Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.34, NO.5, Oct. 2024 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2024.34.5.885

딥페이크 탐지 모델의 검증 방법론 불일치에 따른 성능 편향 분석 연구*

김 현 준,[1†] 안 홍 은,[1] 박 래 현,[1] 권 태 경[2‡]
[1,2]연세대학교 (대학원생, 교수)

On the Performance Biases Arising from Inconsistencies in Evaluation Methodologies of Deepfake Detection Models*

Hyunjoon Kim,[1†] Hong Eun Ahn,[1] Leo Hyun Park,[1] Taekyoung Kwon[2‡]
[1,2]Yonsei University (Graduate student, Professor)

요 약

생성형 AI 기술이 정교해짐에 따라 빈번해지는 악의적 딥페이크 사용에 대응하기 위해 딥페이크 탐지 모델 연구가 활발히 진행되고 있다. 딥페이크 탐지 모델의 성능 평가는 학습 데이터셋 선택, 데이터셋 전처리, 학습 방법, 평가 데이터셋 선택 과정을 순차적으로 거친다. 하지만 기존 딥페이크 탐지 연구들은 각 단계마다 임의로 검증 방법론을 선택하여 논문에서의 성능이 표준화된 환경에서는 재현되지 않는 성능 평향 문제가 발생한다. 본 논문에서는 기존 딥페이크 탐지 연구의 검증 방법론을 분석하여 성능 평가의 신뢰성 저하 원인을 파악한다. 나아가 표준화된 환경에서의 실험을 통해 탐지 모델의 절대적 성능 비교에 어려움이 있음을 보여준다. 본 연구의 실험 결과는 탐지 성능 평가 신뢰성 제고와 절대적 성능 비교를 위해서는 통일된 검증 방법론이 필요함을 제시한다.

ABSTRACT

As deepfake technology advances, its increasing misuse has spurred extensive research into detection models. These models' performance evaluations, which include selecting train and test datasets, data preprocessing, and data augmentation, are often compromised by arbitrarily chosen validation methodologies in existing studies. This leads to biases under standardized conditions. This paper reviews these methodologies to pinpoint what diminishes evaluation reliability. Experiments in standardized environments reveal the difficulties in comparing performance absolutely. The findings highlighted the need for a consistent validation methodology to boost evaluation reliability and enable fair comparisons.

Keywords: Deepfake Detection, Evaluation Methodology, Performance Bias, Deepfake Dataset

I. 서 론

딥페이크 기술은 딥러닝 기술을 이용하여 사람이 눈으로 구분하기 어려울 정도로 현실적이고 자연스러운 가짜 얼굴을 생성한다. 딥페이크 기술이 오픈소스, 어플리케이션 또는 서비스로 제공되어 누구나 쉽게 사용할 수 있다. 하지만 이런 기술은 가짜뉴스, 음란물, 피싱 등 사회적으로 악용 가능성이 있다. 이에 딥

Received(05. 14. 2024), Modified(07. 24. 2024), Accepted(08. 27. 2024)
* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학 ICT연구센터사업의 연구결과로 수행되었음(RS- 2024-00436936)
* 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. RS-2024-00439762, AI 모델 취약성분석, 평가 기술 생성정보 비밀성 판단 도구 개발)
† 주저자, hjkimc@yonsei.ac.kr
‡ 교신저자, taekyoung@yonsei.ac.kr(Corresponding author)

페이크 기술로 생성한 가짜 영상 및 이미지를 탐지하기 위한 다양한 연구가 진행되고 있다. 특히 합성곱 신경망(CNN)을 이용하여 이미지 속에서 인위적인 조작 흔적인 아티팩트를 검출하는 딥페이크 탐지모델이 활발히 연구되고 있다.

많은 딥페이크 탐지 모델은 성능 평가를 위해 학습 및 평가 데이터셋 선택, 데이터 전처리, 학습, 평가 과정을 거친다. 하지만 데이터셋 선택, 전처리 방법, 학습 방법에 대한 통일된 기준이 존재하지 않아, 탐지 모델 검증시 임의의 검증 방법론을 설정하여 검증을 수행한다. 저자의 코드가 공개되지 않은 경우, 논문의 내용을 바탕으로 모델 구조와 하이퍼파라미터를 그대로 구현하여 실험을 진행하게 된다. 하지만 모델 성능 측정시 모델 성능이 상이하는 것을 확인할 수 있다. 이런 변화가 모델 자체의 성능 차이인지, 학습 방법 및 성능 검증 방법에 따라 성능 차이가 발생하는지 명확히 구분하기 어렵게 만든다. 이로인해 저자가 주장한 탐지모델 성능의 신뢰성을 떨어트리고 절대적 성능 비교에 어려움을 초래한다. 이런 문제를 해결하기 위해 표준화된 검증 기준을 설정해야 한다. 표준화된 검증 기준에서의 탐지 모델 평가를 진행해야 객관적이고 절대적인 성능 비교가 가능하며, 각 탐지 모델의 장단점을 명확하게 파악할 수 있다.

본 논문은 통일되지 않은 검증 방법론에 대한 편향 문제를 실험을 통해 보여준다. 이를 위해 2장에서 연구 배경에 대한 간략한 설명을 시작으로 3장에서는 현재 딥페이크 탐지모델 성능 평가 시나리오를 분석한다. 4장에서는 시나리오에 따른 발생가능한 문제점이 무엇인지 설명하고 5장에서 실험을 통해 검증 방법 변화에 따른 성능 편향을 확인한다. 6장에서 딥페이크 탐지 검증을 위한 추후 연구 방향을 제안한다.

II. 배 경

2.1 딥페이크 탐지 모델

딥페이크 탐지 모델은 크게 세가지 방법으로 나눌 수 있다.

1) 분류 탐지기(Naive): CNN을 단순 분류기로 사용하여 딥페이크 여부를 판단하는 방법이다. CNN 기반 이진 분류 탐지기에는 MesoNet, Xception[6] 등이 있다. 이는 팀페이크 탐지기 연구의 초창기 때 많이 연구된 방법이며, 현재는 딥페이크 탐지 모델의 backbone 모델로 많이 사용된다.

2) 공간 기반 탐지기(Spatial): Fig. 1. (a)에서 아티팩트의 대표적인 특징 4가지를 보여준다. 왼쪽부터 landmark 불일치, 합성 경계 표식, 색깔 불일치, 화질 불일치이다. 공간 기반 탐지방법은 CNN을 backbone으로 이용하여 Fig. 1. (a)와 같이 합성한 이미지 내에서 합성 및 조작하는 과정에서 발생하는 미묘한 아티팩트를 탐지하는 방법이다. 혹은 얼굴 모습과 다른 특정 표현을 학습하여 탐지하는 방법이 있다. 전자의 경우 FaceXray[1], UCF[20] 등이 있고 후자인 경우 StyleGRU[10], RECCE [2] 등이 있다.

3) 주파수 영역 기반 탐지기(Frequency): Fig. 1. (b)처럼 합성한 가짜 이미지를 푸리에 변환 등을 이용하여 주파수 영역으로 변환하고 이상 주파수 신호(특징점) 및 아티팩트를 찾아 판별하는 방법이다. 대표 모델로는 SPSL[3], SRM[4], F3Net [5] 등이 있다.

Naive 탐지기를 단순 이진 분류기로 사용했을 때는 다른 모델에 비해 성능이 현저히 낮게 나오지만 딥페이크 데이터셋을 학습시키는 것이 아닌 인위적인 합성 흔적인 남는 이미지를 만들어 학습시키는 방법(BI[1] 등)을 통해 다른 종류의 딥페이크 탐지기와

Fig. 1. (a) shows four types of artifacts. (b) shows the process of converting the image to the frequency domain.

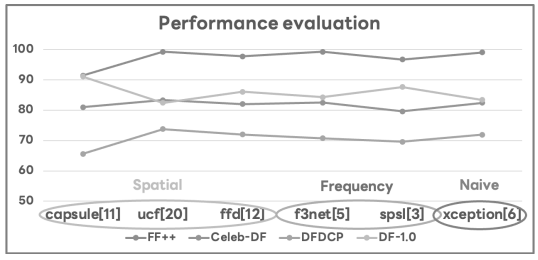

Fig. 2. The result of training three types of deepfake detection models with FF++ and testing their performance on the FF++, Celeb-DF, DFDC, Deeper1.0. The experiments were conducted in a standardized experi-mental environment as claimed in the study.

유사한 성능을 낼 수 있다. Fig. 2.는 3종류의 6가지의 탐지모델을 FaceForensics++(FF++) [6] 데이터셋으로 학습하여 FF++[8], Celeb-DF (CDF)[8], DFDC[7], DeeperForensics-1.0 (DF-1.0)[9] 데이터셋으로 테스트하여 AUC 측정한 결과이다. 이처럼 일반적으로 FF++로 학습한 탐지기는 FF++[6]에서 95%, CDF[8]와 DF-1.0 [9]에서 80%, DFDC[7]에서 70% 내외 성능을 보인다. 이를 통해 딥페이크 탐지 모델은 데이터셋에 의존적임을 유추할 수 있다.

2.2 딥페이크 데이터셋

딥페이크 데이터셋은 제작한 시기를 기준으로 두가지로 나눌 수 있다.

1) 이전 세대 데이터셋 : 딥페이크 생성 기법을 이용하여 무차별적인 생성을 하여 구성한 딥페이크 데이터셋이다. 후처리 작업을 하지 않아 데이터셋을 구성하는 영상/이미지 품질이 고르지 못하는 문제점이 있다. FaceForensics++[6], DFDC[7], Celeb-DF[8] 등이 있다.

FF++의 데이터셋인 경우 원본 영상 1000개와 이를 딥페이크 생성기법으로 합성하여 생성기법 마다 1000개의 딥페이크 영상으로 구성되어 있다. Celeb-DF[8]는 자연스러운 합성 이미지를 위해 행동이 절제된 유명인들의 인터뷰 영상을 이용하여 총 5639개의 영상으로 구성되어 있다. DFDC[7]는 다양한 인종, 성별, 나이의 사람들의 영상 128,154개로 구성되어 있다.

2) 최신 데이터셋 : 실생활 악용되는 딥페이크와 유사한 고품질 딥페이크 영상을 구성하기 위해 제작되었으며, 딥페이크 생성 기법을 통해 생성한 뒤 후처리를 통해 품질이 좋지 못한 영상을 제거하여 데이터셋 품질을 최대한 고르게 데이터셋을 구성하였다. DeeperForensics-1.0[9], DeepfakeDetection (DFD)[6], WildDeepfake(WDF) [13] 등이 있다. Deeper1.0[9]은 다양한 영상 품질과 압축 수준, 왜곡이 포함된 영상 50,000개로 구성되어 있다. DFD[6]는 다양한 시나리오에서 연기한 영상 3,363개, WDF[13]는 실제 환경을 반영하기 위해서 실제 온라인 플랫폼에서 낮은 해상도, 복잡한 배경 등을 포함한 영상 7,314개로 구성되어 있다.

III. 딥페이크 탐지 성능 평가 시나리오 분석

3.1 데이터셋 및 데이터셋 전처리

딥페이크 탐지 모델은 일반적으로 FF++을 학습 데이터셋으로, Celeb-DF(CDF)[8], DFDC[7]을 평가 데이터셋으로 많이 사용한다(Table 1.). 하지만 FF++[6]은 동일한 영상에 대해 압축률에 따라 raw, c23, c40 세가지 품질이 존재한다. CDF[8], DFDC[7]는 초창기 버전과 업데이트 버전으로, 업데이트 버전은 다양한 생성기법과 더 많은 영상으로 구성되어 있다.

대부분 탐지 모델은 프레임 단위로 탐지를 진행하

Table 1. Describes the training and testing dataset used for performance verification in each deepfake detection model. CDF: Celeb-DF[8], DF-1.0: DeeperForensics-1.0[9], DFD: DeepfakeDetection[6], Fsh: FaceShifter[14], WDF: WildDeepfake[13].

Model	Train Dataset	Test Dataset						
	FF++	CDF	DFDC	DF-1.0	DFD	FSh	WDF	etc
AltFreezing[15]	c23	v2	O	O	X	O	X	
OST[16]	c23	v2	O	O	O	X	X	-
UIA-ViT[17]	c23	v1 v2	P	X	O	X	X	-
SLADD[18]	c23	v2	O	O	X	X	X	-
RECCE[2]	c40	v1	O	X	X	X	O	-
SPSL[3]	c40	v1	O	X	X	X	X	-
F3Net[5]	c40	X	X	X	X	X	X	FF++
LAA-Net[19]	U	v2	O	X	O	X	O	-
StyleGRU[10]	C23	v2	X	O	O	O	X	-

기 때문에 전처리 과정을 통해 영상으로 구성된 데이터셋을 프레임 단위로 분할한다. 이때 다양한 얼굴 탐지 기법을 통해 원본 영상에서 찾은 얼굴 영역 주변으로 특정 크기(224*224, 299*299 등)을 잘라 이미지를 재구성한다. 얼굴이 탐지된 프레임의 전체를 사용하는 탐지 모델과, 그중 일부 프레임을 사용하는 탐지 모델이 있다.

3.2 데이터 증강 및 학습

딥러닝 학습과정에서 데이터 증강(data augmenation) 기법을 사용하여 학습하는 데이터셋 수를 늘리고 성능을 향상시킬 수 있다. 이런한 기법으로는 이미지를 수평축으로 무작위로 뒤집어 모델이 좌우 방향에 의존하지 않고 중요한 특성을 학습할 수 있도록 하는 Random Horizontal Flip(RHF), 이미지에서 무작위로 선택한 작은 영역을 제거하여 일부가 가려져 있거나 누락되었을 때도 중요한 특징을 인식하고 객체를 식별하는 능력을 향상시키는 Random Cutout(RCO), 이미지에 가우시안 노이즈를 추가하여 노이즈가 포함된 데이터에서도 정보를 잘 추출할 수 있도록 하는 Add Gaussian Noise (AGN)등이 있다. Table 2.를 참고하면, 대부분의 탐지 모델은 RHF 방식을 사용하지만 그 외 데이터 증강 기법은 저자의 임의의 선택으로 이용되고 사용한 이유에 대한 명확한 근거는 제시되지 않는다.

3.3 교차 데이터셋 평가

교차 데이터셋 평가는 학습한 데이터셋 외의 다른 데이터셋으로 평가하는 일반적인 성능 평가 방법을 말한다. 학습한 데이터셋으로 평가하는 내부 데이터셋 평가 방식하고는 반대되는 검증 방법이다. 현재 딥페이크 탐지 모델은 학습한 데이터셋, 생성기법 외에는 성능이 하락하는 한계점이 존재한다. 이를 극복하여 모든 데이터셋에 대해 일관된 성능을 보일 수 있도록 해야한다. 교차 데이터셋 검증 방법은 탐지 모델의 일반화 성능을 검증하는 중요한 평가 시나리오다.

IV. 기존 딥페이크 탐지 검증 방법론의 문제점

4.1 데이터셋 전처리의 문제점

얼굴 추출 프레임. 성능이 제각각인 얼굴 탐지 모델을 이용하여 영상 데이터셋에서 얼굴이 탐지된 프레임을 추출하는 과정을 거치면, 누락된 얼굴 프레임이 발생하게 되고 이 또한 얼굴 탐지 모델마다 동일하지 않다. Table 2.의 2열을 참조하면, 각 탐지 모

Table 2. Summarizes the implement details of each detection model, including 'face extraction model', 'image size of frames', 'quality of FF++', 'the number of frames', 'data augmentation', and 'metrics(AUC, ACC etc)'.

Model	Face Extraction	Image Size	FF++ Quality	#Frames	Data Augmentation	Metrics		
						AUC	ACC	etc
AltFreezing [15]	MTCNN	224*224	c23	32	RHF, RCO, AGN,	O	X	X
OST[16]	DLIB	256*256	c23	All	Online meta dataset	O	O	EER
UIA-ViT[17]	DLIB	224*224	c23	All	16*16 patch	O	O	X
SLADD[18]	DLIB	256*256	c23, c40	All	-	O	X	X
RECCE[2]	RetinaFace	299*299	c40	-	RHF	O	O	X
SPSL[3]	-	-	c23, c40	-	-	O	O	X
F3Net[5]	-	299*299	All	-	Mixblock	O	O	X
FaceXray[1]	Cascade	256*256	c23, c40	32	Blend image	O	X	EER
LAA-Net[19]	RetinaFace	384*384	Unknown	128	RHF, random Scaling, erasing...	O	X	AP, mF1
StyleGRU[10]	RetinaFace	256*256	c23	All, 32	RCO	O	X	X

델마다 사용하는 얼굴 탐지모델이 다른 것을 확인 할 수 있으며, 이는 추출되는 데이터셋 프레임이 달려져 딥페이크 탐지 성능에 영향을 줄 수 있음을 의미한다.

사용 프레임 수. 프레임 선택 개수 및 방법에 의해 딥페이크 탐지 모델의 학습 및 평가 데이터가 달라진다. Table 2.의 5열을 참조하면 일부 프레임을 사용하는 경우, 첫 번째 프레임에서 순차적으로 특정 개수를 선정하거나, 랜덤으로 특정 개수를 추출한다. 이는 학습 및 평가 데이터셋 변화로 인한 딥페이크 성능의 불일관성을 초래한다.

4.2 데이터 증강 방법 및 학습 데이터셋의 문제점

데이터 증강 방법. Table 2.을 참조하면, 딥페이크 탐지 모델마다 데이터 증강 기법 사용 여부나, 사용 시에도 기법 종류 및 조합이 일관적이지 않다. 이는 탐지 기법이나 학습 기법에 의한 성능 차이가 아닌 학습에 사용된 증강 데이터셋에 의한 성능 차이를 유발한다.

학습 데이터셋 품질. 탐지 모델의 대부분은 FF++ 데이터셋을 사용하여 학습한다. 하지만 학습 데이터셋의 품질은 Table 2.에서 볼 수 있듯이 동일 데이터셋의, 각기 다른 품질을 이용한다. 원본 딥페이크 영상을 각각 23%, 40% 압축해서 만든 저품질의 c23, c40 데이터셋을 구성한다. SPSL[3] 에 따르면, 압축률이 높아 더 저품질을 보이는 c40은 c23보다 주파수 성분이 많이 감소한다. 이처럼 데이터셋 품질에 따라 아티팩트 정도가 달라지고, 이는 학습 성능에 영향을 미치는 것을 알 수 있다.

4.3 평가 데이터셋의 문제점

평가 데이터셋. 동일한 품질의 데이터셋으로 학습을 진행하였더라도 논문에서는 잘 나온 결과의 테스트 데이터셋에서의 성능을 보여준다. 또한 Table 1.을 참조하면, 각 탐지 모델의 평가 데이터셋이 일관되지 않음을 확인 할 수 있다. 이처럼 일관적이지 않은 환경에서 평가된 성능은 직접적인 비교가 어렵다. 논문에서 평가되지 않은 데이터셋에 대해 성능 평가를 진행하기 위해서는 이전 연구를 재구성하여 실험을 진행해야 하지만, 논문에서 제시한 성능을 재현하는데 어려움이 발생한다.

V. 실험을 통한 편향 문제 확인

본 논문에서는 RECCE[2], SRM[4], UCF[20], SPSL[3] 모델을 동일한 c23 품질의 FF++데이터셋[6]으로 학습하고 CDF[8], DFDC[7], DFD[6] 데이터셋으로 평가한다. 이를 통해 학습 및 평가 데이터셋, 데이터 전처리 등에 따라 성능이 어떻게 변하는지 실험을 통해 확인하고 기존 논문과 성능을 비교한다.

5.1 평가 데이터셋 및 학습 시 마다 성능 상이

평가(대상)셋 프레임. 평가 데이터셋 프레임은 크게 얼굴 추출 모델의 성능 차이로 인해 추출된 프레임이 달라지는 경우와, 추출된 프레임 데이터셋에서 평가 데이터셋의 개수를 선정하는 과정에서 프레임수가 변화게 된다.

(1) 얼굴 추출 탐지 모델에 의한 프레임의 수 변화. FaceX-ray[1] 모델은 Cascade 얼굴 탐지 모델을 사용하여 영상 속 얼굴 탐지된 프레임을 추출한다. 하지만 이 모델은 다른 MTCNN, DLIB, RetinaFace에 비해 성능이 떨어져 얼굴 프레임을 덜 추출한다. 이 때, 추출되지 않은 프레임은 딥페이크 탐지가 어려운 프레임이고, 이를 제외하여 성능 평가를 진행하면 성능이 상승한다. 이는 Table 3.에서 확인 할 수 있다. Face X-ray[1]의 전처리 단계에서 DLIB을 통해 보다 많은 프레임을 추출하여 대상 프레임을 다르게 평가했을 때, 내부 데이터셋 실험에서는 AUC가 98.71(공식 논문)에서 90.96(본 논문)으로 하락하는 것을 확인할 수 있다. 또한 교차 데이터셋 모두에서 현저한 성능 저하를 확인 할 수 있다.

Table 3. FaceXray model evaluation in intra-dataset and cross dataset. Comparison between the official paper, the LAA-Net paper, and the experiment in which we verified the performance.

Data set	Test Dataset	Official paper	LAA-Net paper	Ours
Intra	FF++ (c23)	98.71	**99.92**	90.96
Cross	CDF	**95.40**	79.5	57.65
	DFDC	**80.92**	65.5	56.0
	DFD	80.58	**95.40**	53.73

(2) 임의의 프레임수 선정. Table 4.는 랜덤으로 전체 프레임의 3분의 1의 딥페이크 데이터를 선택해서 fake confidnce을 측정한 결과이다. 3번 반복 실험을 하였을 때 평균 fake confidecne이 상이하는 것을 확인할 수 있다. Xception모델[6]로 DF-1.0[9] 데이터셋을 측정한 경우,최대 24.37% 차이를 보이는 것을 확인할 수 있다. 이런 수치는 모델의 성능의 큰차이를 보여주며, 모델간 성능 비교시 순위를 변동시키고 절대적인 성능 비교 분석에 어려움을 준다. 이런 랜덤적인 성능개선 요소를 줄이기 위해 반복실험을 통해 평균 값을 성능으로 제시하거나, 데이터셋의 전체 프레임을 사용하여 랜덤 요소를 제거하는 방법으로 성능 평가에 신뢰성을 높일 수 있다.

학습마다 성능 상이. 공식 모델 코드가 비공개인 경우 모델을 비교하기 위해 논문과 동일한 모델구조, 학습방법, 하이퍼파라미터 등 동일한 환경을 구성하고 실험을 하였음에도 모델의 성능이 다르게 평가가 될 수 있음을 Table 3.를 통해 보여준다. Table 3.을 참조하면, LAA-Net[19] 논문에서 FaceX-ray[1]의 재현 실험 결과, 내부 데이터셋에서 공식 논문과 재현한 코드의 성능 차이보다 교차 데이터셋의 DFDC[7], CDF[8]에서의 성능 차이가 현저하게 떨어짐을 확인할 수 있으며 이를 LAA- Net[19] 모델과 비교했을 때 절대적 성능 비교로 볼 수 없다. 모델 코드 및 모델 파라미터는 공개를 해야 추후 연구에서 모델 성능 비교가 용이하고 새로운 데이터셋에 대한 성능 비교가 가능하다.

Table 4. Using the SPSL, UCF, and Xception detection models trained on the FF++ dataset, we conducted three repeated experiments, measuring the fake confidence on one-third of randomly selected frames from the DF-1.0 and Celeb-DF datsets.

model	Dataset	test1	test2	test3
SPSL	DF-1.0	**84.4**	**78.17**	82.15
	Celeb-DF	**85.52**	**94.82**	90.80
UCF	DF-1.0	33.33	**26.20**	**37.54**
	Celeb-DF	**95.09**	97.5	**99.36**
Xception	DF-1.0	65.42	**66.29**	**53.30**
	Celeb-DF	**94.96**	92.15	**92.05**

5.2 통일된 환경 설정 및 성능 비교

본 논문에서 실험 결과의 일관성과 탐지 성능 비교의 공정성을 높이기 위해 다음과 같이 통일된 실험환경을 설정한다.

1. 얼굴 추출 모델로는 많이 사용되는 DLIB의 HOG 방식을 이용하여 데이터셋에서 프레임을 추출한다.
2. 선택하는 프레임에 따라 성능 변화가 생기므로, 추출된 데이터셋의 전체 프레임을 이용하여 성능 평가를 진행한다.
3. 데이터 증강 방법은 성능 변화를 야기하므로 순수 딥페이크 탐지 모델의 성능을 측정하기 위해 데이터증강 방법은 사용하지 않고 학습을 진행한다.
4. 학습 데이터셋 외의 데이터셋으로 평가하는 Cross-dataset 검증을 통해 각 데이터셋 내에서 탐지모델의 순위를 파악하고 데이터셋 간의 성능 순위를 파악하고 일반화 성능도 평가한다.

Table 5.에서는 탐지 모델 논문에서 주장하는 각 모델의 성능과, 통일된 환경을 구축하여 평가한 성능을 비교한다. 논문 저자의 주장대로 성능을 비교하게 되면 UCF[20]가 CDF[8] 데이터셋에서 가장 높은 탐지 성능을 보여준다. 하지만 통일된 환경에서 실험을 진행했을 때, RECCE[2], SPSL[3], UCF[20] 논문 주장과 동일한 성능을 보여준다. 하지만 SRM 모델인 경우 논문과 달리 더 좋은 성능을 보이며, CDF[8] 데이터셋에서 가장 높은 탐지 성능을 보여준다. DFDC[7] 데이터셋에서는 성능이 논문 주장과 달리 RECCE[2]보다 낮은 성능을 보여준다.

최신 탐지 모델의 성능 보고를 위해 평가 데이터셋

Table 5. Compare and analyze the performance presetnted in the paper and the performance of the reconstructed detector in cross-dataset. The second row is the performance claimed in the paper, and the third row is the result when we reproduced the model.

	Model	Dataset		
		CDF	DFDC	DFD
paper	RECCE[2]	68.71	**69.06**	-
	SPSL[3]	76.88	66.16	-
	UCF[20]	**82.4**	80.5	-
	SRM[4]	**79.4**	**79.7**	91.9
reconstruction	RECCE[2]	73.19	**71.33**	81.19
	SPSL[3]	79.75	65.60	84.75
	UCF[20]	**84.08**	71.91	80.74
	SRM[4]	**84.36**	**68.98**	87.25

을 임의로 선정한다. 따라서 추가 실험 없이는 성능 비교에 어려움이 존재한다. Table 5.를 참조하면, SRM[4] 논문만이 DFD[6] 데이터셋에 대해 성능 평가를 진행하여 다른 모델과의 성능 비교가 어렵다. 이에 본 논문에서는 통일된 검증 환경을 구축하여 성능 평가를 진행하였다. 이때 SRM[4], SPSL[3], RECCE[2] 순으로 높은 탐지 성능을 보여준다.

VI. 결 론

본 논문에서는 실험을 통해 각 탐지 기법 논문의 저자가 주장하는 성능과 본 연구에서 재현한 성능의 차이가 오차범위를 벗어난 결과를 보여주는 것과, 다양한 데이터셋에서의 성능 순위 변화로 절대적 탐지 모델의 순위 평가의 어려움을 보여준다. 이런 문제는 통일 되지 않은 데이터셋 전처리, 학습 및 평가 데이터셋의 차이로 발생함을 실험을 통해 확인했다. 모델 간의 절대적이고 정확한 성능 평가를 위해 명확하게 통일된 기준 마련과 벤치마크가 필요하다. 선택적인 데이터셋을 대상으로 하는 성능 공개가 아닌 평가 데이터셋으로 많이 사용되는 CDF[8], DFDC[7], DF-1.0[9], DFD[6] 등을 반복 실험하여 보다 객관적인 성능을 평가하고 공개해야 한다. 통일된 성능 평가 환경을 통해 추후 탐지 모델의 성능 평가 및 비교를 위한 추가 실험 없이 각 탐지 모델의 장단점을 파악하기 용이해진다.

References

[1] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," Proceedings of the IEEE /CVF conference on computer vision and pattern recognition, pp. 5001-5010, Aug. 2020.

[2] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 18710-18719, Sept. 2022.

[3] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, Hui. Xue, W. Zhange, and N. Yu, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 772-781, Nov. 2021.

[4] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with highfrequency features," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16317-16326, Nov. 2021.

[5] Y. Qian, G. Yin, Z. Chen, and J. Shao, "Thinking in frequency: Face forgerydetection by mining frequency-aware clues," European conference on computer vision. Cham: Springer International Publishing, pp. 86-103, Aug. 2020.

[6] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "Faceforensics++: Learning to detect manipulated facial images," Proceedings of the IEEE/CVF international conference on computer vision, pp. 1-11, Feb. 2019.

[7] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Ferrer, "The deepfake detection challenge dataset," arXiv: 2006.07397, 2020.

[8] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3207-3216, Aug. 2020.

[9] L. Jiang, R. Li, W. Wu, C. Qian, and C. Loy "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition, pp. 2889-2898, Aug. 2020.

〔10〕 J. Choi, T. Kim, Y. Jeong, S. Baek, and J. Choi, "Exploiting Style Latent Flows for Generalizing Deepfake Video Detection," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1133-1143, Sept. 2024.

〔11〕 H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos," IEEE international conference on acoustics, speech and signal processing, pp. 2307-2311, 2019.

〔12〕 H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain, "On the Detection of Digital Face Manipulation," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5781-5790, Oct. 2020.

〔13〕 B. Zi, M. Chang, J. Chen, X. Ma, and Yu. Jiang, "Wilddeepfake: A challenging real-world dataset for deepfake detection," Proceedings of the 28th ACM international conference on multimedia, pp. 2382-2390, Oct. 2020.

〔14〕 L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5074-5083, Aug. 2020.

〔15〕 Z. Wang, J. Bao, W. Zhou, W. Wang, and H. Li, "Altfreezing for more general video face forgery detection," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4129-4138, Aug. 2023.

〔16〕 L. Chen, Y. Zhang, Y. Song, J. Wang, and L. Liu, "Ost: Improving generalization of deepfake detection via one-shot test-time training," Advances in Neural Information Processing Systems, 35, pp. 24597- 24610. Apr. 2022.

〔17〕 W. Zhuang, Q. Chu, Z. Tan, Q. Liu, H. Yuan, C. Miao, Z. Luo, and N. Yu,"UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection," European conference on computer vision. Cham: Springer Nature Switzerland, pp. 391-407, Oct. 2022.

〔18〕 L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 18710-18719, Sept. 2022.

〔19〕 D. Nguyen, N. Mejri, I. Singh, P. Kuleshova, M. Astrid, A. Kacem, E. Ghorbel, and D. Aouada, "LAA-Net: Localized Artifact Attention Network for High-Quality Deepfakes Detection," arXiv preprint arXiv:2401.13856, 2024.

〔20〕 Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "Ucf: Uncovering common features for generalizable deepfake detection," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22412-22423, Jan. 2024.

〈저자소개〉

김 현 준 (Hyunjoon Kim) 학생회원
2023년 2월: 건국대학교 전기전자공학부 졸업
2022년 1월~6월: 한국생산기술연구원 연구생
2023년 3월~현재: 연세대학교 인공지능학과 석사과정
〈관심분야〉 Deepfake, Generative AI, Vehicle Control 등

안 홍 은 (Hong Eun Ahn) 학생회원
2023년 2월: 이화여자대학교 사이버보안 졸업
2023년 3월~현재: 연세대학교 정보대학원 석사과정
〈관심분야〉 LLM Jailbreak, Membership Inference, Adversarial Machine Learning 등

박 래 현 (Leo Hyun Park) 학생회원
2017년 2월: 광운대학교 컴퓨터공학 졸업
2017월 3월~현재: 연세대학교 정보대학원 석박사통합과정
〈관심분야〉 Generative AI, Adversarial Machine Learning, Deepfake, 딥러닝 모델 검증 등

권 태 경 (Taekyoung Kwon) 종신회원
1992년: 연세대학교 컴퓨터과학과 학사
1999년: 연세대학교 컴퓨터과학과 박사
1999년~2000년: U.C, Berkeley, EECS, Post-Doc.
2001년~2013년: 세종대학교 컴퓨터공학과 교수
2007년~2008년: Univ. of Maryland, College Park 교환 교수
2013년~현재: 연세대학교 정보대학원 교수
〈관심분야〉 암호 프로토콜, 시스템 보안, 인공지능 보안, 메타버스 보안.