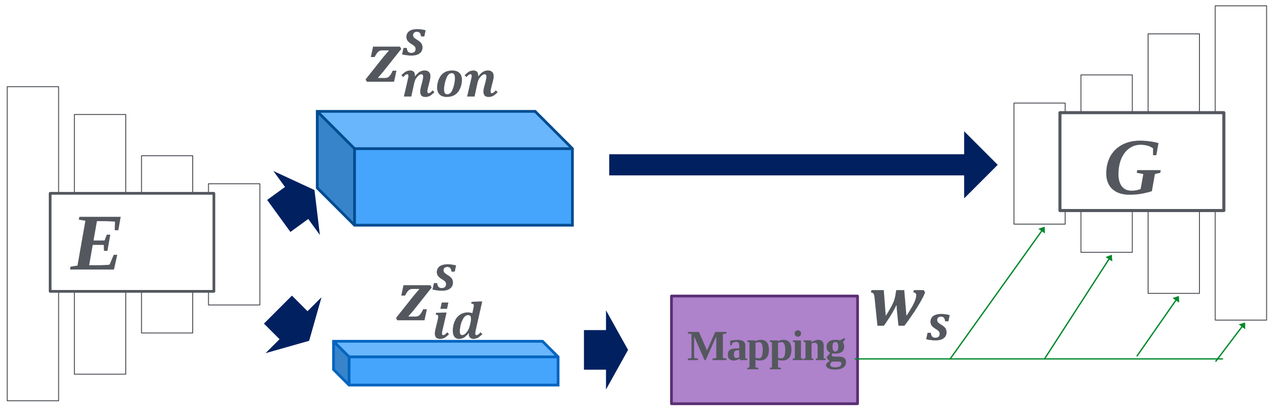
※ Paper: FICGAN: Facial Identity Controllable GAN for De-identification (https://arxiv.org/abs/2110.00740)

2.3 얼굴 이미지 비식별화

얼굴 이미지 비식별화 - Loss

- Stage1: Reconstruction Loss, Adversarial Loss

Source



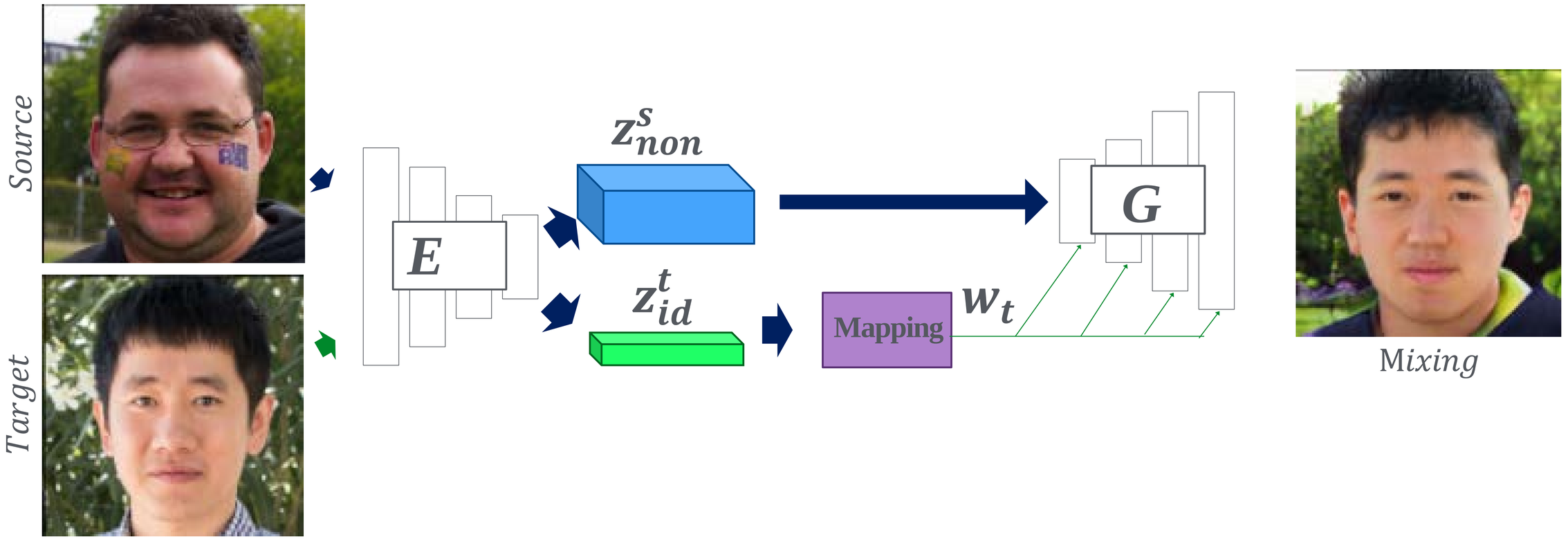
Reconstruction

※ Paper: FICGAN: Facial Identity Controllable GAN for De-identification (https://arxiv.org/abs/2110.00740)

2.3 얼굴 이미지 비식별화

얼굴 이미지 비식별화 - Loss

- Stage2: Adversarial Loss



※ Paper: FICGAN: Facial Identity Controllable GAN for De-identification (https://arxiv.org/abs/2110.00740)

2.3 얼굴 이미지 비식별화

얼굴 이미지 비식별화 – Loss

- Stage2: Contrastive Verification Loss
- L_{contra}= Cos(FaceNet(x_s), FaceNet($\hat{x}_{mix}$))
  +{1- Cos(FaceNet(x_t), FaceNet($\hat{x}_{mix}$))}
- mix된 $\hat{x}_{mix}$는 x_s와 Cosine 유사도가 멀어지도록,
- mix된 $\hat{x}_{mix}$는 x_t와 Cosine 유사도가 가까워지도록.

$Source(x_s)$

$Target(x_t)$

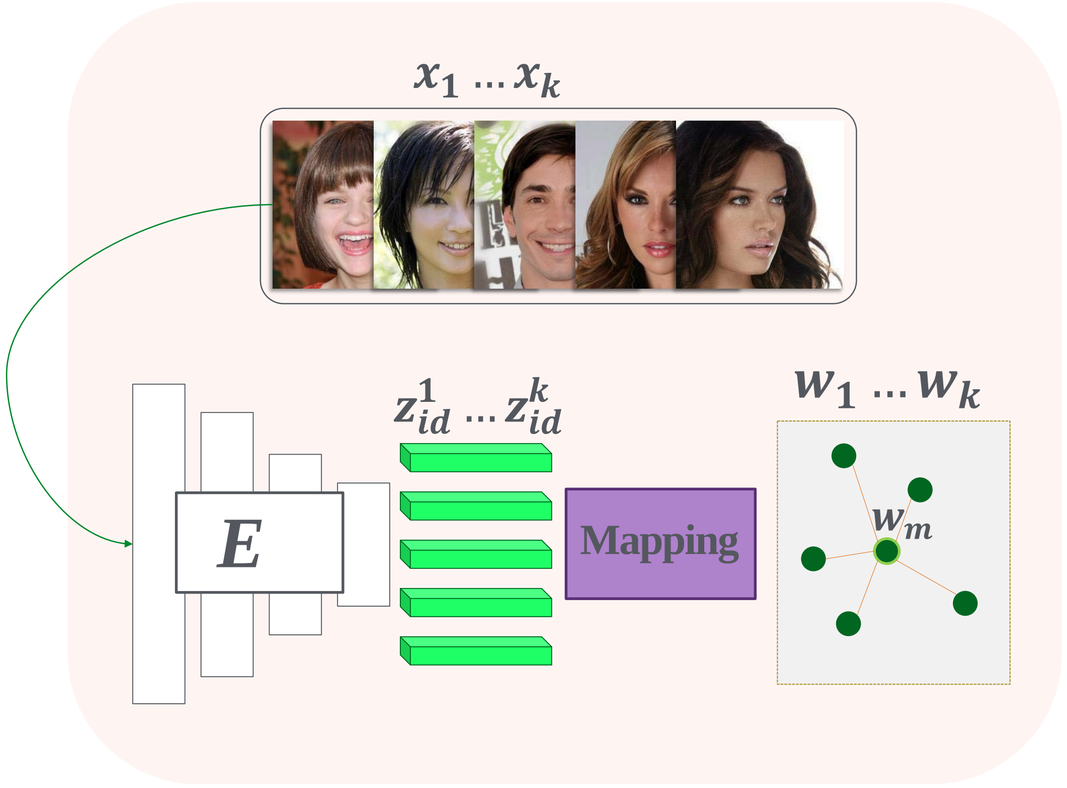
$Mixing(\hat{x}_{mix})$

※ Paper: FICGAN: Facial Identity Controllable GAN for De-identification (https://arxiv.org/abs/2110.00740)

2.3 얼굴 이미지 비식별화

얼굴 이미지 비식별화 – 비식별화(De-id) 수행

- k-익명성 보장을 위한 W의 Centroid 계산



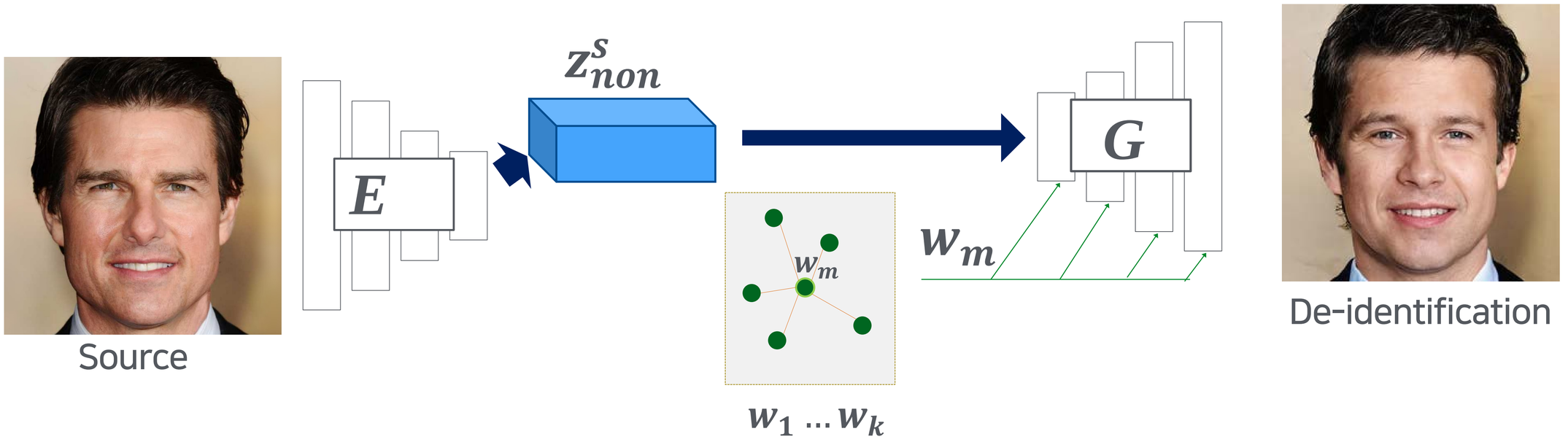
Manifold *k*-same algorithm

※ Paper: FICGAN: Facial Identity Controllable GAN for De-identification (https://arxiv.org/abs/2110.00740)

2.3 얼굴 이미지 비식별화

얼굴 이미지 비식별화 – 비식별화(De-id) 수행

- Manifold k-same algorithm이 적용된 얼굴 이미지 비식별화

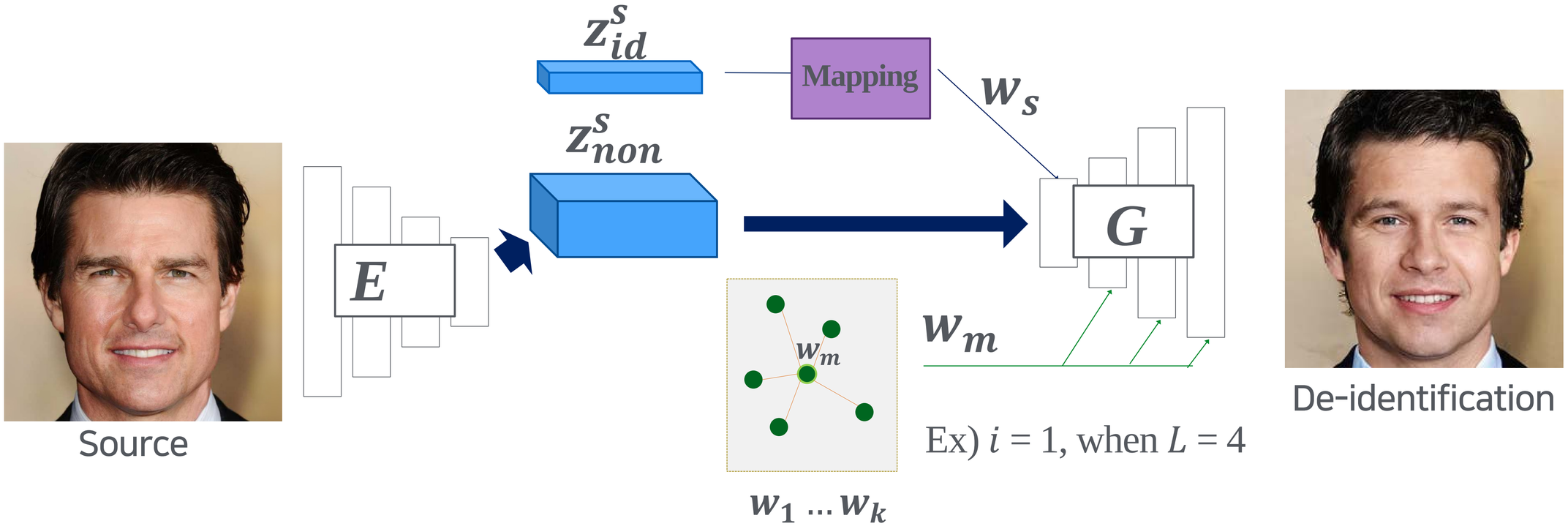


※ Paper: FICGAN: Facial Identity Controllable GAN for De-identification (https://arxiv.org/abs/2110.00740)

2.3 얼굴 이미지 비식별화

얼굴 이미지 비식별화 – 비식별화(De-id) 수행

- 익명화의 수준을 조절.

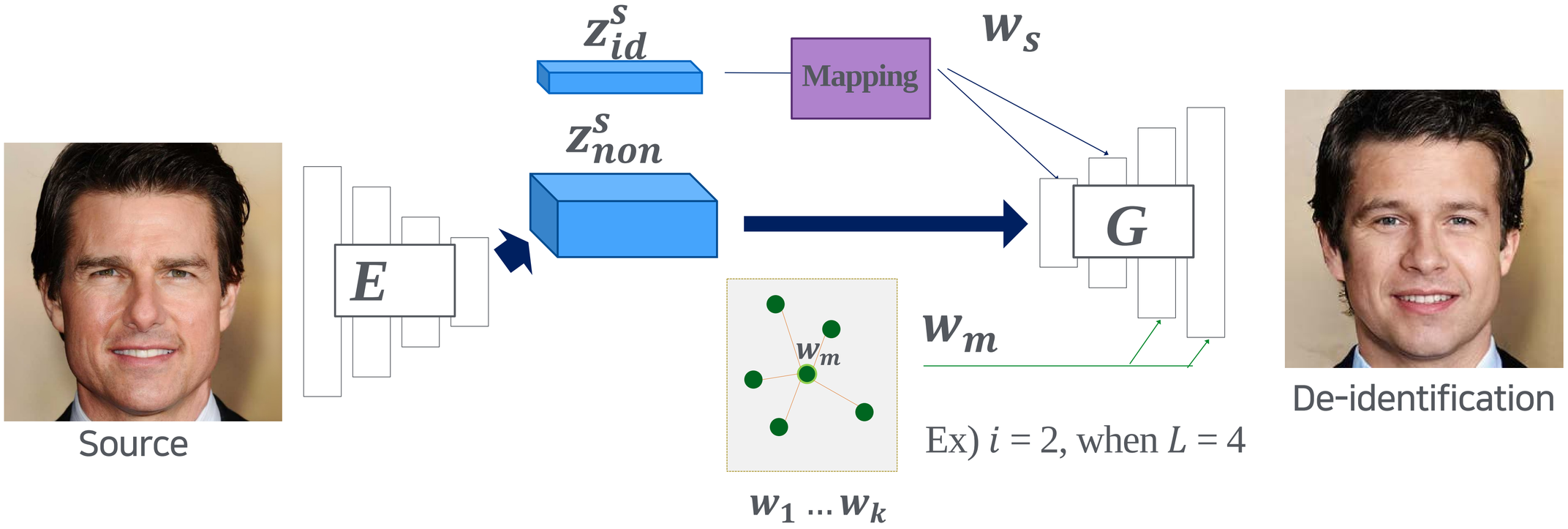


※ Paper: FICGAN: Facial Identity Controllable GAN for De-identification (https://arxiv.org/abs/2110.00740)

2.3 얼굴 이미지 비식별화

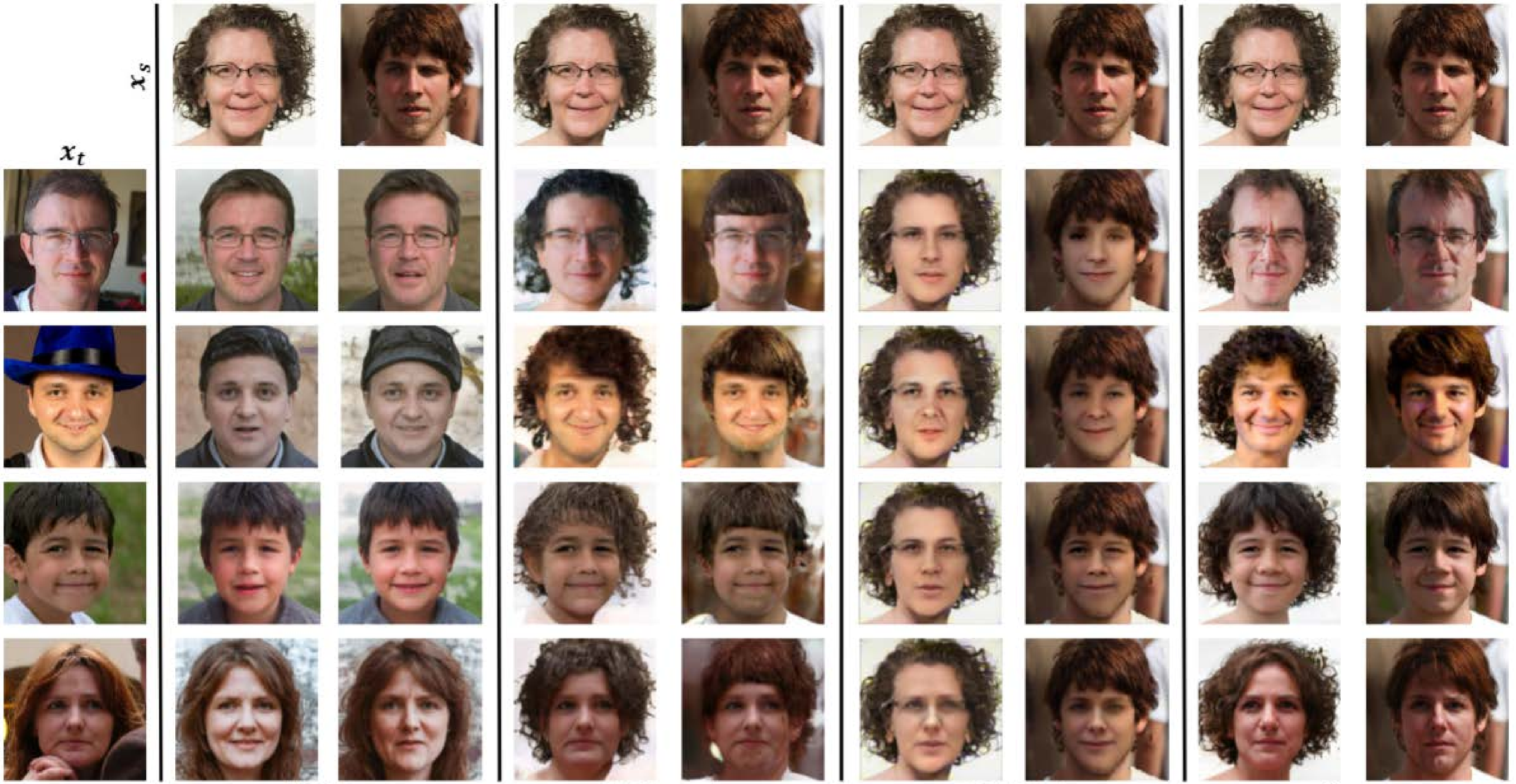
얼굴 이미지 비식별화 – 비식별화(De-id) 수행

- 익명화의 수준을 조절.



※ Paper: FICGAN: Facial Identity Controllable GAN for De-identification (https://arxiv.org/abs/2110.00740)

2.3 얼굴 이미지 비식별화



※ Paper: FICGAN

2.3 얼굴 이미지 비식별화

DEVIEW 2021

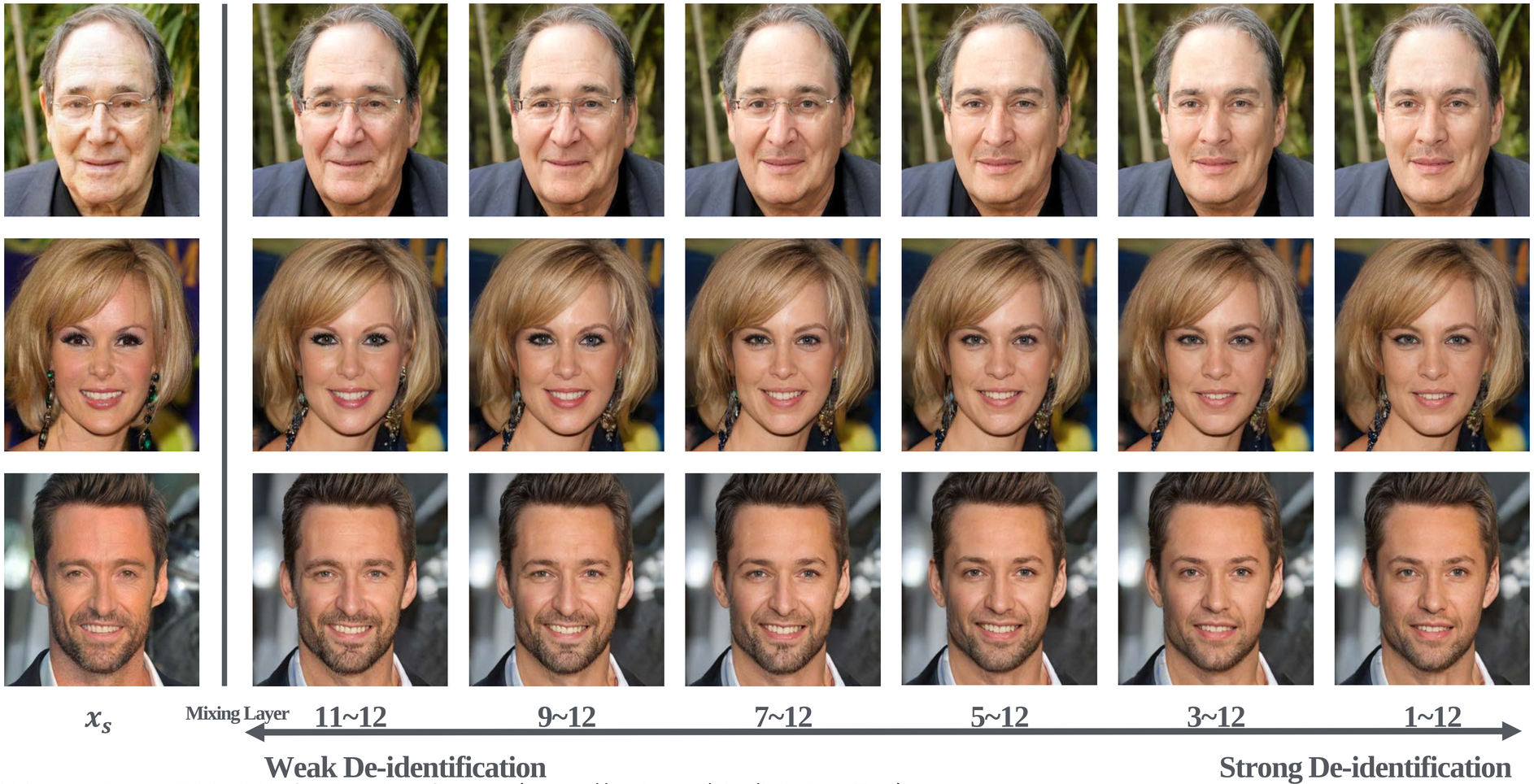
얼굴 이미지 비식별화 - 결과

Source IDDis FSGAN IDInvert Ours

※ Paper: FICGAN: Facial Identity Controllable GAN for De-identification (https://arxiv.org/abs/2110.00740)

2.3 얼굴 이미지 비식별화

얼굴 이미지 비식별화 - 결과



※ Paper: FICGAN: Facial Identity Controllable GAN for De-identification (https://arxiv.org/abs/2110.00740)

3.이미지 딥페이크의 악용

3.1 딥페이크 조작 사례

정치적 악용

- 일본 반정부 여론 조성에 딥페이크 악용 (‘21.2)

개요

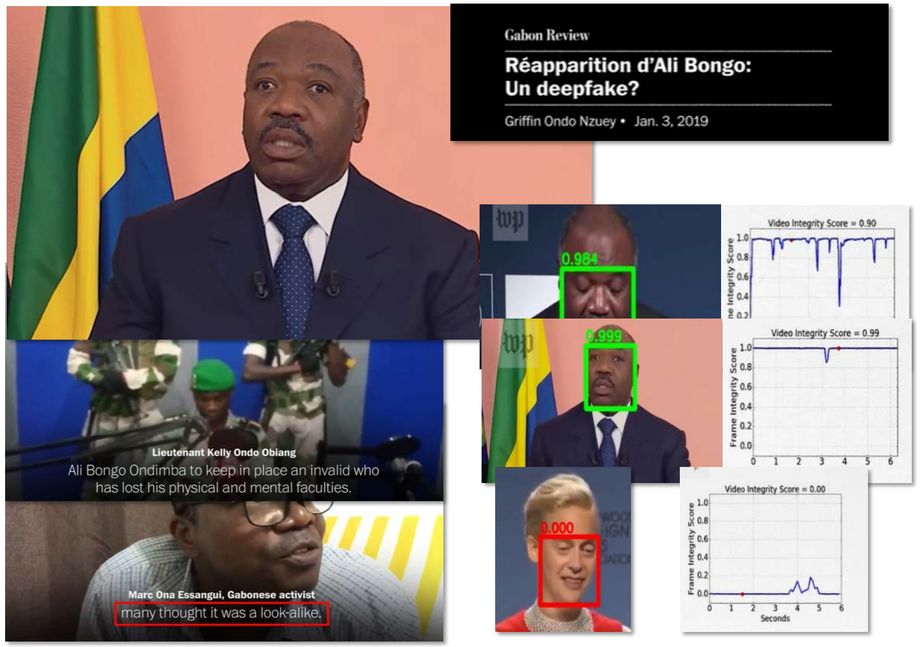
- **후쿠시마 지진 직후 트위터에 웃음 띤 장관 얼굴 유통**
  - 가토 관방 장관 인터뷰 직후 ‘악랄함에도 정도가 있다’는 제목의 웃는 딥페이크 사진을 조작, 유포하여 **반정부 여론 조성**
  - 일부 사진은 회견 30분 내 조작, 트위터에 유포

출처: 요미우리 신문 加藤官房長官、フェイクの笑み…AI で悪意の改変［虚実のはざま］第２部 作られる「真相」(4/28)
https://www.yomiuri.co.jp/pluralphoto/20210425-OYT1I50063

3.1 딥페이크 조작 사례

정치적 악용

- 딥페이크 오해가 낳은 쿠테타('19.1)



개요

- 아프리카 가봉 군부, 자국 대통령의 연설 영상을 딥페이크로 주장하며 쿠데타 촉발
- 알리 봉고 대통령의 뇌졸중 은폐가 불러온 역효과

출처: 워싱턴포스트 https://wapo.st/3w7Bxgb

3.1 딥페이크 조작 사례

얼굴 '바꿔치기'를 통한 경제적 악용

- 중국 정부의 얼굴인식 시스템을 딥페이크로 속여 875억 세금 탈루 ('21.3)

A group of tax scammers hacked a government-run identity verification system to fake tax invoices. The fake tax invoices from the criminal group were valued at US$76.2mil. — SCMP

개요

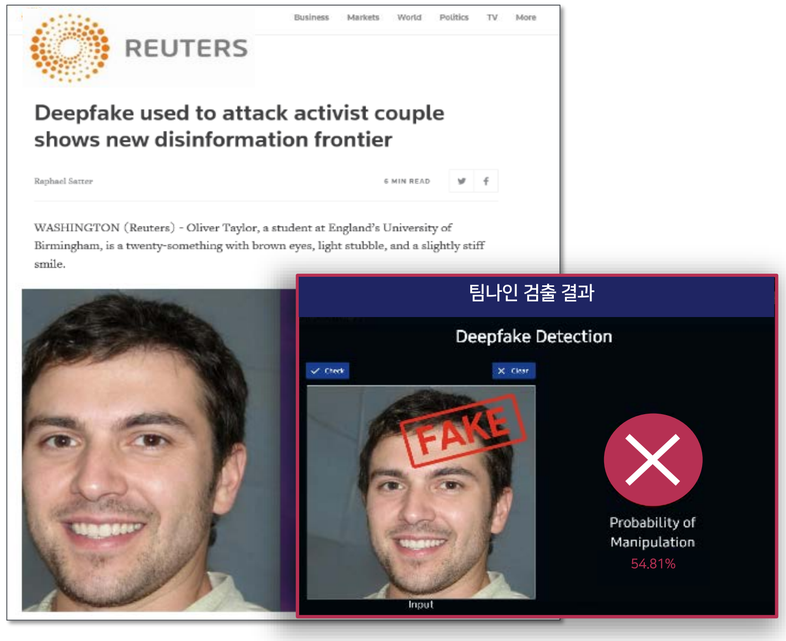
- 중국 정부 운영 안면 인식 시스템을 딥페이크로 속여 875억 세금 탈루 건 적발 ('18~'21)
  - 상하이 검찰, 암시장에서 구한 개인정보와 딥페이크 결합해 세금계산서 허위 발행 범죄자 검거
  - 안면 인증 과정에서 스마트폰 카메라 신호를 탈취, 미리 준비한 딥페이크 사진 및 영상을 수신하도록 조작해서 인증 통과

출처: 사우스차이나 모닝포스트 ('21.3) China's deepfake celebrity porn culture stirs debate about artificial intelligence use | South China Morning Post (scmp.com)

3.1 딥페이크 조작 사례

언론 왜곡

- 로이터, 딥페이크로 만든 가상 기고자 통한 미디어 왜곡 시도 적발 (‘20.7)



개요

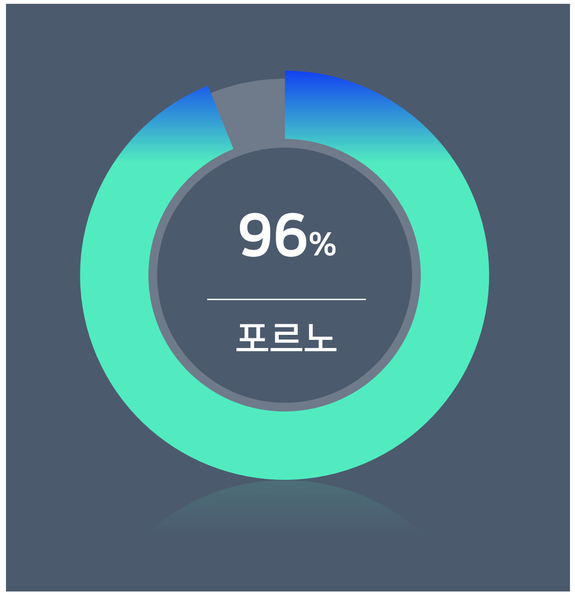
- 로이터는 최근 Jerusalem Post, Israel Times 등 미디어에서 팔레스타인 인권 운동가를 비난한 프리랜서 Oliver Taylor가 실재하지 않는 가상 인물임을 확인 (‘20.7)

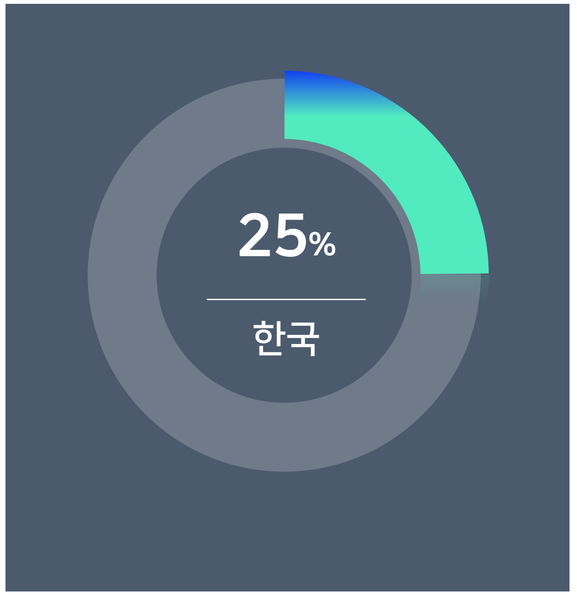
출처: 로이터 뉴스 https://reut.rs/399Q5kL

3.2 딥페이크로 인한 피해와 대응

딥페이크 피해 대상의 1/4이 여성 K-Pop 스타

- 네덜란드 딥페이크 탐지업체 Deeptrace, 딥페이크 영상의 96%가 음란물
- 미국 Rolling Stone Magazine, 딥페이크 피해 대상의 ¼가 여성 K-Pop 스타





출처: DeepTrace, RollingStone https://www.rollingstone.com/culture/culture-news/deepfakes-nonconsensual-porn-study-kpop-895605/

3.2 딥페이크로 인한 피해와 대응

K-pop과 N번방 사건으로 인한 불법 음란물 피해

- 딥페이크 조작 영상 대량 발견

※Image : https://www.imbc.com (MBC NEWS)

3.2 딥페이크로 인한 피해와 대응

딥페이크 처벌을 위한 국민 청원

- 2021년 1월 13일 ~ 2월 12일 한달 간 39만명 청원

여성 연예인들을 고통받게 하는 불법 영상 '딥 페이크' 를 강력히 처벌해주세요.

참여인원 : [390,415명]

카테고리	청원시작	청원마감	청원인
인권/성평등	2021-01-13	2021-02-12	kakao - ***

딥페이크 기술 등을 악용해 불법합성물을 제작해 반포하는 행위는 명백한 범죄행위입니다. 지난해 6월 성폭력처벌법이 개정되면서 관련 규정이 신설된 후 처벌이 가능해졌으며, 경찰은 딥페이크 기술을 악용한 불법합성물 근절을 위해 지난해 12월부터 「허위영상물 제작·유포사범 집중단속」을 실시 중입니다. (청와대 답변 원고 내용 중)

출처: 청와대 국민청원 https://www1.president.go.kr/petitions/595595

3.2 딥페이크로 인한 피해와 대응

DEVIEW 2021

허위영상물 제작·유포사범 집중단속 및 처벌

- 전년 대비 처벌 건수 2.5배 증가

<표-1> 딥페이크 성적 허위영상 처리 현황

(단위 : 건)

구분	심의(차단)	자율규제(삭제)	계
2020년 6월~12월	473	75	548
2021년 1월~9월	537	871	1,408

※ 방송통신심의위원회 제출자료(기간 : 2020.6.25.~2021.9.24.)
성폭력범죄처벌특례법 제14조의2(허위영상물 등의 반포 등)는 2020. 6. 25.부터 시행

방송통신심의위원회에 따르면, 올해 1~9월 딥페이크 처리 건수는 1,408건으로 지난해 6~12월(548건)에 비해 256% 증가했다.

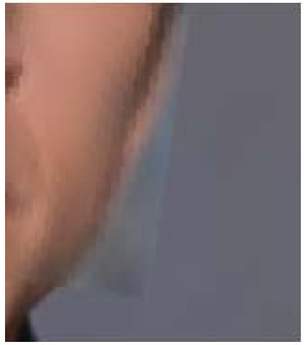
출처: 뉴스워커 http://www.newsworker.co.kr/news/articleView.html?idxno=130611

4.이미지 딥페이크 검출

4.1 딥페이크를 검출하는 방법

DEVIEW 2021

부분적 변경을 수행한 이미지들의 부자연스러운 경계선

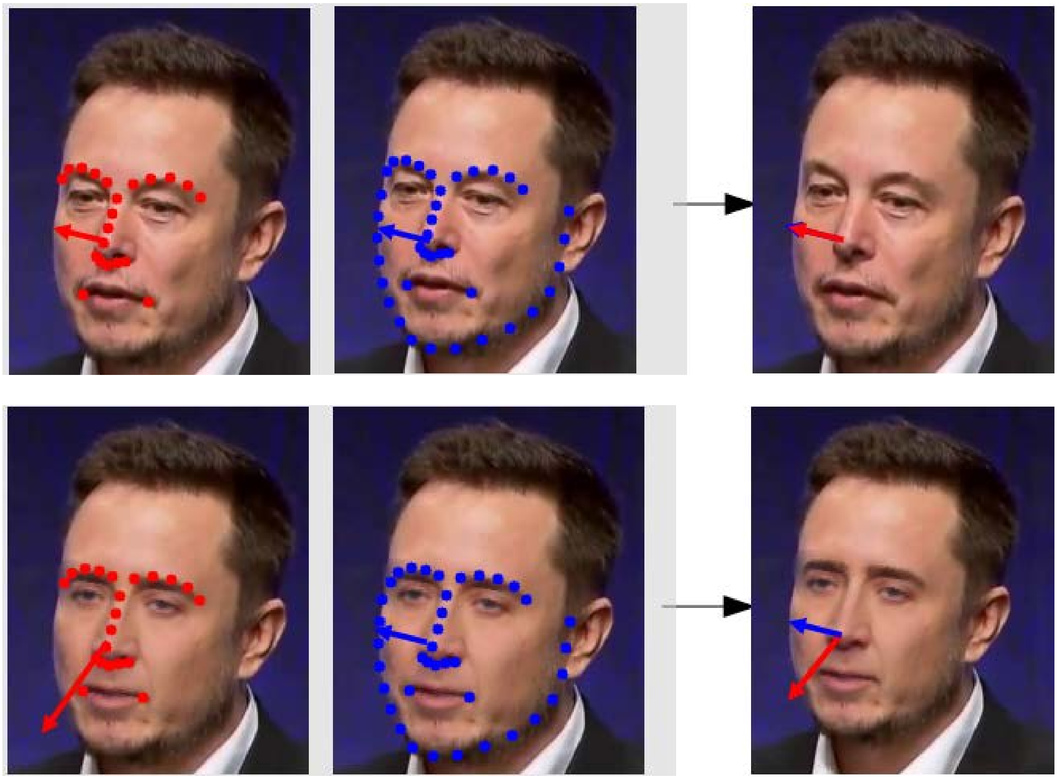


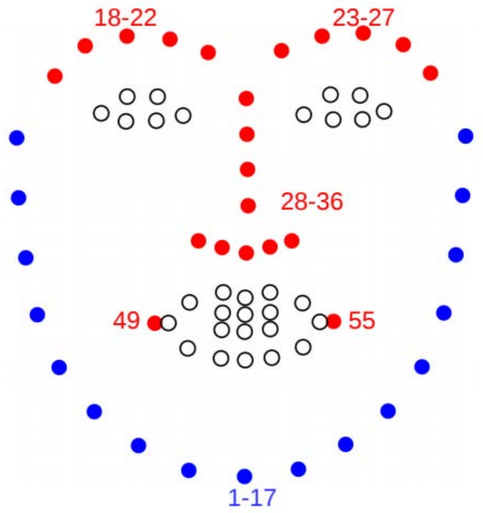
※ Image : Face X-ray for More General Face Forgery Detection (CVPR 2020)

Exposing AI Created Fake Videos by Detecting Eye Blinking (WIFS, 2018)

4.1 딥페이크를 검출하는 방법

Head pose를 통한 검출 방법





※ Image : Exposing Deep Fakes Using Inconsistent Head Poses (ICASSP, 2019)

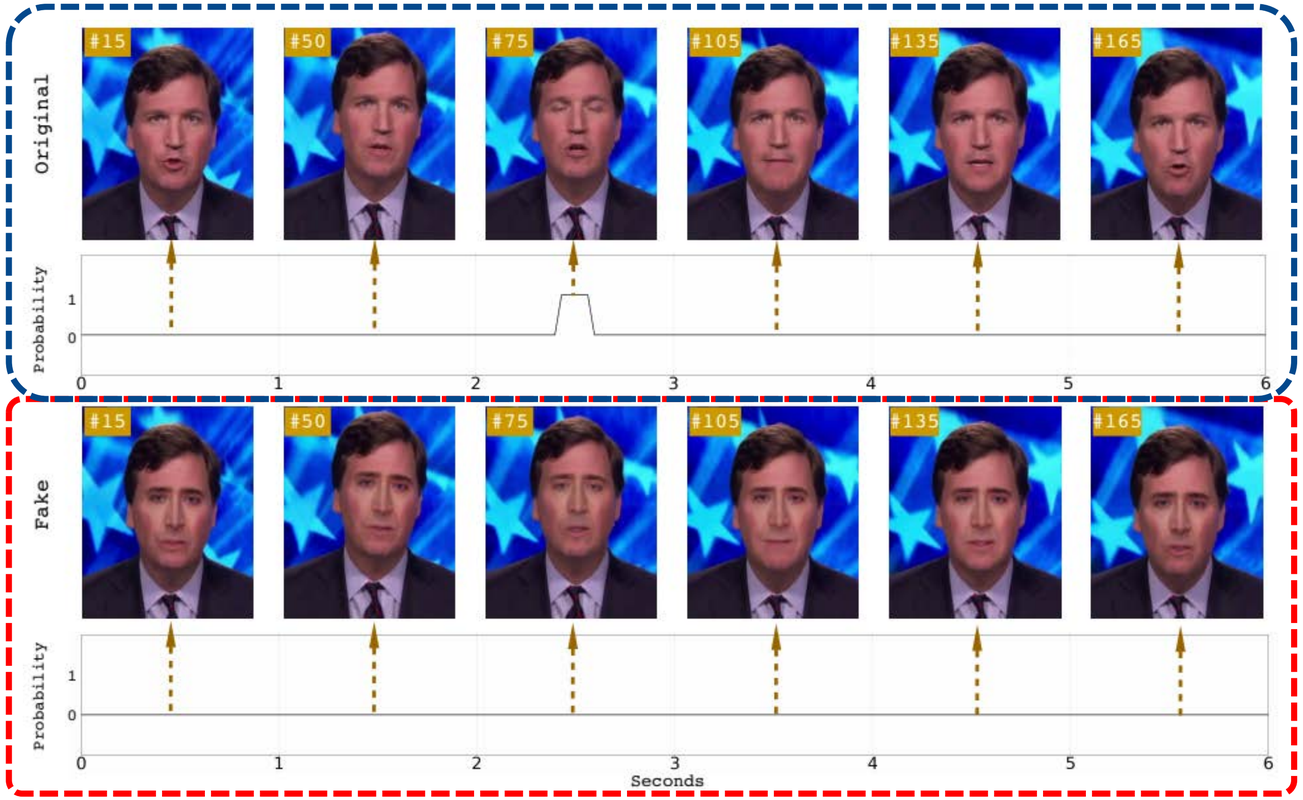
4.1 딥페이크를 검출하는 방법

눈동자의 특징을 활용한 검출 방법

※ Image : Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations (WACVW, 2019)

4.1 딥페이크를 검출하는 방법

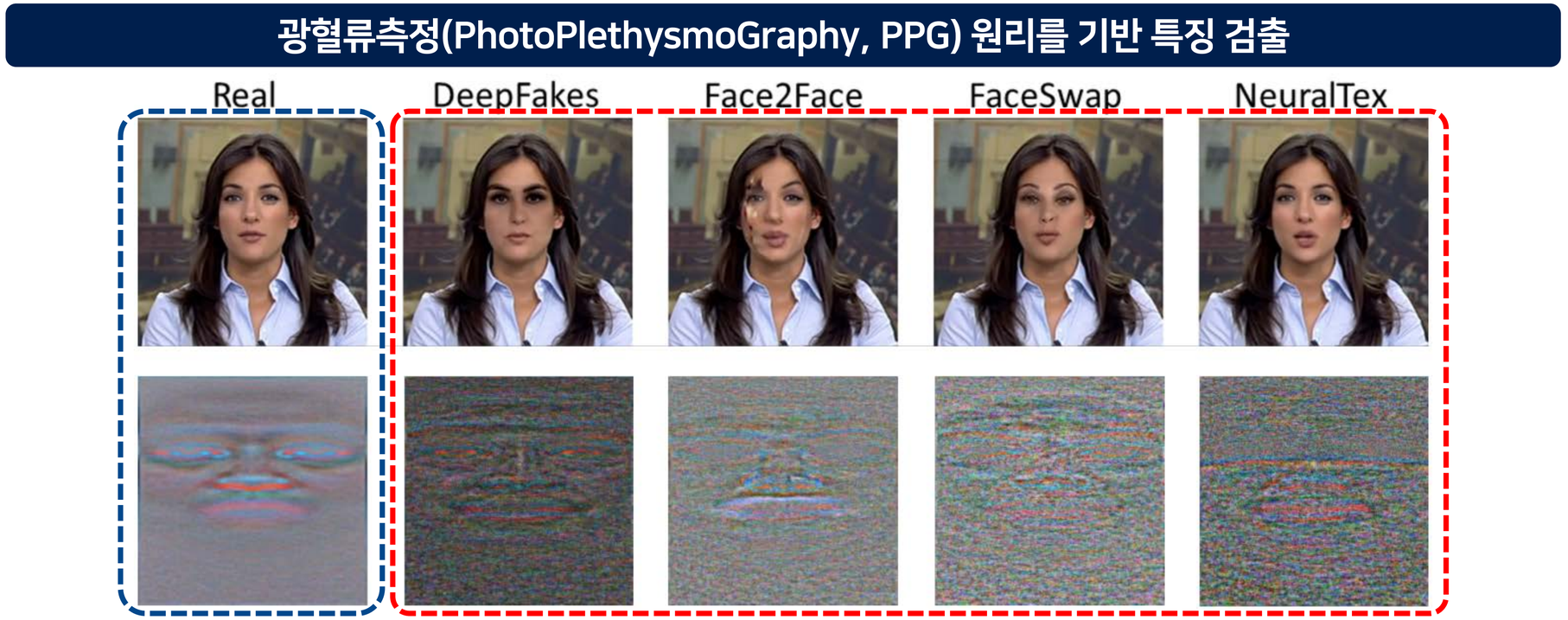
심박수, 눈 깜박임, 그림자 분석 등 생리적 특징을 통해 검출



※ Image : Exposing AI Created Fake Videos by Detecting Eye Blinking (WIFS, 2018)

4.1 딥페이크를 검출하는 방법

이미지 내 사람의 혈류가 어떻게 흐르는지 관찰
광혈류 측정 (PPG) 원리 기반 특징 분석



※ Image : How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals (arXiv 2020)

4.2 생리적 특징 검출의 한계점

최근 얼굴 외 다양한 오브젝트로 생성 범위 확대 중,
그러나 생리적 특징 검출은 얼굴 위주의 검출 범위 한계 존재

StyleGAN2가 생성한 가짜 이미지

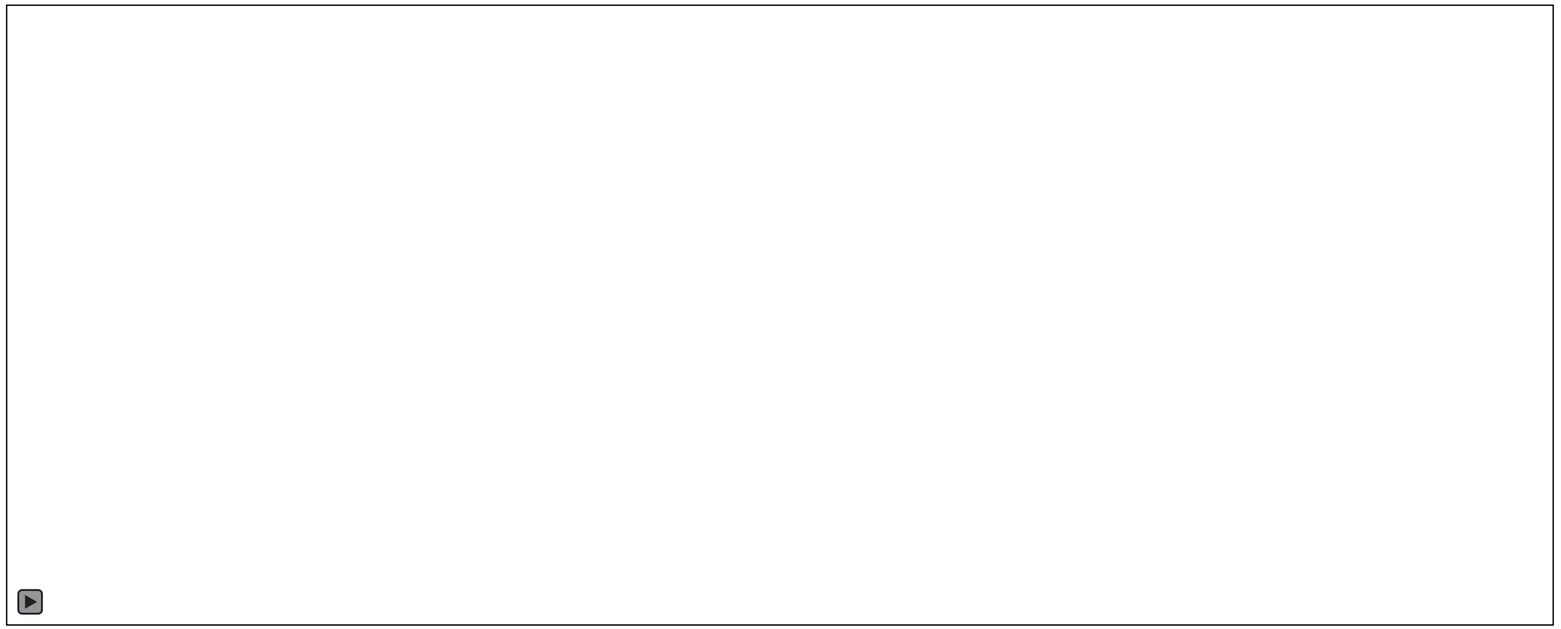
자동차 고양이 교회 말

※ Image: Analyzing and Improving the Image Quality of StyleGAN (CVPR 2020)

4.3 뉴럴네트워크가 만드는 Fingerprint

DEVIEW 2021

Convolutional Neural Network에 의해 생성된 특징

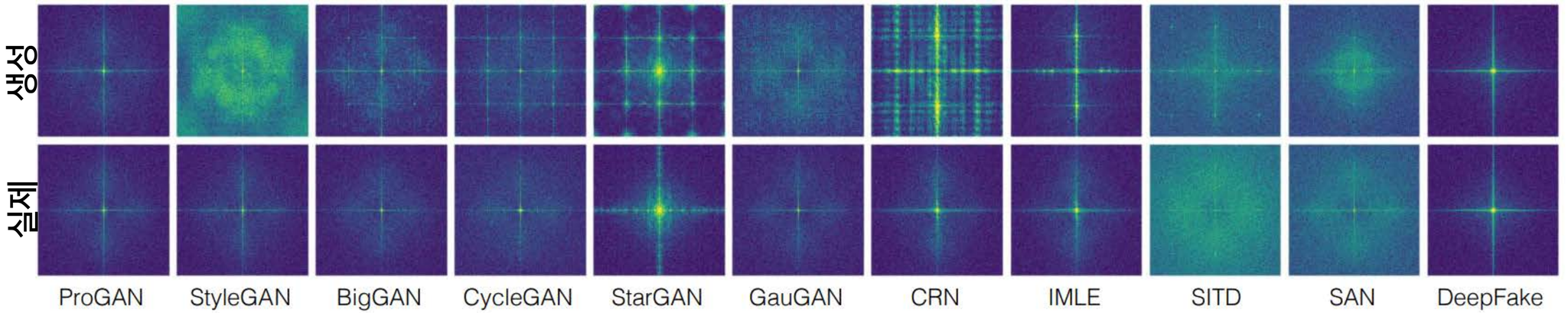


※ Paper: Odena, et al., "Deconvolution and Checkerboard Artifacts", Distill, 2016. http://doi.org/10.23915/distill.00003

4.3 뉴럴네트워크가 만드는 Fingerprint

딥페이크에 의해 만들어진 이미지는 주파수 성분에서 특징 발생

High-pass filter를 통과한 이미지의 평균 스팩트럼



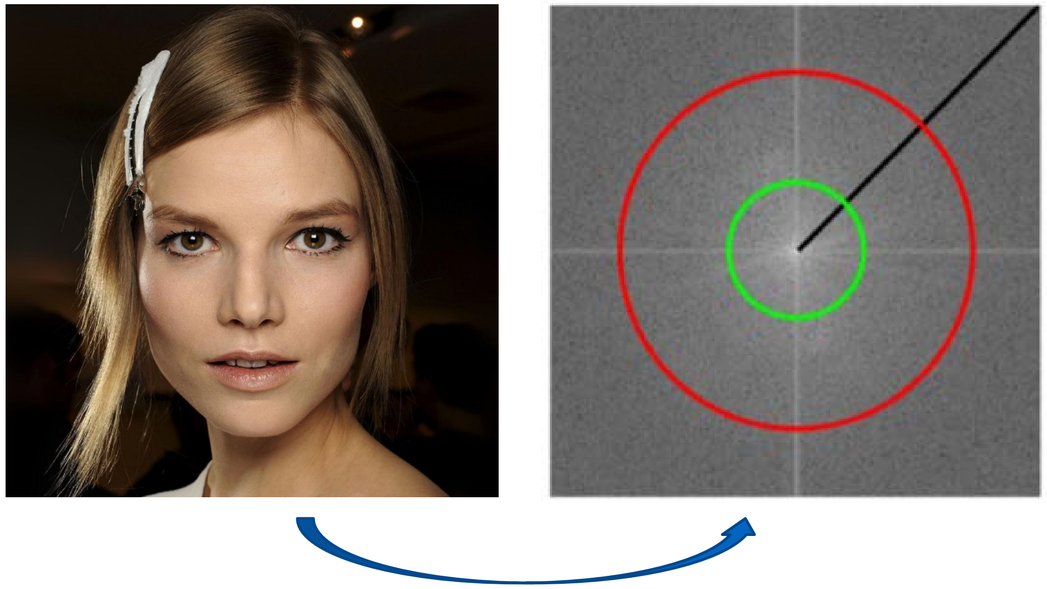
체크 무늬 (격자 무늬) Artifact 존재

※ Image : CNN-generated images are surprisingly easy to spot... for now (CVPR 2020)

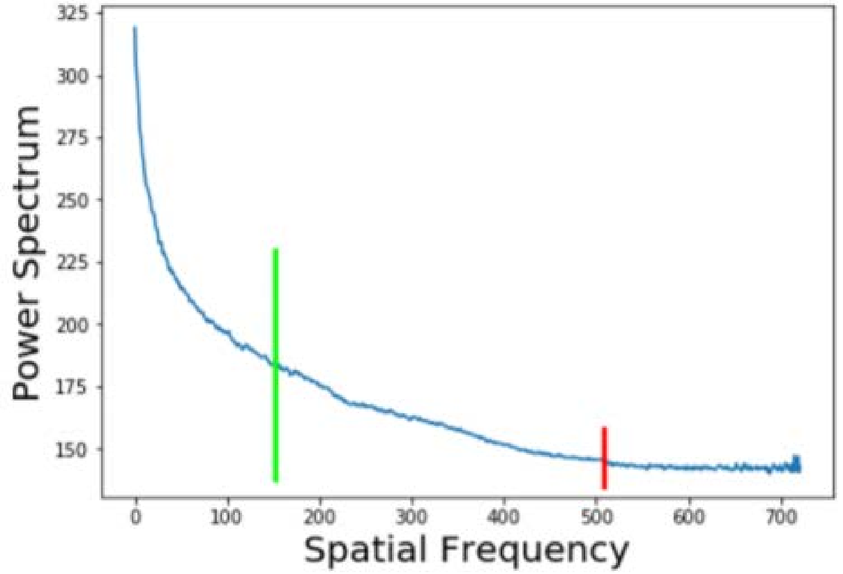
4.4 고주파 성분 기반 검출

주파수 도메인에서 1차원 스펙트럼 분석

2D power spectrum 성분을 1D power spectrum으로 변환



이미지를 주파수 성분으로 변경 (퓨리에 변환)

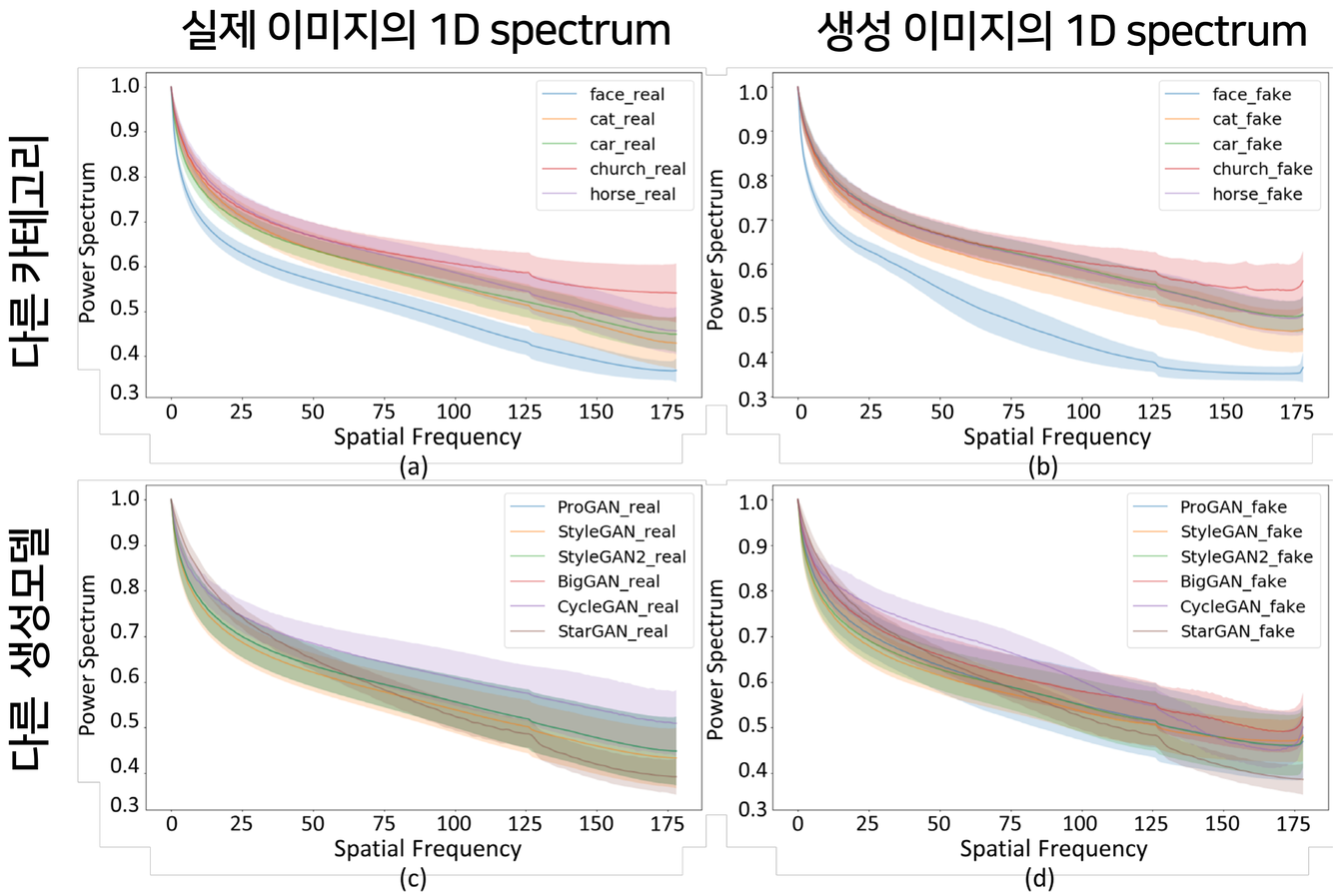


1D- Power spectrum

※ Image : Unmasking DeepFakes with simple Features (arXiv 2019)

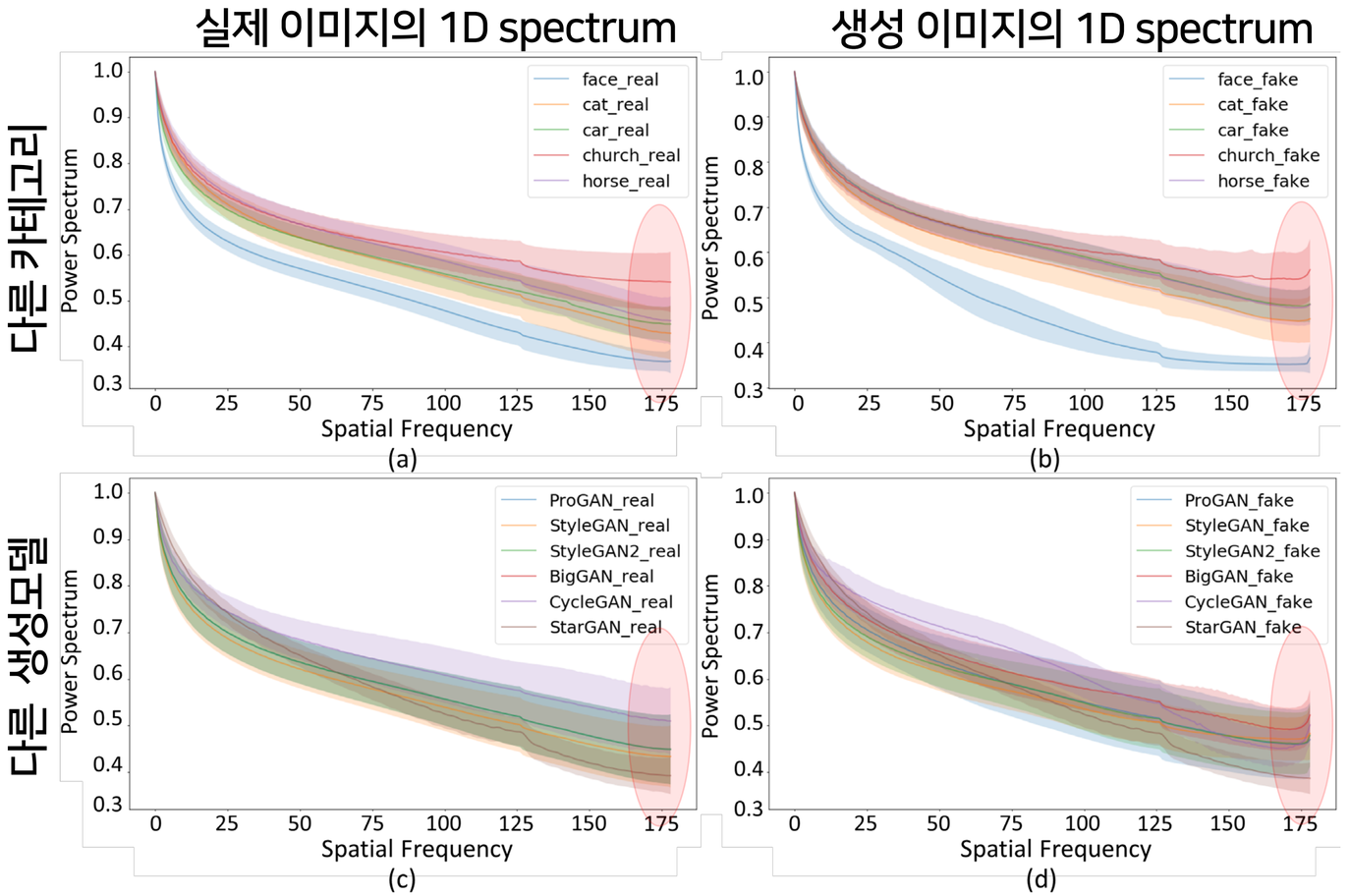
4.4 고주파 성분 기반 검출

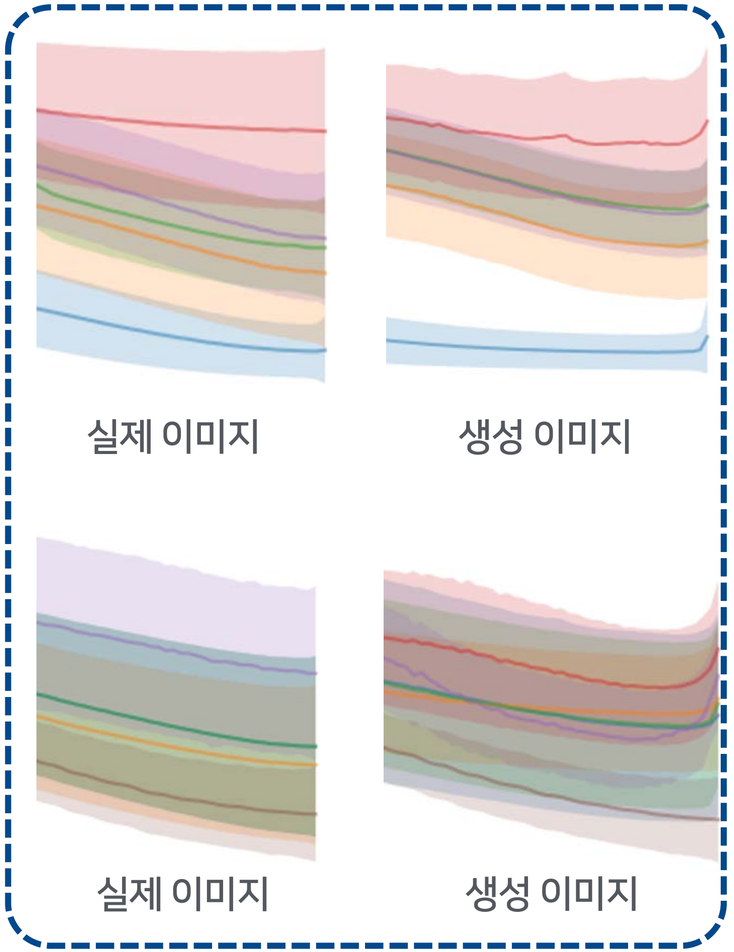
고주파 성분에서 진짜와 가짜 이미지 간 상이한 특징 발견



※ Paper: BiHPF: Bilateral High-Pass Filters for Robust Deepfake Detection (WACV 2022, https://arxiv.org/pdf/2109.00911)

4.4 고주파 성분 기반 검출





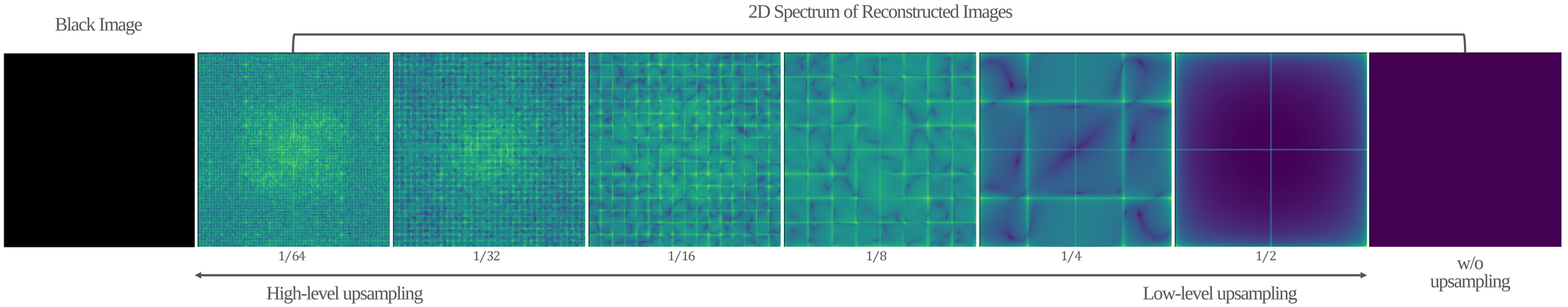
확대 영역

※ Paper: BiHPF: Bilateral High-Pass Filters for Robust Deepfake Detection (WACV 2022, https://arxiv.org/pdf/2109.00911)

4.5 체크 보드 아티팩트 생성

DEVIEW 2021

CNN의 업샘플링 수에 따라 변화하는 격자 무늬 아티팩트

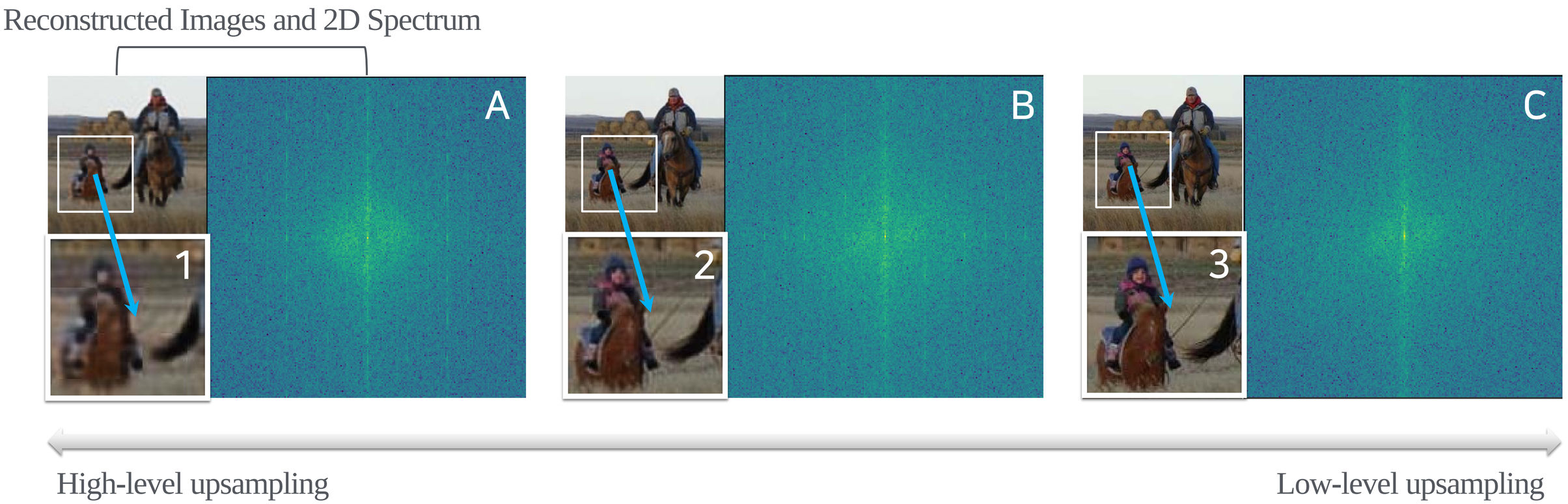


※ Patent: Apparatus for deep fake image discrimination and learning method thereof. (Journal submitted under review)

4.5 체크 보드 아티팩트 생성

DEVIEW 2021

같은 이미지 같지만, CNN의 다운샘플링 깊이에 따라 변화하는 격자 무늬 아티팩트

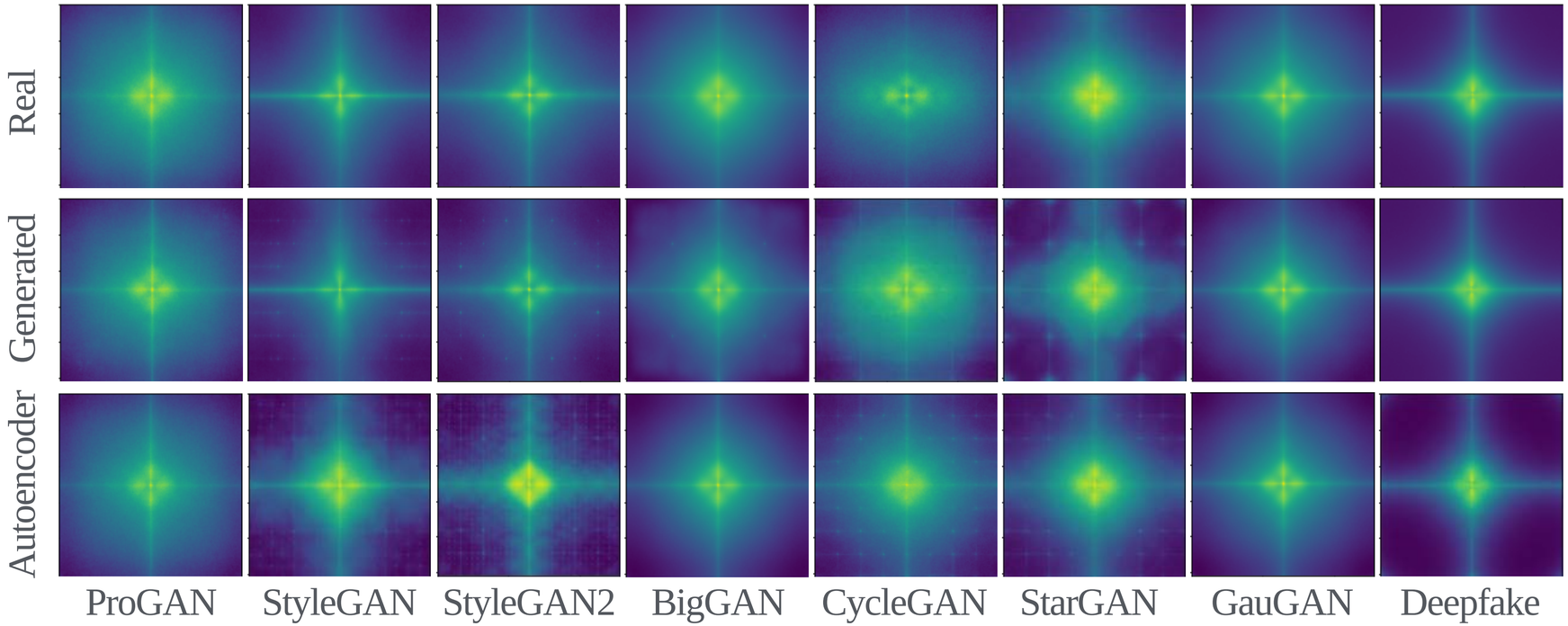


※ Patent: Apparatus for deep fake image discrimination and learning method thereof. (Journal submitted under review)

4.5 체크 보드 아티팩트 생성

DEVIEW 2021

다양한 업샘플링을 통해 여러 개의 GAN과 비슷한 주파수 분포를 찾을 수 있다.



※ Patent: Apparatus for deep fake image discrimination and learning method thereof. (Journal submitted under review)

4.5 체크 보드 아티팩트 생성

기존의 한계점을 개선하여, 타 모델 대비 가장 범용적이고 높은 검출 성능 확보

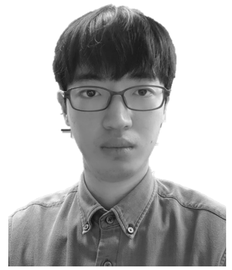
Model	StyleGAN		StyleGAN2		BigGAN		CycleGAN		StarGAN		GauGAN		Deepfake		Mean	
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
Wang (CVPR 2020)	63.8	91.4	76.4	97.5	52.9	73.3	72.7	88.6	63.8	90.8	63.9	92.2	51.7	62.3	63.6	85.2
Frank (ICML 2020)	72.2	82.1	64.2	80.1	68.9	82.4	53.7	66.2	89.1	99.2	65.3	90.3	51.1	49.6	66.4	78.6
Durall (CVPR 2020)	63.9	58.4	69.0	62.7	58.5	54.7	63.6	63.1	99.0	98.1	57.0	53.8	50.4	50.1	66.8	63.0
Ours	71.5	79.1	70.0	79.3	77.3	90.0	57.5	86.3	99.8	100.	70.9	96.9	69.2	74.0	73.7	86.5

※ Patent: Apparatus for deep fake image discrimination and learning method thereof. (Journal submitted under review)

함께 연구하신 분들

포항공과대학교
AI대학원
오태현 교수

서울대학교
DSAIL 연구실
최주영 연구원



삼성SDS
보안사업부
김평건 프로

중앙대학교
AI대학원
최종원 교수

서울대학교
DSAIL 연구실
김성원 연구원

삼성SDS
보안사업부
김도연 프로

삼성SDS
AI 연구센터
노영민 프로

서울대학교
DSAIL 연구실
하한석 연구원

삼성SDS
AI 연구센터
정용현 프로

Q & A

Thank You