AI 기반 딥페이크 탐지 기술

26th NetSec 2020
July 17, 2020

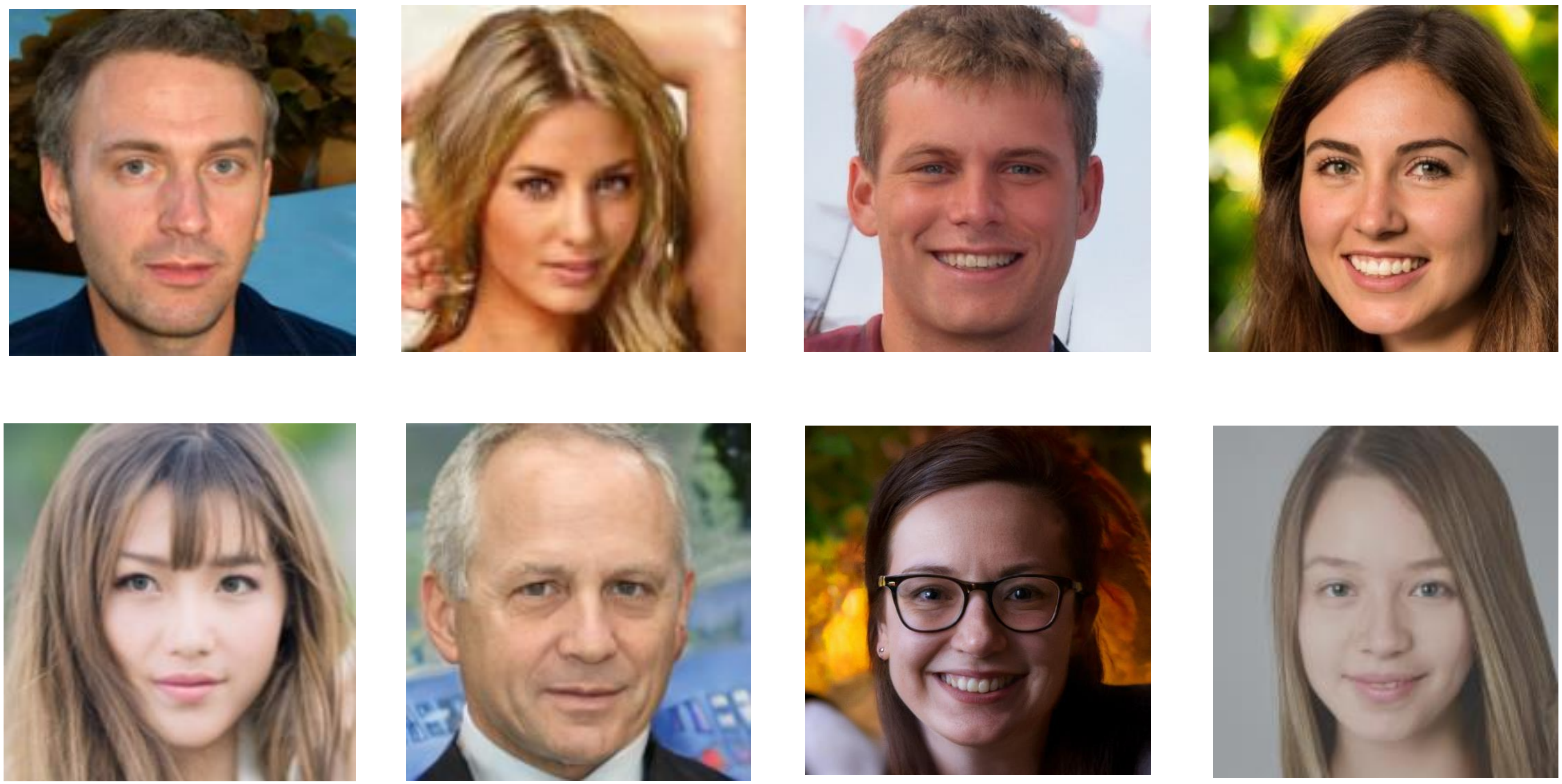
우사이먼성일
데이터사이언스융합학과
소프트웨어학과
인공지능 융합학과
성균관대학교

Do you know them?

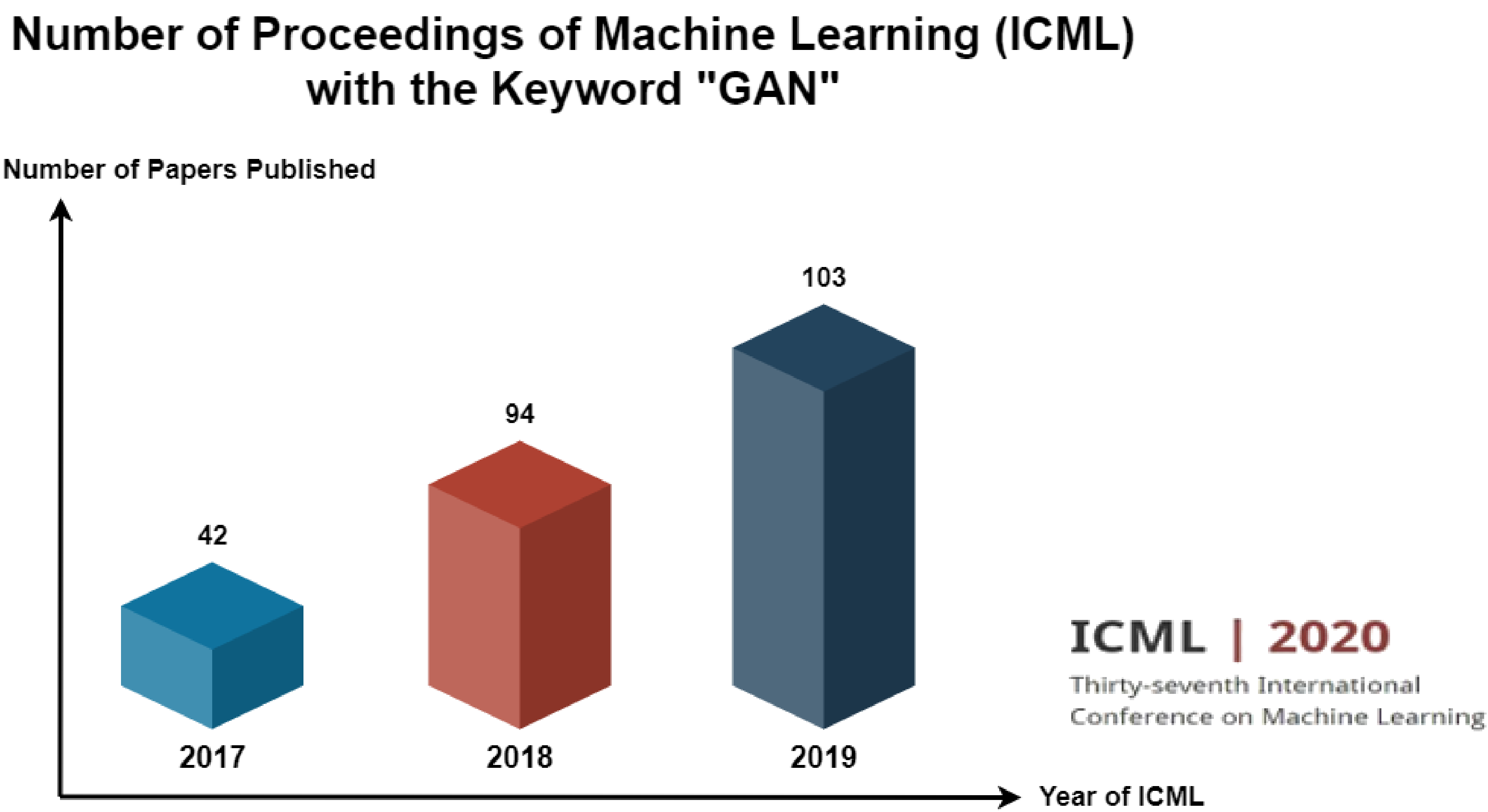
Progressive Growing of GANs for Improved Quality, Stability, and Variation (PGAN) by Nvidia Team

Image source: https://research.nvidia.com/publication/2017-10_Progressive-Growing-of

GAN generated images: PGGAN,StarGAN,StyleGAN,StyleGAN2



GAN architecture continues to be developed and published.



Why is this a problem?

실태조사

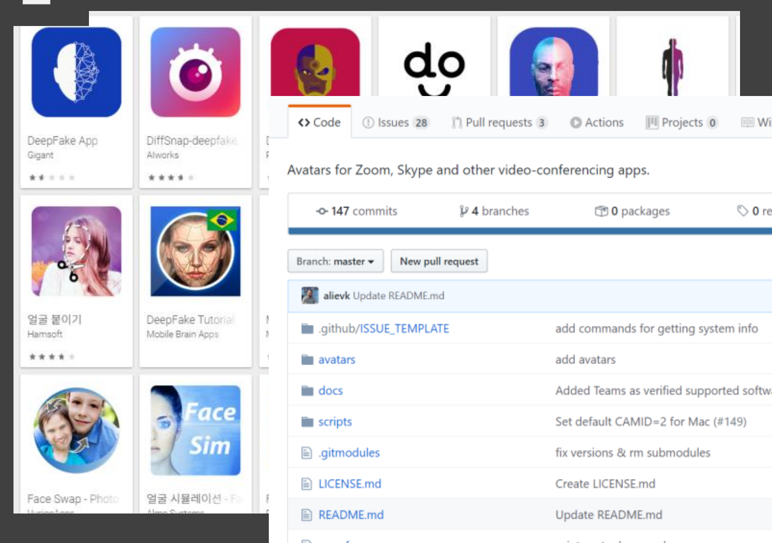
최근 사례 및 현황

1	코로나-19 관련 벨기에 총리의 딥페이크 영상
2	누구나 쉽게 사용할 수 있는 어플리케이션 및 프로그램 코드
3	음란물사이트의 딥페이크 영상 - 국가별 분류

1

실제 공개된 영상 캡쳐화면

2



3

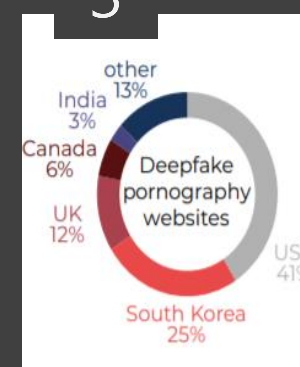


Image source: http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=102&oid=008&aid=0003748739 (Money Today)
Image source: http://post.naver.com/viewer/postView.nhn?volumeNo=13993925&memberNo=225195 (News 1)

심각한 문제

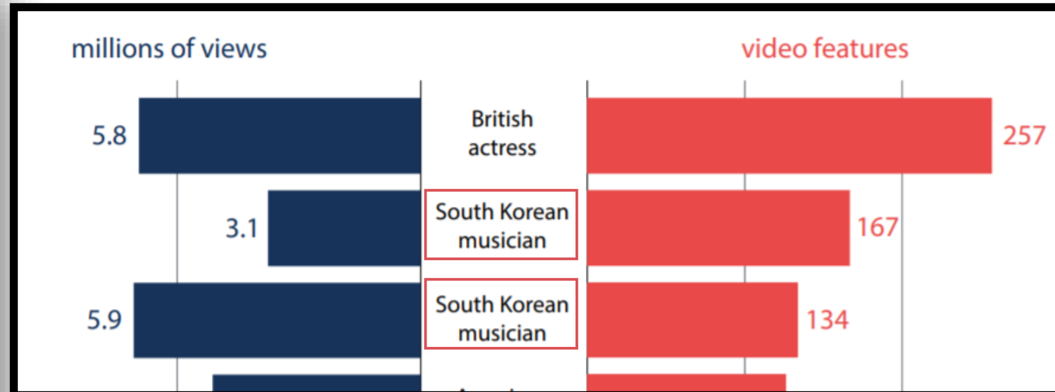
01

국내인 대상 딥페이크 현황 조사

피해 대상별 불법행위 분류

- 유명인 대상
- 일반 성인대상
- 아동/청소년 대상

딥페이크 불법 음란물 피해사례(유명인)



출처: THE STATE OF DEEPFAKES, Deeptrace 2019

딥페이크 불법 음란물 피해사례(일반인)

n번방을 잇는 '지인능욕' 가해자들을 조사 해주세요.

참여인원 : [54,203명]

딥페이크 불법 음란물 피해사례 (아동/청소년)→

"Deepfakes don't hurt people, People using deepfakes hurt people."

사람을 해치는 것은 딥페이크가 아니라 딥페이크를 "악용" 하는 사람이다

Current Talk

- Intro to Deepfake Generation and Detection Methods
- Towards the Universal Detection
  - Few-shots/Unbalanced Dataset
  - One-Class detection
  - Transfer Learning
- Government/Industry Efforts
- Conclusions

Relevant Research Publications on DeepFakes Detection (2018-19)

[1] Shahroz Tariq, Sangyup Lee, Youjin Shin, Ho Young Kim, and **Simon S. Woo* "Detecting Both Machine and Human Created Fake Face Images In the Wild",** 2nd International Workshop on Multimedia Privacy and Security (MPS 2018), co-located with 25th ACM Conference on Computer and Communications Security (CCS 2018), Toronto, USA, 2018

[2] Shahroz Tariq, Sangyup Lee, Youjin Shin, Ho Young Kim, and **Simon S. Woo*, "GAN is a Friend or Foe? A Framework to Detect Various Fake Face Images", ACM SAC** Cyprus April 2019,

(**BK Computer Science 우수학회 IF=1**)

[3] Hyeonseong Jeon, Youngoh Bang, and **Simon S. Woo*, "FakeTalkerDetect: Effective and Practical Realistic Neural Talking Head Detection with a Highly Unbalanced Datase",** 10th International Workshop on Human Behavior Understanding (HBU), held in conjunction with ICCV'19 Nov, 2019 - Seoul, S. Korea

[4] Junyaup Kim, Siho Han, and Simon S. Woo*, **"Poster: Classifying Genuine Face images from Disguised Face Images,"** 2019 IEEE International conference on Big Data (IEEE BigData 2019), Los Angeles, CA, USA

Relevant Research Publications on DeepFakes Detection (2020)

[5] Hyeonseong Jeon, Youngoh Bang, and **Simon S. Woo*, "FDFtNet: Facing Off Fake Images using Fake Detection Fine-tuning Network",** SEC 2020 International Conference on Information Security and Privacy Protection (**IFIP-SEC**), Solvenia, Sept 2020 (**BK Computer Science IF=1**)

[6] Hasam Khalid and **Simon S. Woo*, "OC-FakeDect: Classifying Deepfakes Using One-class Variational Autoencoder"**, Workshop on **Media Forensics**, CVPR 2020, Monday, 15th June 2020, Seattle, USA

[7] Hyeonseong Jeon, Youngoh Bang, Junyaup Kim, and **Simon S. Woo*, "T-GD: Transferable GAN-generated Images Detection Framework**." Thirty-seventh International Conference on Machine Learning (**ICML**), Vienna, Austria, 2020 (**BK Computer Science IF=4**)

[8] 전소원, 강준형, 황진희, 우사이먼성일, "국내 딥페이크 기술 현황 및 제도적 대응방안 연구", 한국정보보호학회 하계학술대회 (CISC-S), 2020

Real or Fake? (not the focus of this talk)

Created by our team using Photo editing tools

Image source: License free image from Google, photoshopped by our team

Introduction to Deepfake Generation Methods

가짜 딥페이크 영상 생성 기본원리

< 기본 원리 >

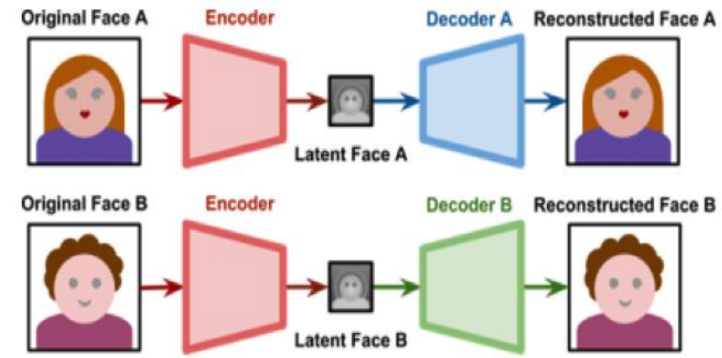
●디코더 A는 A의 얼굴로만 학습하고, 디코더 B는 B의 얼굴로만 학습

●모든 Latent Face는 같은 인코더를 통해 생성됨

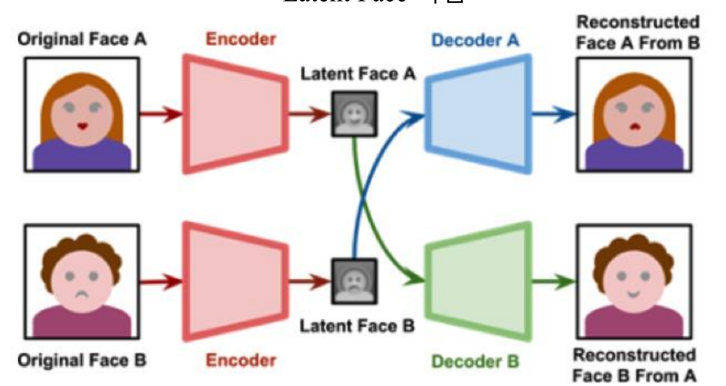
●이것으로 양쪽 얼굴의 공통적인 특징을 정의

●학습 과정이 완료 후, A에서 생성된 Latent Face를 디코더 B에 전달

●디코더 B는 A의 얼굴 움직임으로 B의 이미지를 재구성



A와 B 얼굴을 사용하여
Latent Face 학습



디코더를 변환하여 얼굴 재구성

가짜 영상 데이터셋 구축

- **가짜 영상 생성 methods**
  - DeepFakes – (https://github.com/deepfakes/faceswap)
  - Face2Face – (https://github.com/ondyari/FaceForensics)
  - FaceSwap – (https://github.com/ondyari/FaceForensics/tree/master/dataset/FaceSwapKowalski)
  - Neural Textures – (https://github.com/ondyari/FaceForensics)
  - DeepfakeDetection – (https://github.com/ondyari/FaceForensics)
- **가짜 영상 총 7,000개 dataset: FaceForensics++ (TUM - Visual Computing Group[1])**
  - DeepFake (동영상 1,000개)
  - Face2Face (동영상 1,000개)
  - FaceSwap (동영상 1,000개)
  - Neural Textures (동영상 1,000개)
  - DeepfakeDetection (동영상 3,000개 made by Google)
  - Real (source) – 가짜 영상 생성에 사용된 진짜 동영상

<실시간 Face2Face 예시
배우 얼굴의 facial expression을 푸틴의 얼굴에 입힘>

츄http://www.niessnerlab.org/projects/roessler2019faceforensicspp.html

가짜 영상 생성 예시

000 org

003 org

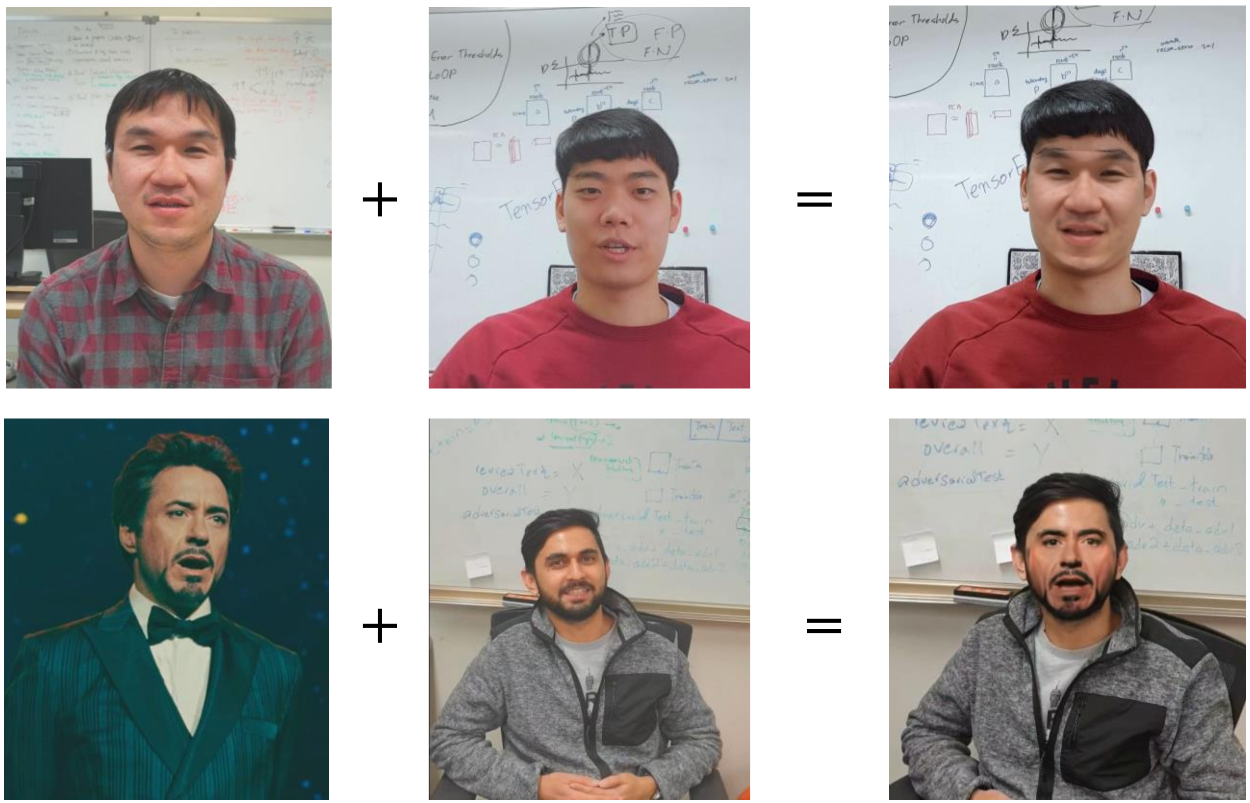
< 원본 영상 >

Deepfake 000 - 003

Face2Face 000 - 003

FaceSwap 000 - 003

가짜 영상 생성 예시



< Deepfakes 생성 예시 >

Deepfake Detection Models

(딥페이크 탐지기법)

[1] Shahroz Tariq, Sangyup Lee, Youjin Shin, Ho Young Kim, and **Simon S. Woo* "Detecting Both Machine and Human Created Fake Face Images In the Wild"**, 2nd International Workshop on Multimedia Privacy and Security (MPS 2018), co-located with 25th ACM Conference on Computer and Communications Security (CCS 2018), Toronto, USA, 2018

[2] Shahroz Tariq, Sangyup Lee, Youjin Shin, Ho Young Kim, and **Simon S. Woo*, "GAN is a Friend or Foe? A Framework to Detect Various Fake Face Images", ACM SAC** Cyprus April 2019,

(**BK Computer Science 우수학회 IF=1**)

Data Collection - GAN

CelebA-HQ

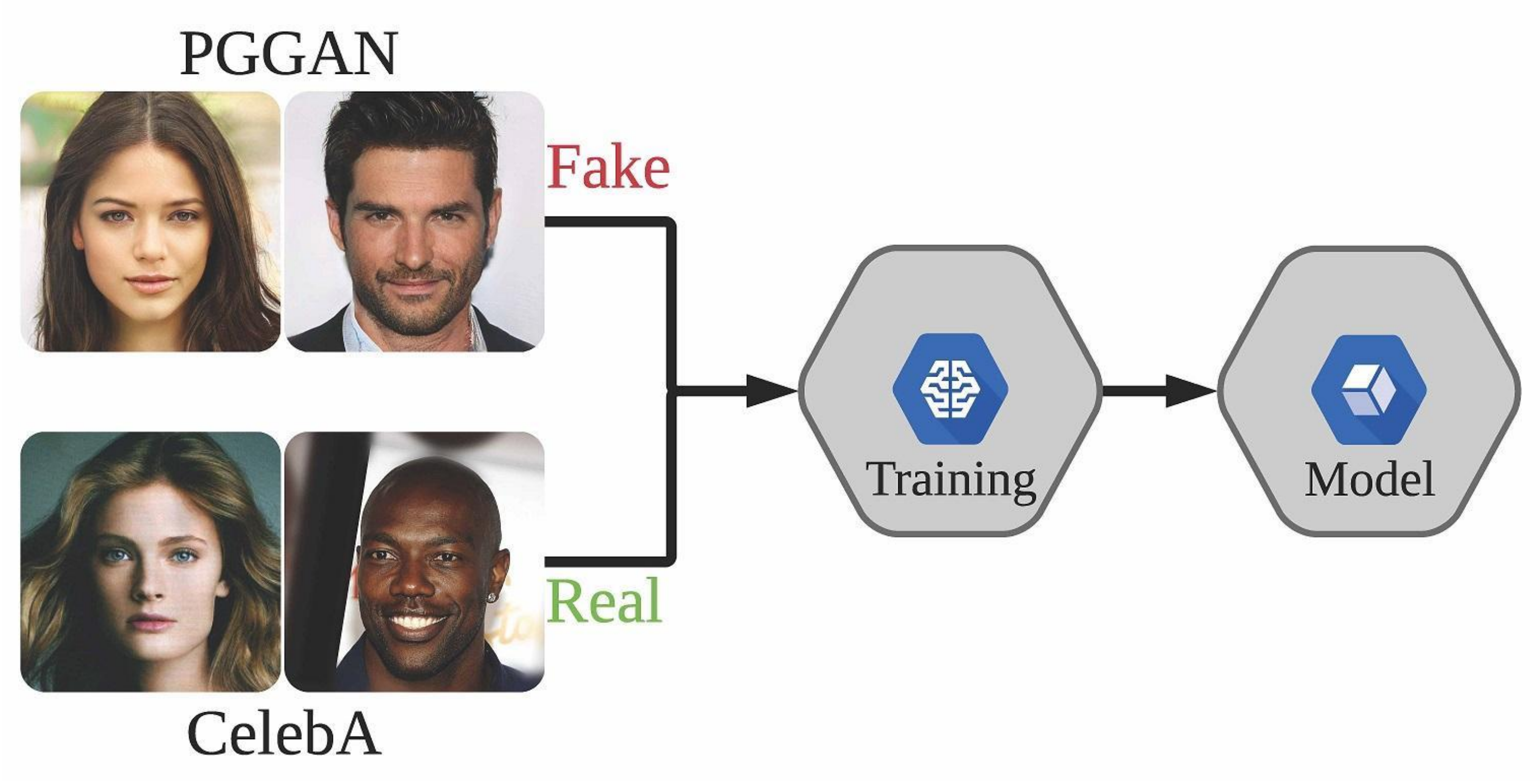
PGAN

MTCNN - Face Detection & Noise Filtering

- Red Boxes
  - Input for the classifier after alignment.
- Yellow Boxes
  - Marked by our filtering algorithm to ignore.

$$\frac{maxbox_{width} + maxbox_{height}}{\sqrt{3}} > box_{width} + box_{height}$$

Detection Methodology - GAN



Baselines & ShallowNet

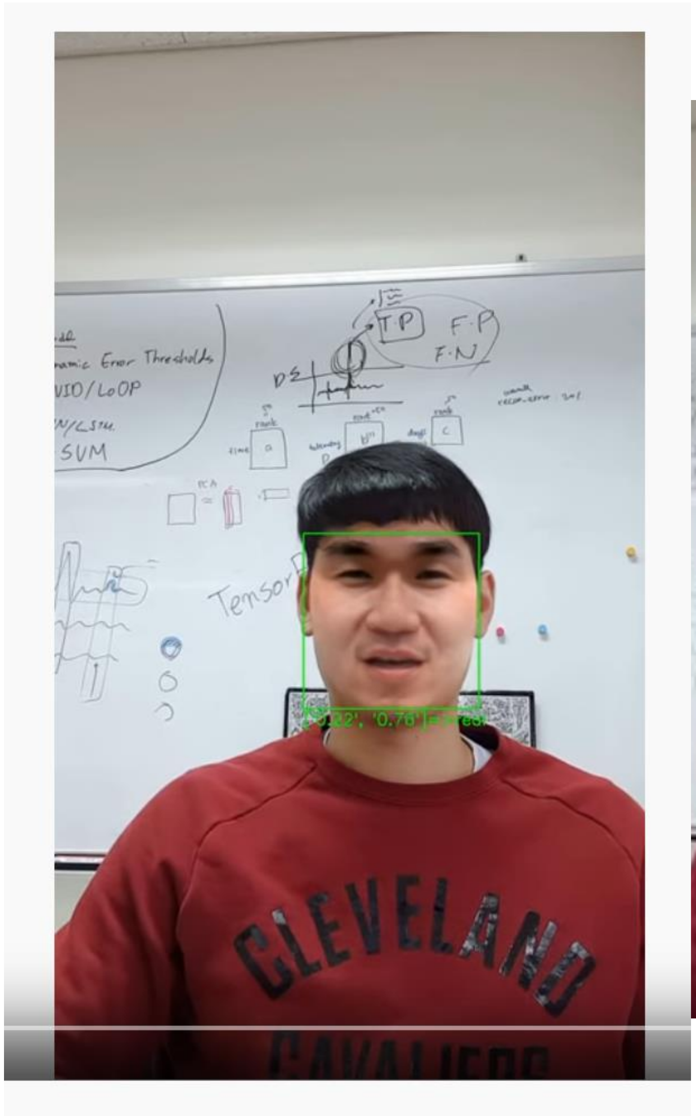
- Baselines
  a. VGG 16 & 19
  b. ResNet50
  c. InceptionV3
  d. InceptionResNetV2
  e. DenseNet121
  f. XceptionNet
- Our Method
  a. ShallowNet V1, V2 & V3

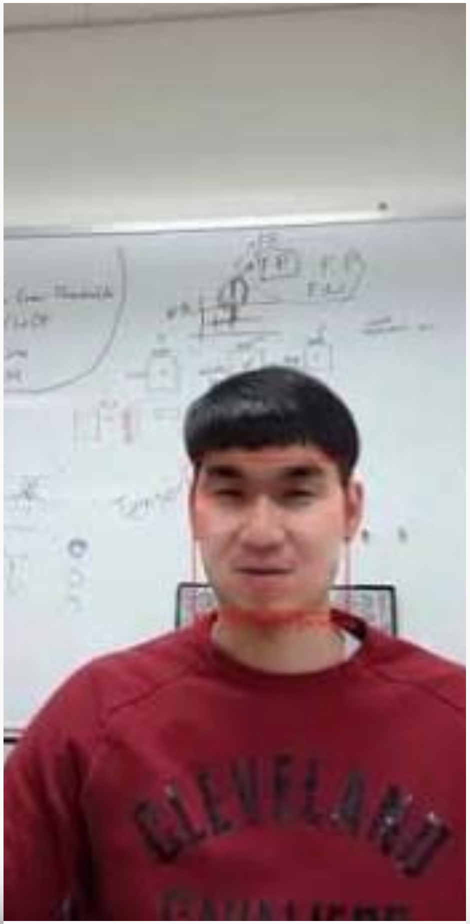
Evaluation - GAN

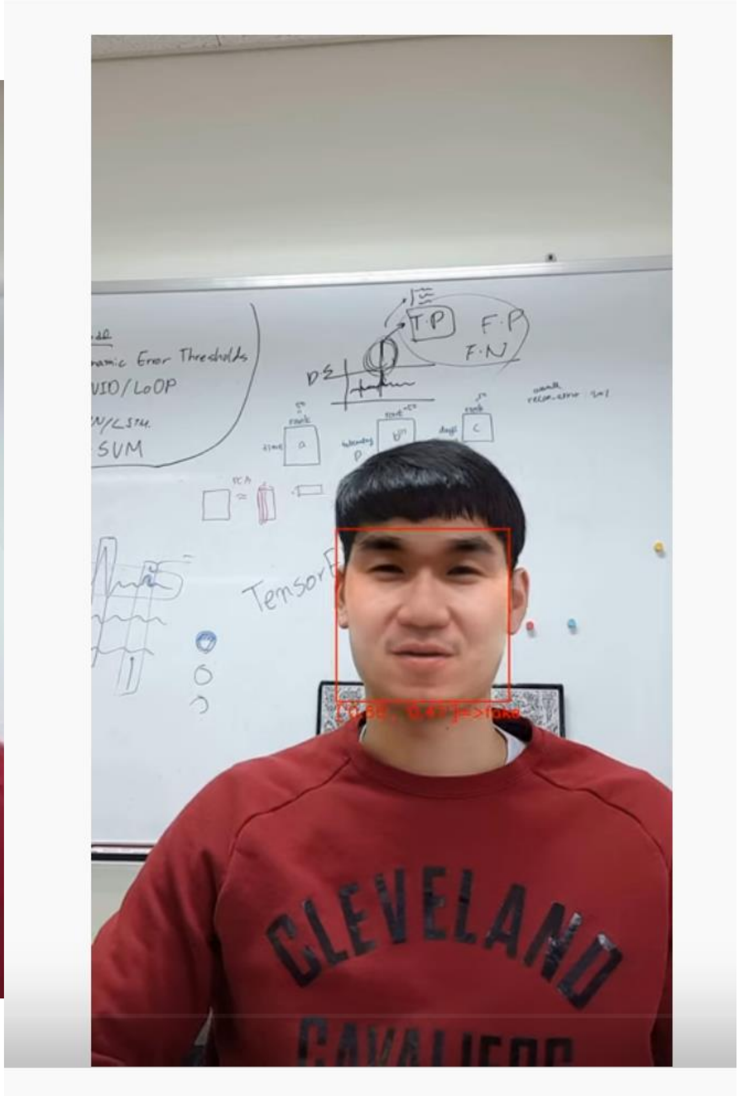
Method	AUROC (%)			
	64x64	128x128	256x256	1024x1024
VGG16	56.69	55.13	57.13	60.13
XceptionNet	79.32	79.03	82.03	85.03
NASNet	83.55	90.55	92.55	96.55
ShallowNetV1	84.94	98.12	99.82	**99.99**
ShallowNetV2	79.82	99.98	**99.99**	**99.99**
ShallowNetV3	90.85	**99.99**	**99.99**	**99.99**
Ensemble ShallowNet (V1 & V3)	**93.99**	**99.99**	**99.99**	**99.99**

Demo

- https://www.youtube.com/watch?v=kHUb6XVO0B4&feature=youtu.be







Few-Shot Learning for Talking Head Detection

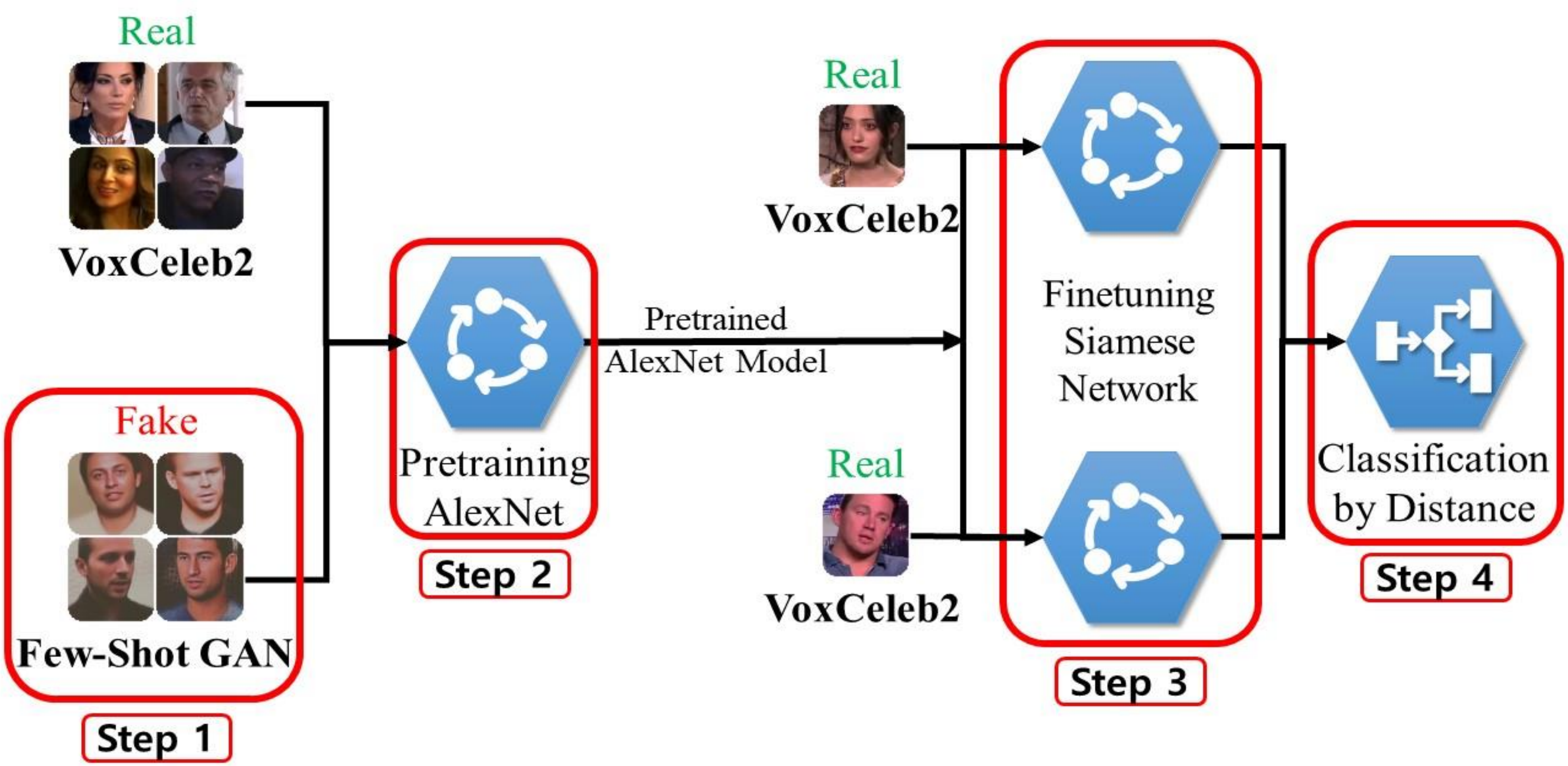
[3] Hyeonseong Jeon, Youngoh Bang, and **Simon S. Woo*, "FakeTalkerDetect: Effective and Practical Realistic Neural Talking Head Detection with a Highly Unbalanced Datase",** 10th International Workshop on Human Behavior Understanding (HBU), held in conjunction with ICCV'19 Nov, 2019 - Seoul, S. Korea

Main Challenges

- New generation methods
  - How to handle new attacks and generation methods?
  - Is there a way to leverage existing architectures or pre-trained models?
- Too long to generate new training dataset
  - Lack of training dataset?
  - Leverage existing dataset?

➔ Few-Shot Learning with re-usable approach?

FakeTalkerDetect
(Pre-training and Siamese Network)



[3] Hyeonseong Jeon, Youngoh Bang, and **Simon S. Woo*, "FakeTalkerDetect: Effective and Practical Realistic Neural Talking Head Detection with a Highly Unbalanced Datase",** 10th International Workshop on Human Behavior Understanding (HBU), held in conjunction with ICCV'19 Nov, 2019 - Seoul, S. Korea

Pre-training and Siamese Network

- Step 1 and 2. First, pre-trains well-known fake image classification model such as AlexNet, using real and fake image pairs.
- After pre-training, we further focus on improving the detection performance.
- Step 3. the Siamese network learns two input pairs (e.g., real-real) and evaluates sum of square error of each pair, where the higher error means that they are different classes.
- We use mean squared error loss function for fine-tuning, where this loss function runs over pairs of samples.

FDFtNet: Facing Off Fake Images using Fake Detection Fine-tuning Network

Hyeonseong Jeon, Youngoh Bang, and **Simon S. Woo*,** SEC 2020
International Conference on Information Security and Privacy Protection (**IFIP-SEC**), Solvenia, Sept 2020 (**BK Computer Science IF=1**)

Objectives

In real world:

- Deepfake dataset is small (imbalanced dataset)
- New methods will be coming
- Need to reuse existing architectures and datasets as much as possible
- Can the existing methods can be fine-tuned on a few dataset?
- Need for a new robust fine-tuning neural network-based architecture

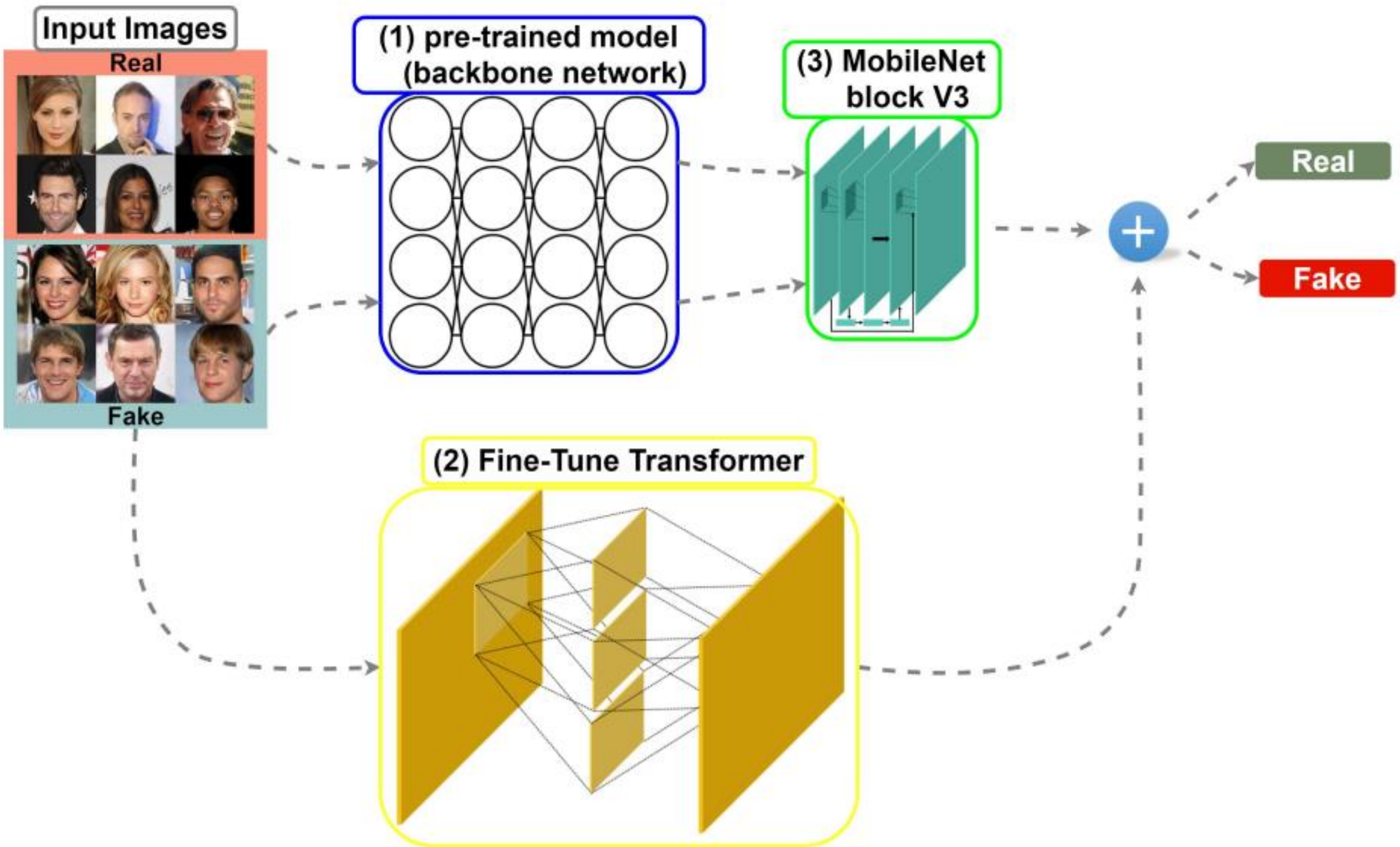
Fake Detection Fine-tuning Network (FDFtNet)

FDFtNet

- Explore Fine-Tune Transformer that uses only the **attention module** and the down-sampling layer.
- This module is added to the pre-trained model and fine-tuned on a few data to search for new sets of feature space to detect fake images.
- We experiment with our FDFtNet on the GANs based dataset (Progressive Growing GAN) and Deepfake-based dataset (Deepfake and Face2Face) with a small input image resolution of 64 x 64 that complicates detection.

Our FDFtNet achieves an overall accuracy of 90.29%

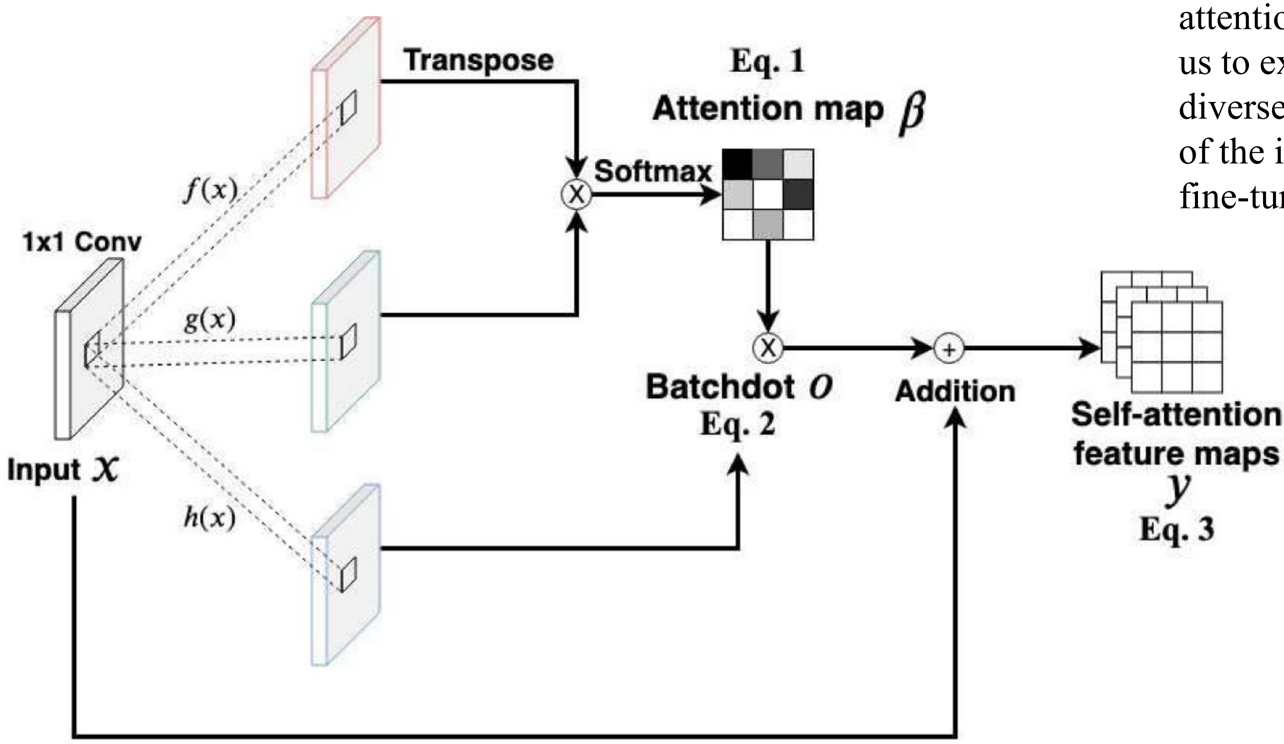
Architecture



The main reason we apply self-attention modules in FTT is to overcome the limitations of CNN in achieving long-term dependencies, caused by the use of numerous Conv filters with a small size.

Fine Tune Transformer (FTT)

A three-time application of self-attention modules allows us to explore and learn diverse deep features of the input images via fine-tuning.



Use different feature extraction from images using the self-attention

Evaluation results

Model	Dataset	PGGAN		Deepfake		Face2Face	
	Backbone	ACC (%)	AUROC	ACC (%)	AUROC	ACC (%)	AUROC
SqueezeNet	baseline	50.00	50.00	50.00	50.00	50.00	50.00
FDFtNet (Ours)	SqueezeNet	88.89	92.76	92.82	97.61	87.73	94.20
ShallowNetV3†	baseline	85.73	92.90	89.77	92.81	83.35	88.49
FDFtNet (Ours)	ShallowNetV3	88.03	94.53	94.29	97.83	84.55	93.28
ResNetV2	baseline	84.80	88.58	81.52	89.72	58.83	62.47
FDFtNet (Ours)	ResNetV2	84.83	94.05	91.03	96.08	85.15	92.91
Xception	baseline	87.12	94.96	95.10	98.92	85.78	93.67
FDFtNet (Ours)	Xception	**90.29**	**95.98**	**97.02**	**99.37**	**96.67**	**98.23**

Our approach provides a reusable fine-tuning network, improving the existing backbone CNN architectures. FDFNet requires only small amount data for fine-tuning and can be easily integrated with popular CNN architectures.

Extremely
Highly Imbalanced
Dataset
(No deepfake training data at all)

OC-FakeDect: Classifying Deepfakes Using One-class Variational Autoencoder

Hasam Khalid, and Simon S. Woo

CVPR Workshop on Media Forensics 2020 - Seattle, USA

DASH Lab, Sungkyunkwan University, South Korea

hasam.khalid@g.skku.edu, swoo@g.skku.edu

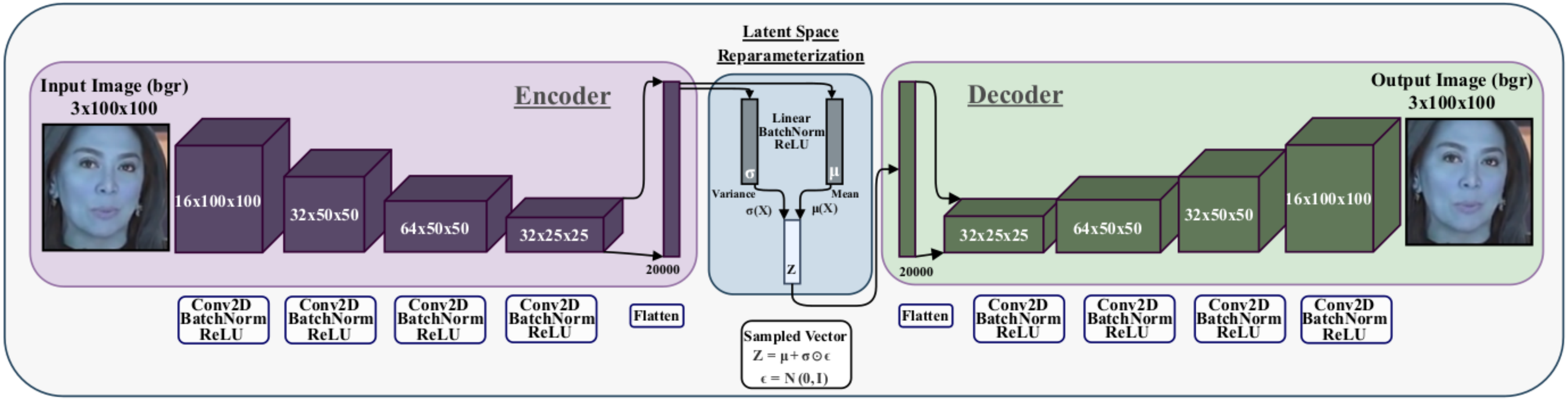
June 15 2020

Introduction

- We present **One-class classification** based deep-learning approach (**OC-FakeDect**)
  - Classifying Real and Fake Images using **One-class Variational Autoencoder**
  - Trained only on **Real** images
  - More generalizable approach

OC-FakeDect: One-class Deepfake Detection

- **One-class Variational Autoencoder (OC-VAE) Architecture Diagram with latent space reparameterization**



Dataset:

- Used **FaceForensics++ HQ** dataset.
- Used **Real** images for training, and **Real** and **Fake** images for testing.

T-GD: Transferable GAN-generated Images Detection Framework

Hyeonseong Jeon[1], Youngoh Bang[1], Junyaup Kim[2], and **Simon S. Woo[1]**
DASH Lab, Sungkyunkwan University,
South Korea

ICML 2020

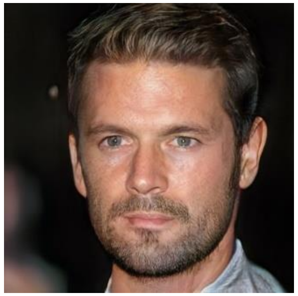
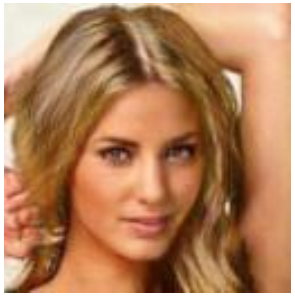
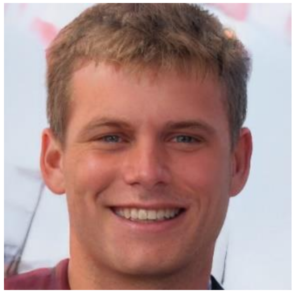
ICML | 2020
Thirty-seventh International
Conference on Machine Learning

Focus on Detecting GAN generated images through Transfer Learning

- Real images *[CelebA, CelebA-HQ, FFHQ]*



- GAN generated images *[PGGAN,StarGAN,StyleGAN,StyleGAN2]*


Getting harder to detect GANs

GAN-images are getting sophisticated that erasing artifacts, genuine patterns on the image.

- Example image

StarGAN (2017.11)

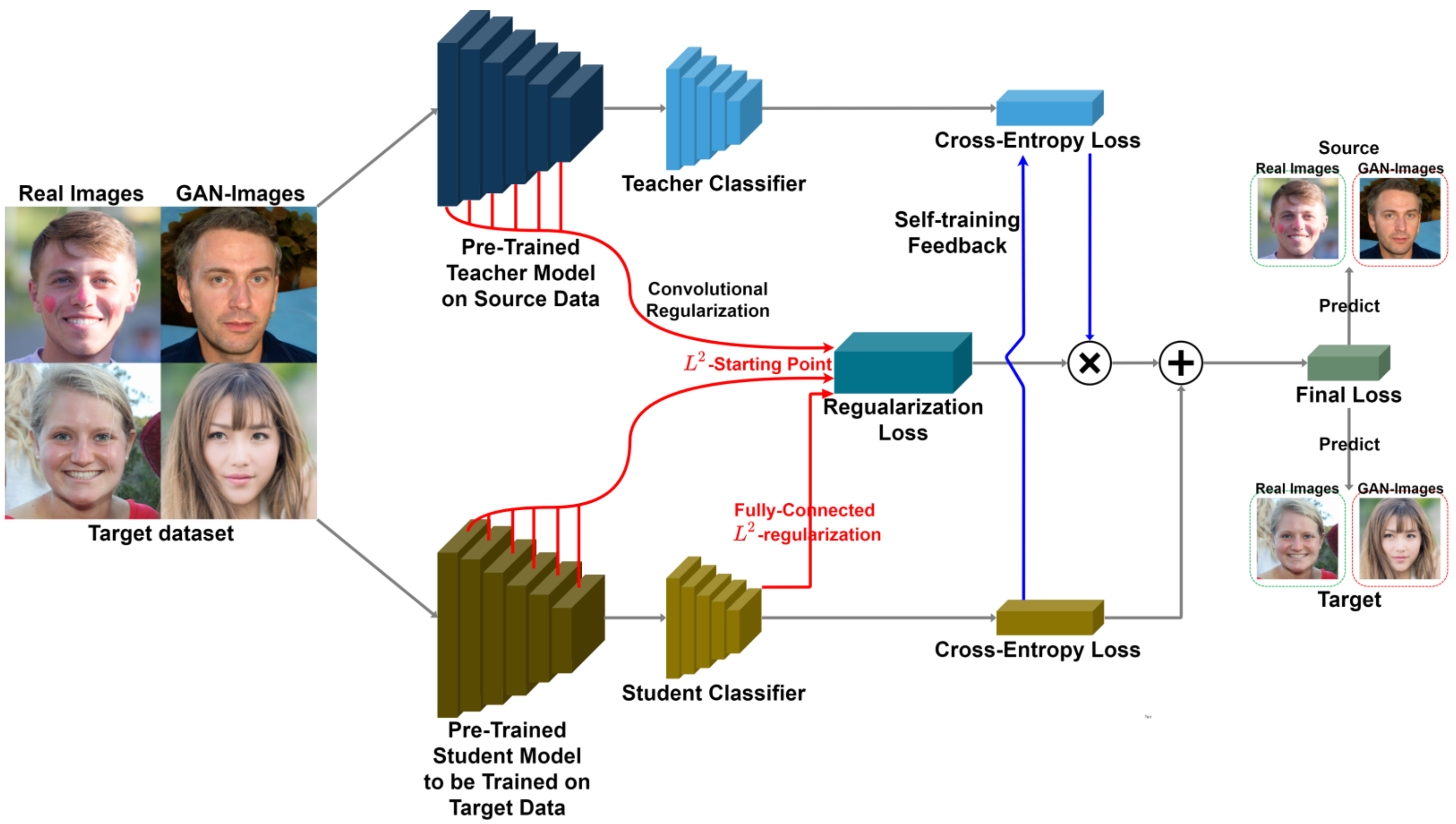
StyleGAN2 (2019.12)

Most approaches show relatively weak results for **transfer learning** ability

Motivations

1. High performance in different GAN-image detection
2. Transfer-learning with small target data
3. No catastrophic forgetting in this transfer process
4. Generalized way to augment input image to detect GAN-image

Proposed Architecture



Results

Method	Category / Dataset	Zero-shot (Pre-trained model) PGGAN	Zero-shot StarGAN	Zero-shot StyleGAN	Zero-shot StyleGAN2	Transfer Learning PGGAN	Transfer Learning StarGAN	Transfer Learning StyleGAN	Transfer Learning StyleGAN2
GeneralTransfer	PGGAN	99.91%	56.81%	49.47%	49.32%	99.86%	87.06%	54.17%	54.18%
EfficientNet-B0	StarGAN	66.47%	99.88%	52.01%	52.10%	95.90%	89.87%	99.03%	99.04%
(Base model)	StyleGAN	49.80%	50.04%	99.96%	99.97%	66.89%	51.12%	99.94%	99.95%
	StyleGAN2	45.23%	49.00%	99.99%	99.99%	91.33%	88.16%	45.26%	47.37%
ForensicTransfer†	PGGAN	97.15%	50.27%	53.57%	53.27%	69.35%	72.40%	76.50%	76.50%
	StarGAN	47.09%	85.34%	49.51%	49.48%	90.14%	51.32%	53.14%	53.14%
	StyleGAN	49.23%	49.66%	99.12%	99.97%	76.57%	58.93%	65.83%	65.85%
	StyleGAN2	49.22%	49.66%	99.12%	99.12%	76.58%	58.94%	65.84%	65.84%
T-GD	PGGAN	99.91%	56.81%	49.47%	49.32%	**95.87%**	91.61%	**98.12%**	**98.13%**
EfficientNet-B0	StarGAN	66.47%	99.88%	52.01%	52.10%	94.94%	**97.32%**	97.29%	93.34%
(Base model)	StyleGAN	49.80%	50.04%	99.96%	99.97%	84.92%	90.00%	**97.83%**	97.71%
	StyleGAN2	45.23%	49.00%	99.99%	99.99%	84.91%	90.01%	97.83%	**97.71%**
T-GD	PGGAN	99.81%	61.25%	49.76%	49.91%	94.91%	93.21%	87.37%	87.58%
ResNext32×4d	StarGAN	41.43%	99.78%	48.37%	48.50%	98.88%	96.15%	91.48%	91.26%
(Base model)	StyleGAN	41.05%	49.16%	99.99%	99.99%	85.93%	79.69%	94.31%	94.31%
	StyleGAN2	38.90%	50.31%	99.90%	99.88%	87.20%	80.19%	**98.39%**	95.38%

Conclusions

1. High performance on GAN-image detection without metadata.

2. Transfer learning method with little target dataset to prevent catastrophic forgetting.

3. General augmentation method on GAN-image detection.

Current Government &Industry Efforts

국내기업

제2의 'n번방' 막는다…카카오, 아동 성범죄에 '무관용'조항 신설

오로라 기자

본문듣기

입력 2020.06.26 08:58

26일 카카오, 운영정책에 관련 조항 신설
n번방 금지법 시행 앞두고 선제적 조치

기업 뿐만 아니라, 국내 다 부처 간에 협력 및 공동대응 필요

Concluding Remarks

- Balance between advancement of AI vs. Security/Privacy
- Malicious use
- Good AI vs. Bad AI

Deepfakes don’t hurt people, people using deepfakes hurt people.”

사람을 해치는 것은 딥페이크가 아니라 딥페이크를 “악용” 하는 사람이다

Questions & Comments
Thank You!

우사이먼성일 (성균관대 데이터사이언스융합학과)
Contact us: swoo@g.skku.edu
https://dash.skku.edu/
https://dash-lab.github.io/

Student Paper Authors at SKKU