이미지/비디오에서 나타나는 가짜 특징에 기반한 딥페이크 탐지모델

허란*, 조은진**
*동양미래대학교 생명화학공학과
**Dept. of Mechanical Engineering, RMIT University
hran9462@gmail.com, jintazcho@gmail.com

요 약

딥페이크는 딥러닝을 사용해서 만들어진 가짜 이미지를 의미한다. 딥페이크 비디오가 SNS를 비롯한 인터넷에 확산되면서 이와 관련된 사회문제가 발생하고 있다. [4,5,6] 딥페이크를 탐지할 수 있는 모델의 중요성이 대두되면서 많은 탐지모델이 제안되었지만, 딥페이크 생성 알고리즘 중 일부만 탐지할 수 있어 현실에 적용하기 어렵다. (표2 참조) 본 논문에서는 현실에 있는 다양한 딥페이크 생성 알고리즘에 대응할 수 있도록 CNN(Convolutional Neural Network)으로 가짜 특징을 탐지하고 RNN(Recurrent Neural Network)으로 가짜인지 판단하는 모델을 제안하며, 실험을 통해 제안된 구조의 탐지모델이 현실에서 사용될 수 있음을 입증하고자 한다. 갈수록 다양해지는 생성 알고리즘을 탐지 가능한지 입증하기 위해 탐지모델을 보완하면 탐지율이 향상되는지 실험을 진행했다. 또한, 다양한 딥페이크 생성 알고리즘에 대응하기 위해 여러 연구자의 참여가 필요하다.[1] 이를 위해 제시된 구조로 딥페이크 탐지모델을 학습하여 탐지했을 때의 탐지율이 고사양 GPU를 사용한 모델[37, 38]과 비슷함을 증명하고자 했다. 실험 결과는 탐지모델 보완 후 Recall 2%에서 75%으로 증가, FPR 1.6%에서 0.1%으로 감소, AUC 0.02에서 0.77으로 증가하였으며, 고사양의 GPU를 사용한 모델[37, 38]과 비슷한 탐지율을 보였다.

키워드 – Deepfake Detection, Face synthesis Detection, Convolutional Neural Network, Recurrent Neural Network

1. 서론

2017년 말 Reddit에 Deepfakes라는 유저가 딥러닝을 사용한 얼굴 조작 방식으로 주목받은 이후, 가짜 얼굴을 생성하는 알고리즘이 갈수록 정교해지며 다양화되고 있다. (표1 참조) 대표적인 가짜 얼굴 생성 방식은 얼굴을 포함한 머리 전체를 생성하는 방식(Face synthesis)과 얼굴이나 표정을 바꾸는 방식(Face swap)이 있다. 이 방식을 사용한 딥페이크 영상은 오바마, 푸틴, 힐러리 같은 정치인의 가짜 연설 영상이나 포르노에 유명인 얼굴을 합성한 영상이 있다. [1, 2, 3] 이런 영상은 가짜 뉴스 생성, 명예 훼손 등 사회문제[4, 5, 6]를 야기할 수 있으므로 딥페이크 관련 영상물 법제화[7, 8, 9]와 함께 가짜 얼굴을 탐지하는 것이 중요한 문제로 대두되었다.

표2의 현재 발표된 탐지모델은 학습 이미지와 같은 알고리즘으로 생성된 가짜 이미지에 대해 높은 탐지율을 보인다. 하지만 이러한 탐지모델들은 대부분 알고리즘 하나에만 특화되어있기 때문에 다른 알고리즘으로 만들어진

<표1> 다양한 딥페이크 생성 알고리즘

딥페이크 생성 알고리즘	생성 방식
Began[11]	Face synthesis
CausalGAN[12]	Face synthesis
faceswap/deepfake[13]	Face swap
StarGAN[14]	Face synthesis
Enrique Sanchez and Michel Valstar[15]	Face synthesis
MWGAN[16]	Face synthesis
ALAE[17]	Face synthesis
StyleGAN[18]	Face synthesis
MSG-GAN[19]	Face synthesis
FQGAN[20]	Face synthesis
ProGAN[21]	Face synthesis
StyleGAN v2[22]	Face synthesis
COCO-GAN[23]	Face synthesis
VAEGAN[24]	Face synthesis
HoloGAN[25]	Face synthesis
SPA-GAN[26]	Face synthesis
FTGAN[27]	Face synthesis
SEGAN[28]	Face synthesis
StarGAN V2[29]	Face synthesis
LSGAN[30]	Face synthesis
DCGAN[31]	Face synthesis
WGAN[32]	Face synthesis
GAN2play[33]	Face synthesis
Glow[34]	Face synthesis
GANnotation[35]	Face synthesis
deferred neural rendering[36]	Face synthesis
neural texture[36]	Face synthesis

1) github 주소모델 전체 코드, 사용 방법, 작동 과정 비디오 공개) : https://github.com/teamnova-ailab/Deepfake-detection-model-based-on-fake-attributes-shown-in-image-video/network/dependencies

가짜 얼굴에 대해서는 탐지율을 보장하지 못한다. [10] 가짜 얼굴을 생성하는 알고리즘이 갈수록 다양해지고 있으므로 기존의 탐지모델은 실제로 가짜 얼굴을 탐지할 때 사용하기 어렵다. 따라서 가짜 얼굴을 생성하는 새로운 알고리즘이 등장하더라도 탐지율이 보장되는 탐지모델이 필요하다.

그림1에서 볼 수 있듯이 가짜 얼굴을 생성하는 과정에서 생기는 결점 때문에 부자연스러운 이미지가 나타나는데, 이를 가짜 특징이 나타난 이미지라고 규정했다. 본 논문에서는 가짜 특징별로 CNN 모델을 만들어 가짜 특징을 검출하였으며, 이를 기반으로 Real/Fake를 판단하는 RNN 모델로 탐지모델을 구성했다. 이렇게 탐지모델을 구성함으로써 가짜 얼굴을 생성하는 새로운 방식이 등장해도 새로운 가짜 특징을 학습한 CNN 모델을 기존 탐지모델에 추가해서 사용할 수 있다. 또한, github에 모델 및 코드를 공개하여 오픈소스처럼 여러 사람이 모델 정확도 향상에 기여할 수 있다.

다른 탐지모델들이 Nvidia Titan X GPU[37], Nvidia Tesla P40 GPU[38]를 사용한 데 비해, 본 논문에서 제시하는 모델은 Nvidia RTX 2070 SUPER로도 충분히 학습할 수 있다. 각 가짜 특징에 적합한 필터가 적용된 이미지로 학습을 진행함으로써 Convolution 과정에서 불필요한 특징 추출은 줄이고 필요한 특징 추출은 쉽게 만들었기 때문이다. 따라서 100장, 200장의 적은 데이터로도 학습할 수 있어 학습 시간이 짧고, 하드웨어가 고사양이 아니더라도 모델을 만드는 게 가능하다. 이로 인해 딥페이크 탐지 연구를 할 때 있던 제약을 완화하고 더 많은 사람이 모델을 발전시킬 수 있는 발판을 제공한다.

(a)

(b)

(c)

(그림 1) 가짜 특징이 나타난 이미지 예시 (a) 배경에 비해 얼굴이 흐려 눈코입이 보이지 않는 이미지 (b) 얼굴에 선으로 된 노이즈가 생긴 이미지 (c) 안경을 쓰고 있으나 안경다리가 없는 이미지

<표2> 탐지모델 및 탐지모델 학습에 사용한 데이터 정리

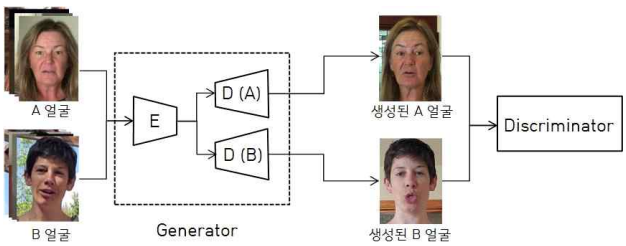
탐지모델	사용한 데이터셋
McCloskey and Albright (2018) [39]	NIST MFC2018
Yu et al. (2019) [40]	Own (ProGAN, SNGAN, CramerGAN, MMDGAN)
Wang et al. (2019) [38]	FF++, DFDC, Own (PGGAN, StyleGAN2, StarGAN, STGAN, StyleGAN, STGAN)
Stehouwer et al. (2019) [41]	DFFD (ProGAN, StyleGAN)
Nataraj et al. (2019) [42]	100K-Faces (StyleGAN)
Neves et al. (2019) [43]	100K-Faces (StyleGAN) FSRemovalDB (StyleGAN)
Marra et al. (2019) [44]	Own (CycleGAN, ProGAN, Glow, StarGAN, StyleGAN)
Zhou et al. (2018) [45]	Own
Afchar et al. (2018) [46]	Own
Güera and Delp (2018) [47]	Own
Yang et al. (2019) [48]	UADFV
Li et al. (2019) [49]	UADFV DeepfakeTIMIT
Rössler et al. (2019) [50]	FF++
Matern et al. (2019) [51]	Own
Nguyen et al. (2019) [52]	FF++
Agarwal and Farid (2019) [53]	Own (FaceSwap, HQ)
Sabir et al. (2019) [54]	FF++
Bharati et al. (2016) [55]	Own (Celebrity Retouching, ND-IIITD Retouching)
Tariq et al. (2018) [56]	Own (ProGAN, Adobe Photoshop)
Wang et al. (2019) [57]	Own (InterFaceGAN/StyleGAN)
Jain et al. (2019) [58]	Own (ND-IIITD Retouching, StarGAN)
Marra et al. (2019) [59]	Own (Glow/StarGAN)
Zhang et al. (2019) [60]	Own (StarGAN/CycleGAN)
Amerini et al. (2019) [61]	FF++

2. 관련 연구

2.1. 가짜 얼굴 생성 방식

2.1.1. Face synthesis

Face synthesis는 GAN(Generative Adversarial Networks)[62]을 사용해서 머리카락을 포함한 얼굴 전체를 생성하는 방식이다. GAN은 이미지를 생성하는 Generator와 Real/Fake를 판단하는 Discriminator로 구성되어 있다. Generator는 다시 인코더와 디코더로 구성되어 있으며, 인코더에서 이미지의 특징을 학습하고 디코더에서 그 특징을 토대로 이미지를 재구성한다. 구체적인 동작 방식은 다음과 같다. A 얼굴과 B 얼굴 데이터셋으로 인코더에 각 얼굴의 특징을 학습시킨다. 학습된 특징을 기반으로 A 디코더는 A 얼굴을, B 디코더는 B 얼굴을 생성한다. (그림 2 참조) 최근에는 머리 스타일, 머리색, 눈동자 색, 수염, 얼굴 표정 등 개인의 특성을 보존하는 연구[63, 64, 65]가 활발히 진행되고 있다.



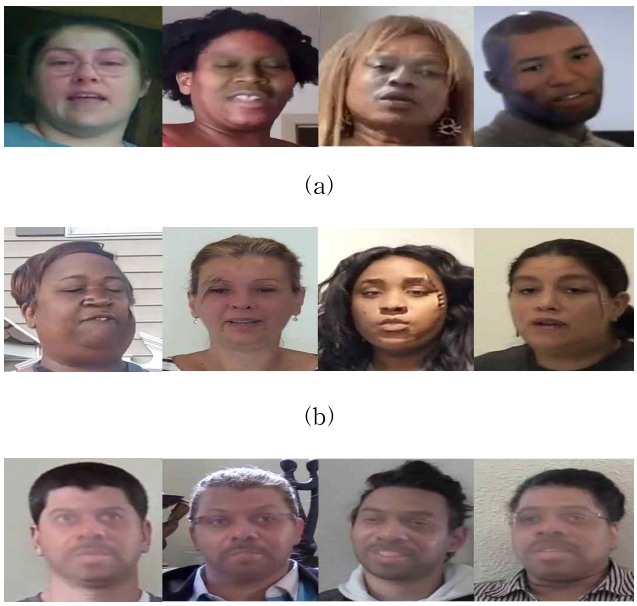
(그림2) face synthesis 생성 알고리즘 E : 인코더, D (A) : A 디코더, D (B) : B 디코더

2.1.2. face swap

Face swap은 원본 이미지/비디오에서 Source 얼굴을 Target 얼굴로 변경해서 가짜 얼굴을 생성하는 방식이다. 이 과정에서 Target 얼굴로 변경할 때 GAN을 사용하여 생성한 얼굴을 이용한다. 먼저 이미지/비디오에서 원본 얼굴을 탐지한 후 얼굴을 정렬(Face Align)한 Source 얼굴을 인코더의 입력값으로 사용한다. 인코더에서 추출한 특징을 토대로 디코더에서 Target 얼굴을 생성한다. 생성한 얼굴을 원본 이미지에 조정해서 넣은 뒤에 경계를 자연스럽게 한다. (그림3 참조) 이 방식은 이미지/비디오 모두에 적용될 수 있으며, 그림3의 (c) 부분에서 다양한 GAN 알고리즘을 사용하여 얼굴을 생성할 수 있다. 그러나 생성한 얼굴과 원본 이미지 간에 얼굴색 불일치 및 화질 불일치, 합성된 경계 부분의 부자연스러움 등이 나타날 수 있다. (그림4 참조)

2.2. 탐지모델

2.2.1. 본 논문과 비슷한 구조로 가짜 얼굴 탐지를 시도한 탐지모델

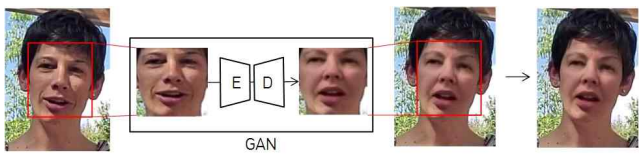


(a)

(b)

(c)

(그림4) face swap에서 나타날 수 있는 현상 (a) 생성한 얼굴과 원본 이미지 간에 얼굴색 불일치 (b) 합성된 경계 부분의 부자연스러움 (c) 생성한 얼굴과 원본 이미지 간에 화질 불일치



(a) (b) (c) (d) (e)

(그림3) face swap 생성 알고리즘 (a) 얼굴 검출 (b) 얼굴 자르기(Face Crop) 및 정렬(Face Align) (c) 얼굴 생성(GAN) (d) 얼굴 조정 (e) 합성된 경계를 자연스럽게 하기

표2에 정리된 것처럼, 가짜 얼굴을 탐지하기 위한 다양한 탐지모델이 존재한다. 그중 Li et al.[66], Güera and Delp[47], Sabir et al[48] 은 CNN과 RNN을 사용해 가짜 비디오에서 가짜 얼굴 탐지를 시도했다. Li et al.[66]은 가짜 비디오가 진짜 비디오보다 눈 깜빡임이 적다는 관찰에 기초해서 눈 깜빡임 간격으로 가짜 비디오를 탐지하고자 했다. 이를 위해 CNN으로 눈 이미지에서 특징을 추출한 후 RNN으로 눈 깜빡임 간격을 체크한다.

Güera and Delp[47]은 비디오에서 나타나는 특징을 토대로 가짜 비디오를 검출하고자 했다. 비디오 특징을 학습하기 위해 먼저 CNN으로 프레임에서 특징을 추출한 후 RNN으로 시간 흐름에서 가짜 비디오 탐지를 시도했다.

Sabir et al.[48]은 Güera and Delp[47]와 모델구조가

비슷하나, 얼굴 자르기(Face Crop) 및 정렬(Face Align) 전처리를 추가해서 CNN 특징을 추출하기 쉽게 만들었다는 점이 다르다. 세 탐지모델은 공통으로 비디오라는 특성을 이용해 시간적 흐름에서 나타나는 부자연스러움을 탐지하고자 CNN 모델에 RNN 모델을 통합해 사용했다.

2.2.2. 본 논문보다 좋은 성능의 GPU를 사용하여 가짜 얼굴 탐지를 시도한 모델

Dang et al.[37]은 데이터 불균형 문제가 발생했을 때도 가짜 얼굴 탐지를 잘할 수 있는 모델을 만들고자 했다. 현실에서는 가짜 이미지/비디오의 수보다 진짜 이미지/비디오의 수가 많기 때문이다. MANFA에 XGBoost을 통합한 HF-MANFA 모델을 제안하여 실험을 통해 입증을 시도했다.

Wang et al.[38]은 가짜 이미지 및 진짜 이미지에 네 가지 변형(노이즈, 블러, 압축, 리사이즈)이 일어나도 잘 탐지할 수 있는 모델을 제안했다. 학습 시 뉴런의 행동을 파악해서 가짜 얼굴을 탐지하는 모델을 제안하였으며, 네 가지 변형을 얼마나 탐지하는지 실험을 진행하였다. 또한, 학습에 사용하지 않은 새로운 데이터에 대해 탐지 모델이 얼마나 탐지할 수 있는지 확인하였다.

2.2.1. 탐지모델에 사용된 기술

2.2.1.1. CNN (Convolutional Neural Network)

가짜 특징 판정을 위한 CNN Detector는 SSD[67]와 Faster-RCNN[68]을 복합적으로 사용하였다. Faster-RCNN[68]은 2 Stage-Detector로 두 단계로 가짜 특징을 검출한다. 이미지가 입력값으로 들어오면 첫 번째 단계에서 RPN(Region Proposal Network)을 이용하여 가짜 특징이 있는 것 같은 영역에 박스를 친다. 두 번째 단계에서 ROI pooling 과정을 거쳐 박스 친 부분에 어떤 가짜 특징이 있는지 확인한다. Faster-RCNN[68]은 Convolution 과정을 두 번 거치므로 정확도가 높은 반면, 속도가 느린 단점이 있다.

SSD[67]는 1 Stage-Detector로 한 단계로 가짜 특징을 검출한다. 입력 이미지가 들어왔을 때 Convolution 과정마다 미리 크기를 정해놓은 Default box 안에 가짜 특징이 있는지 확인한다. Convolution 과정이 한 번이므로 속도가 빠른 반면, Convolution 과정을 덜 거친 이미지는 저차원 특징(직선, 곡선, 점 등)을 가지므로 고차원 특징(눈, 코, 입, 얼굴 등)이 필요한 이미지는 탐지율이 떨어질 수 있다.

이와 같은 특성을 고려하여 탐지모델 속도를 높이기 위해 눈, 코, 입에서 발견되는 특징에는 SSD 모델[67]을 사용하였으며 얼굴에서 발견되는 특징 중 일부에만 Faster-RCNN 모델[68]을 사용하였다. 눈, 코, 입에서 발견되는 특징은 눈, 코, 입을 탐지 후 다시 가짜 특징을 탐지해야하기 때문이다. 또한, 얼굴에서 나타나는 특징 중 저차원적인 특징이 나타나는 곳에 SSD를, 고차원적인 특징이 나타나는 곳에는 Faster-RCNN을 사용하였다.

2.2.1.2. RNN (Recurrent Neural Network)

Real/Fake 판단을 위한 RNN 모델은 BERT(Bidirectional Encoder Representations from Transformers)[69]를 사용하였으며 BERT가 가짜 및 진짜 이미지를 판단하는 과정은 다음과 같다. Transformer[70]의 Self-Attention(1), (2) 및 Feed-Forward Network(3)를 통해 가짜 특징들의 연관성을 토대로 Real/Fake 를 판단한다.

$$Attention(Q,\ K,\ V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

$$MultiHead(Q,\ K,\ V\ ) = Concat(head_1,\cdots,head_h)W^O \quad (2)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V\ )$$

$$FFN(x) = max(0,\ xW_1\ + b_1\ )W_2\ + b_2 \quad (3)$$

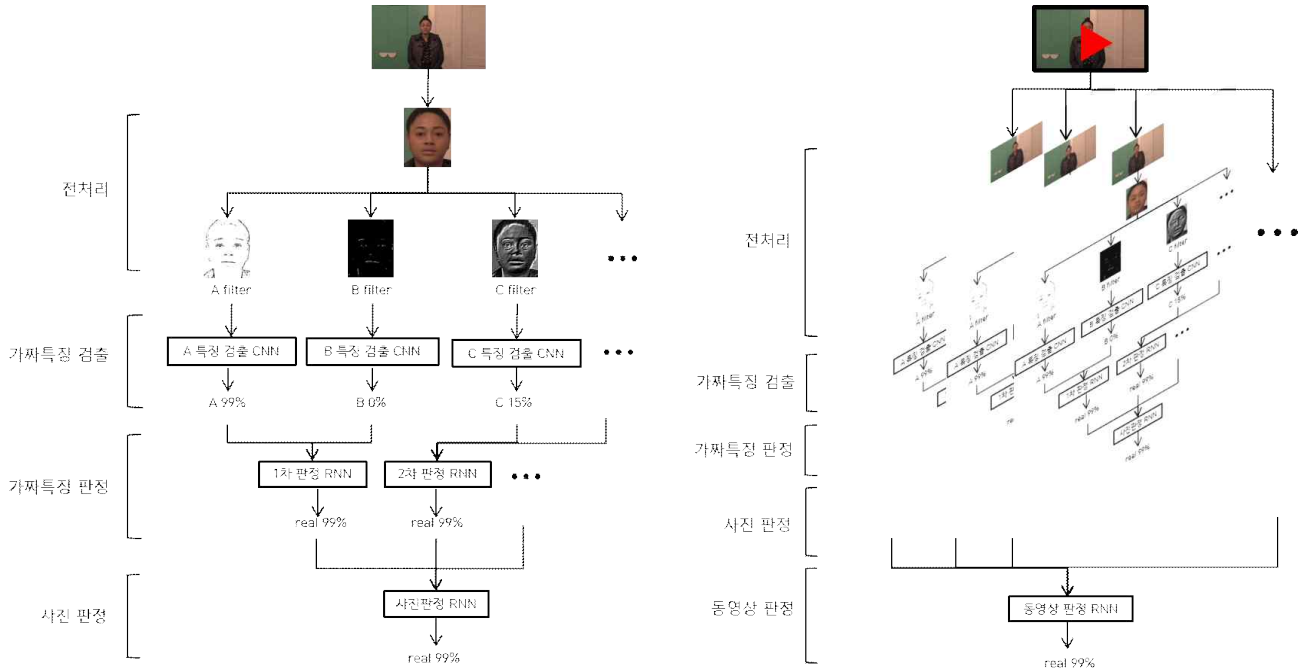
자세한 과정은 다음과 같다. CNN Detector로 검출된 가짜 특징들을 i번 Scaled Dot-Product Attention(1) 연산하여 어떤 특징들이 연관성이 있는지 파악한다. i번 연산한 결과값을 병합해 다시 가중치 행렬을 곱하여(Multi-Head Attention(2)) 결과값들을 통합한다. 각 특징의 Multi-Head Attention 결과값을 Position-Wise Feed-Forward Networks(3) 입력값으로 사용하여 최종적으로 가짜 특징이 서로 얼마나 영향을 주는지 파악한다.

3. 제안 방법

본 논문에서 제시하는 탐지모델은 크게 전처리, 가짜 특징 검출, 가짜 특징 판정, 이미지 판정, 비디오 판정 다섯 단계로 이루어져 있다. 전체적인 모델 구성은 그림5에 제시되어 있으며, 단계별로 나누어 모델의 구조를 자세히 설명한다.

3.1. 전처리

이미지/비디오가 들어오면 먼저 얼굴 검출을 통해 얼굴 부분만 잘라낸다. 얼굴 이미지에서 각 특징에 최적화된



(a) (b)

(그림 5) 전체 탐지모델 구조 (a) 이미지 탐지 과정 (b) 비디오 탐지 과정

필터를 적용한 후 가짜 특징 검출 단계로 넘어간다. 최적화된 필터조합을 찾기 위해 가짜 특징 데이터를 연구자가 보고 직접 분류하여 단일 특징에 맞는 필터조합을 찾는다. 필터조합은 한 개 필터부터 다중 필터까지 다양하며, 진짜 이미지와 가짜 이미지 모두 필터를 적용했을 때 구분이 가능한 상태가 되면 적합한 필터로 선정한다.

3.2. 가짜 특징 검출

이 단계에서는 CNN 모델을 사용하여 이미지/비디오에서 발견되는 가짜 특징을 검출한다. 데이터를 가짜 특징별로 분류하여 이를 탐지하는 CNN 모델을 구성한다. 각 특징에 맞게 필터를 적용한 얼굴 이미지가 입력값으로 사용되어 CNN 모델을 거쳐 각 특징이 존재하는지 판단한다. 처리속도를 고려해서 가짜 특징을 탐지하는 각 모델이 병렬로 처리되어 처리결과를 반환한다. 생성 방식이 다양하고, 생성 방식 안에서도 가짜 특징이 다양하게 나타날 수 있으므로 직렬 처리 시 실시간 탐지가 어렵다.

3.3. 가짜 특징 판정

가짜 특징 판정에서는 검출된 가짜 특징 정보를 토대로 RNN 모델이 Real/Fake를 판단한다. 가짜 특징 정보는 관련 있는 것끼리 묶여 RNN 모델의 입력값으로 사용된다. 가짜 특징 중 단일로 Real/Fake 판단이 가능한 것도 있지만 다수의 특징이 합쳐졌을 때 Real/Fake를 판단할 수 있는 것도 있으므로 이 과정이 필요하다.

3.4. 이미지 판정

이미지 판정은 3.3(가짜 특징 판정)의 결과값을 입력으로 사용하여 판단한다. 가짜 특징 판정이 여러 상황으로 나누어져 있으므로 RNN을 통한 이미지 판정 시 종합적인 판단이 가능하다. 입력값으로 이미지가 들어왔을 경우 이 과정에서 종료하며, 입력값이 비디오일 경우 3.1(전처리)부터 3.4(이미지 판정)까지의 과정을 선택한 프레임 수만큼 반복한 후 비디오 판정을 실행한다.

3.5. 비디오 판정

비디오 판정은 선택한 프레임 수만큼의 이미지 판정 결과값을 입력으로 사용한다. 비디오는 이미지의 집합이기 때문에 이미지 한장 한장에 대한 정보가 모여 비디오의 Real/Fake를 판정하게 된다.

4. 실험

4.1. 실험환경 구성

4.1.1. 데이터셋

탐지모델 학습 데이터는 표3에 기재되어있는 모델로 데이터를 생성하여 구성했다. 현실에는 다양한 모델로 생성된 가짜 이미지가 있으므로 다양한 가짜 특징을 탐지할수록 현실에서 사용 가능한 탐지모델에 가까워진다.

따라서, 탐지모델의 학습 데이터로 많은 생성 알고리즘의 다양한 데이터를 포함하고자 했다. 생성된 데이터가 100장 이하이거나, 가짜 특징이 100장 이하인 데이터는 제외하였으며 표3에 제시된 알고리즘으로 생성한 총 28개 데이터 중 17개 데이터를 사용하여 CNN Detector를 만들었다.

테스트를 위한 가짜 얼굴 및 진짜 얼굴 데이터셋은 접근이 편리하고, 편향된 결과 방지를 위해 공공 데이터인 generated photo[71]와 celeb-A[72]를 사용하였다.

4.1.2. 성능 평가지표

성능 평가지표는 FPR(False positive rate), Recall, AUC(Area Under Curve)를 사용하였다. FPR은 진짜 데이터를 가짜 데이터로 잘못 탐지했을 확률을 나타내는 지표이다. Recall은 가짜 데이터를 가짜라고 탐지했을 확률을 나타내는 지표이다. AUC는 ROC curve (Receiver Operating Characteristic curve) 면적을 나타낸 지표로, x축은 FPR(1) y축은 Recall(2)이다. 모델의 FPR, Recall을 Threshhold 별로 좌표평면에 나타낸 것을 ROC curve라고 한다. 이 그래프를 다른 모델과 정량적으로 비교하기 위해 그래프의 밑면적을 계산한 것이 AUC이다.

$$FPR = \frac{FP}{FP + TN} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

딥페이크 탐지는 가짜 얼굴과 진짜 얼굴을 각각 정확하게 탐지하는 게 중요하다. 따라서 가짜 데이터를 가짜라고 예측한 확률을 나타내는 Recall(2)이 높게 나타나고, 진짜 데이터를 가짜라고 예측한 확률을 나타내는 FPR(1)이 낮게 나타나야 좋은 탐지모델이라고 할 수 있다. 또한, threshold에 따라 이를 종합적으로 표현하는 ROC curve의 면적인 AUC가 1에 가까울수록 좋은 탐지모델이다.

4.1.3. 환경 구성

학습 및 테스트 환경 구성은 다음과 같다.

OS : Ubuntu 18.04

CPU : AMD Ryzen 5 2400G

GPU : Nvidia RTX 2070 SUPER

RAM : DDR4 32GB

4.2. 가짜 특징을 추가했을 때 탐지율이 향상되는지 테스트

현실에는 다양한 생성 알고리즘으로 생성한 이미지가 존재하며, 새로운 생성 알고리즘이 계속 발표되고 있다. 반면에, 탐지모델은 학습에서 고려하지 못한 특징이 나타나면 탐지율이 낮아지게 된다. 따라서 새로운 특징의 이미지에 대비하기 위해 탐지모델을 보완했을 때 탐지율이 보장되어야 한다. 이를 위해 탐지모델에 CNN Detector를 추가했을 때에 탐지율이 높아지는지 테스트를 진행하였다. 테스트 방법은 다음과 같다.

<표3> 딥페이크 생성 알고리즘으로 생성한 총 데이터 수와 가짜 특징이 나타난 데이터 수

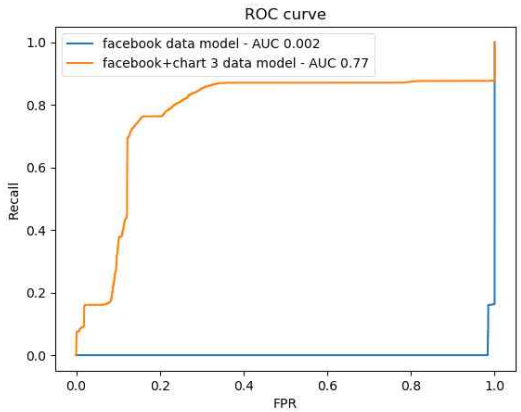
딥페이크 생성 알고리즘	총 데이터 수	학습에 사용한 데이터 수
Began[11]	이미지 512장	512장
CausalGAN[12]	이미지 300장	140장
faceswap/deepfake [13]	이미지 1,576장	1,576장
StarGAN[14]	이미지 2,240장	1,747장
Enrique Sanchez and Michel Valstar[15]	이미지 9,216장	8,694장
MWGAN[16]	이미지 1,281장	500장
ALAE[17]	이미지 184장	184장
StyleGAN[18]	이미지 1,000장	1,000장
MSG-GAN[19]	이미지 1,000장	1,000장
FQGAN[20]	이미지 1,000장	581장
ProGAN[21]	이미지 1,000장	501장
StyleGAN v2[22]	이미지 1,000장	579장
COCO-GAN[23]	이미지 1,024장	941장
VAEGAN[24]	이미지 700장	637장
HoloGAN[25]	이미지 640장	-
SPA-GAN[26]	이미지 401장	-
FTGAN[27]	이미지 310장	-
SEGAN[28]	이미지 300장	-
StarGAN V2[29]	이미지 362장	-
LSGAN[30]	이미지 100장	-
DCGAN[31]	이미지 100장	-
WGAN[32]	이미지 100장	-
GAN2play[33]	이미지 64장	
Glow[34]	이미지 3장	-
GANnotation[35]	비디오 1개	-
deferred neural rendering[36]	웹캠	-
neural texture[36]	웹캠	-

a. 페이스북에서 주최한 대회[73]에서 제공한 데이터로 학습된 기존의 탐지모델로 테스트셋에 대한 탐지를 진행한다.

b. 표1의 데이터로 CNN Detector 학습을 진행하여 모델을 보완해 같은 데이터셋에 대해 테스트를 다시 진행한다.

c. 모델에 가짜 특징을 추가하기 전과 후의 탐지율을 비교하여 CNN Detector를 추가할수록 논문에서 제시한 모델의 탐지율이 증가함을 입증한다.

페이스북 데이터셋 및 페이스북 데이터셋과 표3의 데이터셋을 사용하여 학습한 모델로 테스트를 진행한 결과는 그림6과 같다. 가짜 얼굴 데이터를 잘 탐지했는지 나타내는 Recall이 2%로 낮고, 진짜 얼굴 데이터를 잘못 탐지했는지 나타내는 FPR은 1.6%으로 낮게 나타났다. Threshold에 따라 탐지모델의 Recall과 FPR을 보여주는 ROC curve의 아래 면적을 계산한 AUC도 0.002로 낮게 나타났다. 그러나 표1의 데이터로 CNN Detector를 추가하여 모델을 보완한 후엔 Recall 75%로 증가, FPR 0.1%로 감소, AUC 0.77으로 증가했다.



(그림6) 탐지모델 테스트 결과

테스트 결과를 바탕으로 탐지할 수 있는 가짜 특징이 많아지면 탐지율이 상승함을 알 수 있다. facebook 모델은 14개 가짜 특징 탐지가 가능하였으나 표3의 데이터로 CNN Detector를 추가한 모델은 35개의 가짜 특징 탐지가 가능하기 때문이다. 따라서 가짜 특징을 탐지하는 CNN Detector가 많아지면 탐지율이 보장됨을 증명하였다.

4.3. 고사양 GPU를 사용한 모델과 비교

고사양 GPU를 사용한 모델과 비교를 통해 학습에서 고사양의 하드웨어를 사용하지 않은 탐지모델도 탐지율이 높을 수 있음을 증명하고자 한다. 현실에서 나타나는 다양한 가짜 이미지를 탐지하기 위해 여러 연구자의 참여가 필요한데, 고사양 GPU로만 탐지모델 학습이 가능하면 연구에 제한이 생기기 때문이다. 따라서 이 실험을 통해 딥페이크 탐지 연구의 제약을 완화시키고 더 많은 연구자가 가짜 이미지 탐지 연구에 참여할 수 있도록 하고자 한다.

이에 대한 실험으로 모델 학습을 위해 사용한 RTX 2070 GPU보다 성능이 좋은 GPU를 사용하여 학습한 모델들[37, 38]과 비교를 진행했다. 표4에 정리된 것처럼 비교하려는 각 모델은 통일된 데이터셋을 사용해 실험을 진행하지 않았다. 또한, 코드가 공개되어 있지 않아 논문에 기재되어있는 성능평가지표인 AUC를 기준으로 비교를 진행했다. (표4 참조)

<표4> 비교하려는 탐지모델이 사용한 GPU 및 데이터셋 정리

탐지모델	사용한 GPU	사용한 데이터셋
Dang et al. [11]	Titan X 12GB GPU	- MANFA dataset - SwapMe and FaceSwap dataset
Wang et al. [12]	Tesla P40 GPU	REAL - CelebA - Flicker-FacesHQ(FFHQ) - FaceForensics++ - DFDC - Celeb-DF FAKE - PGGAN - StyleGAN2 - StarGAN - STGAN - StyleGAN - FF++ - DFDC

[37], [38]의 탐지모델과 제시한 탐지모델의 평가지표 AUC를 비교한 결과 본 논문에서 제시한 탐지모델이 약 0.13 낮게 나타났다. (표5 참조) 그러나 4.2.에서 볼 수 있듯이 [37], [38]과 달리 제시된 탐지모델은 가짜 특징을 추가할수록 탐지모델 성능이 좋아질 수 있다.

<표5> 고사양 GPU를 사용한 모델과 논문에서 제시한 모델 간 AUC 값 비교

탐지모델	AUC
Dang et al. [11]	0.93
Wang et al. [12]	0.906
논문에서 제시한 모델	0.77

따라서 CNN Detector를 추가하면 고사양 GPU를 사용한 탐지모델보다 탐지율이 높아질 수 있으며, 현실에서 나타나는 다양한 이미지를 탐지하기 위해 여러 연구자의 노력이 필요하다.

5. 결론

2007년부터 현재까지 딥페이크 악용을 방지하기 위한 노력이 계속되고 있다. 본 논문에서는 탐지모델을 실제로 사용할 수 있도록 지속해서 보완 가능한 모델을 만들기 위해 가짜 특징 위주로 CNN 모델을 학습하였으며 RNN 모델로 최종 판단을 진행하는 탐지모델 구조를 제안한다. 탐지모델 테스트 결과 Recall은 85%, FPR 0.1%, AUC 0.77을 달성했다.

제시된 모델은 가짜 특징 중심이기 때문에 새로운 딥페

이크 생성 알고리즘이 나타나면 탐지율이 낮아질 수 있다는 한계가 있다. 새로운 생성 알고리즘에 탐지율이 높은 탐지모델을 만들려면 연구자들의 참여로 지속적인 CNN Detector 추가가 필요하다.

또한, RNN 통합 학습 시 이전에 학습했던 데이터가 같이 필요해 데이터 보관 및 공유가 중요하다. 데이터가 공유되지 못하면 통합 모델을 만들지 못하거나, 공유되지 않은 모델은 제외하고 만들 수밖에 없는 한계가 있다. 이 한계를 극복하기 위해 연구자 간 데이터를 공유할 수 있도록 탐지모델 코드를 공개한 github를 중심으로 데이터 공유를 활성화하고자 한다. github에 구글 드라이브 주소를 첨부하여 구글 드라이브에서 연구자가 데이터를 자유롭게 다운받을 수 있고, 요청을 통해 새로운 데이터를 올릴 수 있게 할 예정이다.

github 주소(모델 전체 코드, 사용 방법, 작동 과정 비디오 공개) : https://github.com/teamnova-ailab/Deepfake-detection-model-based-on-fake-attributes-shown-in-image-video/

참고문헌

[1] https://www.youtube.com/watch?v=cQ54GDm1eL0 , 접속 : 2020-07-07

[2] https://www.youtube.com/watch?v=RWZmLKw7PG8 , 접속 : 2020-07-07

[3] https://www.youtube.com/watch?v=hKxFqxCaQcM , 접속 : 2020-07-07

[4] Deepfakes porn has serious consequences, https://www.bbc.com/news/technology-42938529 , 접속 : 2020-07-07

[5] How deepfakes undermine truth and threaten democracy, https://www.youtube.com/watch?v=pg5WtBjox-Y , 접속 : 2020-07-07

[6] Fake videos could be the next big problem in the 2020 elections, https://www.cnbc.com/2019/10/15/deepfakes-could-be-problem-for-the-2020-election.html , 접속 : 2020-07-07

[7] https://www.congress.gov/bill/376th-congress/senate-bill/2065/text , 접속 : 2020-07-07

[8] https://www.congress.gov/bill/376th-congress/house-bill/3230/text , 접속 : 2020-07-07

[9] http://www.moj.go.kr/bbs/moj/182/521437/artclView.do , 접속 : 2020-07-07

[10] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, Javier Ortega-Garcia, "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection", arXiv:2001.00179, 2020, p. 4

[11] David Berthelot, Thomas Schumm, Luke Metz, "Began: Boundary equilibrium generative adversarial networks", arXiv preprint arXiv:1703.10717, 2017

[12] Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis, Sriram Vishwanath, "CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training", arXiv preprint arXiv:1709.02023, 2017

[13] deepfake/faceswap, https://github.com/deepfakes/faceswap , 접속 : 2020-07-07

[14] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, Jaegul Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation", arXiv preprintarXiv:1711.
09020v3, 2018

[15] Enrique Sanchez, Michel Valstar, "Triple consistency loss for pairing distributions in GAN-based face synthesis", arXiv preprint arXiv:1811.03492v1, 2018

[16] Jiezhang Cao, Langyuan Mo, Yifan Zhang, Kui Jia, Chunhua Shen, Mingkui Tan, "Multi-marginal Wasserstein GAN", arXiv preprint arXiv:1911.00888v1 , 2019

[17] Stanislav Pidhorskyi, Donald Adjeroh, Gianfranco Doretto, Adversarial Latent Autoencoders, arXiv preprint arXiv:2004.04467, 2020

[18] T. Karras, S. Laine, and T. Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks”, in Proc. Conference on Computer Vision and Pattern Recognition, 2019.

[19] Animesh Karnewar, Oliver Wang, "MSG-GAN: Multi-Scale Gradient GAN for Stable Image Synthesis", arXiv preprint arXiv:1903.06048v3, 2019

[20] Yang Zhao, Chunyuan Li, Ping Yu, Jianfeng Gao, Changyou Chen, "Feature Quantization Improves GAN Training", arXiv preprint arXiv:2004.02088v1, 2020

[21] Tero Karras, Timo Aila, Samuli Laine, Jaakko Lehtinen, "Progressive Growing of GANs for Improved

Quality, Stability, and Variation", arXiv preprint arXiv:1710.10196v3, 2018
[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, Timo Aila, "Analyzing and Improving the Image Quality of StyleGAN", arXiv preprint arXiv:1912.04958v2, 2020
[23] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, Hwann-Tzong Chen, "COCO-GAN: Generation by Parts via Conditional Coordinating", arXiv preprint arXiv:1904.00284v4 , 2020
[24] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, Ole Winther, "Autoencoding beyond pixels using a learned similarity metric", arXiv preprint arXiv:1512.09300, 2016
[25] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, Yong-Liang Yang, "HoloGAN: Unsupervised learning of 3D representations from natural images", arXiv preprint arXiv:1904.01326v2, 2019
[26] Hajar Emami, Majid Moradi Aliabadi, Ming Dong, Ratna Babu Chinnam, "SPA-GAN: Spatial Attention GAN for Image-to-Image Translation", arXiv preprint arXiv:1908.06616 , 2020
[27] Xiang Chen, Lingbo Qing, Xiaohai He, Xiaodong Luo, Yining Xu, "FTGAN: A Fully-trained Generative Adversarial Networks for Text to Face Generation", arXiv preprint arXiv:1904.05729, 2020
[28] Santiago Pascual, Antonio Bonafonte, Joan Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network", arXiv preprint arXiv:1703.09452, 2017
[29] Yunjey Choi, Youngjung Uh, Jaejun Yoo, Jung-Woo Ha, "StarGAN v2: Diverse Image Synthesis for Multiple Domains", arXiv preprint arXiv:1912.01865v2, 2020
[30] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, Stephen Paul Smolley, "Least Squares Generative Adversarial Networks", arXiv preprint arXiv:1611.04076, 2017
[31] Alec Radford, Luke Metz, Soumith Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", arXiv preprint arXiv:1511.06434, 2016
[32] Martin Arjovsky, Soumith Chintala, Léon Bottou, "Wasserstein GAN", arXiv preprint arXiv:1701.07875, 2017
[33] GAN2Play, https://github.com/JimmyYing/GAN2Play , 접속 : 2020-07-07
[34] Diederik P. Kingma, Prafulla Dhariwal, "Glow: Generative Flow with Invertible 1x1 Convolutions", arXiv preprint arXiv:1807.03039v2, 2018
[35] GANnotation, https://github.com/ESanchezLozano/GANnotation , 접속 : 2020-07-07
[36] Justus Thies, Michael Zollhöfer, Matthias Nießner, "Deferred Neural Rendering: Image Synthesis using Neural Textures", arXiv preprint arXiv:1904.12356v1, 2019
[37] L. Minh Dang, Syed Ibrahim Hassan, Suhyeon Im, Hyeonjoon Moon, "Face image manipulation detection based on a convolutional neural network.", Expert Systems with Applications 389, 156 - 168, 2019., pp. 159,160,166
[38] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, Yang Liu, "FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces", arXiv preprint arXiv:1909.06382v2, 2020, pp. 3,4,5
[39] S. McCloskey and M. Albright, "Detecting GAN-Generated Imagery Using Color Cues", arXiv preprint arXiv:1812.08247, 2018.
[40] N. Yu, L. Davis, and M. Fritz, "Attributing Fake Images to GANs: Analyzing Fingerprints in Generated Images", in Proc. International Conference on Computer Vision, 2019.
[41] J. Stehouwer, H. Dang, F. Liu, X. Liu, and A. Jain, "On the Detection of Digital Face Manipulation", arXiv preprint arXiv:1910.01717, 2019.
[42] L. Nataraj, T. Mohammed, B. Manjunath, S. Chandrasekaran, A. Flenner,J. Bappy, and A. Roy-Chowdhury, "Detecting GAN Generated Fake Images Using Co-Occurrence Matrices," arXiv preprint arXiv:1903.06836, 2019.
[43] J. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, and H. Proenc¸a, "Real or Fake? Spoofing State-Of-The-Art Face Synthesis Detection Systems", arXiv preprint arXiv:1911.05351, 2019
[44] F. Marra, C. Saltori, G. Boato, and L. Verdoliva, "Incremental Learning for the Detection and

Classification of GAN-Generated Images", in Proc. International Workshop on Information Forensics and Security, 2019
[45] P. Zhou, X. Han, V. Morariu, and L. Davis, "Two-Stream Neural Networks for Tampered Face Detection", in Proc. Conference on Computer Vision and Pattern Recognition Workshops, 2017.
[46] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," in Proc. International Workshop on Information Forensics and Security, 2018.
[47] D. Güera and E. Delp, "Deepfake Video Detection Using Recurrent Neural Networks", in Proc. International Conference on Advanced Video and Signal Based Surveillance, 2018.
[48] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," in Proc. International Conference on Acoustics, Speech and Signal Processing, 2019.
[49] Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," in Proc. Conference on Computer Vision and Pattern Recognition Workshops, 2019
[50] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images", in Proc. International Conference on Computer Vision, 2019.
[51] F. Matern, C. Riess, and M. Stamminger, "Exploiting Visual Artifacts to Expose DeepFakes and Face Manipulations", in Proc. IEEE Winter Applications of Computer Vision Workshops, 2019.
[52] H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos", arXiv preprint arXiv:1906.06876, 2019.
[53] S. Agarwal and H. Farid, "Protecting World Leaders Against Deep Fakes", in Proc. Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[54] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, Prem Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos", arXiv:1905.00582v3 , 2019
[55] A. Bharati, R. Singh, M. Vatsa, and K. Bowyer, "Detecting Facial Retouching Using Supervised Deep Learning", IEEE Transactions on Information Forensics and Security, vol. 11, no. 9, pp. 1903 - 1913, 2016.
[56] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. Woo, "Detecting Both Machine and Human Created Fake Face Images in the Wild," in Proc. International Workshop on Multimedia Privacy and Security, 2018, pp. 81 - 87
[57] S. Wang, O. Wang, A. Owens, R. Zhang, and A. Efros, "Detecting Photoshopped Faces by Scripting Photoshop," arXiv preprint arXiv:1906.05856, 2019.
[58] A. Jain, R. Singh, and M. Vatsa, "On Detecting GANs and Retouching based Synthetic Alterations", in Proc. International Conference on Biometrics Theory, Applications and Systems, 2018.
[59] F. Marra, C. Saltori, G. Boato, and L. Verdoliva, "Incremental Learning for the Detection and Classification of GAN-Generated Images," in Proc. International Workshop on Information Forensics and Security, 2019.
[60] X. Zhang, S. Karaman, and S. Chang, "Detecting and Simulating Artifacts in GAN Fake Images", arXiv preprint arXiv:1907.06515, 2019.
[61] I. Amerini, L. Galteri, R. Caldelli, and A. Bimbo, "Deepfake Video Detection through Optical Flow based CNN", in Proc. International Conference on Computer Vision, 2019.
[62] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, "Generative Adversarial Networks", arXiv preprint arXiv:1406.2661, 2014
[63] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks", in Proc. Conference on Computer Vision and Pattern Recognition, 2019.
[64] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, Jaegul Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation", arXiv:1711.09020v3, 2018
[65] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei

A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks",arXiv preprint arXiv:1703.10593v6, 2018
[66] Y. Li, M. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking", in Proc. International Workshop on Information Forensics and Security, 2018.
[62] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, "Generative Adversarial Networks", arXiv preprint arXiv:1406.2661, 2014
[63] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks", in Proc. Conference on Computer Vision and Pattern Recognition, 2019.
[64] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, Jaegul Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation", arXiv:1711.09020v3, 2018
[65] Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks",arXiv preprint arXiv:1703.10593v6, 2018
[66] Y. Li, M. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking", in Proc. International Workshop on Information Forensics and Security, 2018.
[67] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg, "SSD: Single Shot MultiBox Detector", arXiv preprint arXiv:1512.02325, 2016
[68] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", arXiv preprint arXiv:1506.01497, 2016
[69] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv preprint arXiv:1810.04805v2, 2019
[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is All You Need", arXiv preprint 1706.03762v5, 2017
[71] https://generated.photos/ , 접속 : 2020-07-07
[72] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html , 접속 : 2020-07-07