Journal of KIIT. Vol. 24, No. 1, pp. 1-13, Jan. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.1.1

ER-FusNet: RGB - 주파수 이중 스트림의 MLP 후기 융합을 이용한 딥페이크 탐지

하유진*[1], 박수진*[2], 박종찬*[3], 김건우**

ER-FusNet: Two-Stream RGB - Frequency with MLP Late Fusion for Deepfake Detection

Yu-Jin Ha*[1], Su-Jin Park*[2], Jong-Chan Park*[3], and Gun-Woo Kim**

본 논문은 교육부와 경상남도의 재원으로 지원을 받아 수행된 경상남도 지역혁신중심 대학지원체계(RISE) 연구결과 및 2025년도 산업통상자원부 및 한국산업기술기획평가원(KEIT)의 연구비 지원(RS-2025-02633048)에 의한 연구결과임

요 약

최근 딥페이크 고도화에 따른 탐지 모델의 일반화 성능 확보를 위해, 본 연구는 RGB와 DCT 기반 주파수 단서를 교차 어텐션으로 융합하는 이중 스트림 모델 ER-FusNet을 제안한다. 본 모델은 전역 시각 정보를 추출하는 EfficientNet과 주파수 영역의 미세 변조 흔적을 분석하는 RepLKNet을 결합하여 특징 간 상보적 시너지를 극대화한다. 교차 데이터셋 실험 결과, Celeb-DF v2 등에서 0.92~0.99의 F1 스코어를 달성하였으며, 실제 환경인 WildDeepfake에서는 단일 스트림 대비 F1을 0.53에서 0.65로 약 12% 향상시키며 범용적 탐지 역량을 입증하였다. 이는 공간·주파수 도메인의 적응적 융합이 미학습 데이터에 대한 강건성 확보의 주요 기제임을 시사하며, 진화하는 딥페이크 위협에 대응하여 탐지 신뢰성을 높이는 데 기여할 수 있음을 보여준다.

Abstract

To address generalization degradation from domain shifts in deepfake detection, we propose ER-FusNet, a dual-stream model that fuses RGB and DCT-based frequency cues via cross-attention. ER-FusNet couples an RGB EfficientNet with a DCT-based RepLKNet and, in cross-dataset tests, achieves F1 scores of 0.92 - 0.99 on benchmarks like Celeb-DF v2 while raising WildDeepfake F1 from 0.53 to 0.65 (~12% over the single-stream average). These results show that jointly leveraging global RGB signals and fine-grained frequency artifacts yields robust real-world deepfake detection.

Keywords

deepfake detection, dual-stream architecture, frequency domain, late fusion, cross-dataset generalization

* 경상국립대학교 AI융합공학부
- ORCID1:https://orcid.org/0009-0004-8963-1913
- ORCID2:https://orcid.org/0009-0008-2573-3901
- ORCID3:https://orcid.org/0000-0002-1886-534X
** 경상국립대학교 컴퓨터공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0001-5643-4797

· Received: Sep. 16, 2025, Revised: Sep. 30, 2025, Accepted: Oct. 03, 2025
· Corresponding Author: Gun-Woo Kim
Dept. of ComputerScience and Engineering, College of IT Engineering, Gyeongsang National University, JinJu, Korea
Tel.: +82-55-772-3323, Email: gunwoo.kim@gnu.ac.kr

Ⅰ. 서 론

최근 인공지능의 발전으로 딥러닝 기반 생성 모델의 성능이 고도화되었다. 이는 얼굴·신체·음성을 정교하게 합성·조작하는 딥페이크 기술의 고도화로 이어졌다. 특히 대규모 데이터로 학습된 모델의 생성 품질이 크게 향상되면서, 단일 이미지나 짧은 영상만으로도 신원 이전, 표정 및 입 모양 동기화 등 다양한 조작을 낮은 비용으로 수행할 수 있게 되었다[1]-[5].

딥페이크 기술은 여러 분야에서 긍정적으로 활용될 수 있다. 교육 분야에서는 역사적 인물을 생생하게 재현하여 학생들에게 높은 몰입감의 학습 경험을 제공할 수 있다. 의료 분야에서는 실어증이나 발음 장애 환자의 목소리 복원 등 개인맞춤형 치료에 기여한다[6]. 영화·게임 분야에서는 배우의 외적 나이를 조절하거나 고인이 된 배우를 복원하는 시각효과에 활용된다[7]. 마케팅·광고 분야에서는 고객 맞춤형 콘텐츠를 제작하고 기존 영상·이미지를 재활용하여 비용 절감에 효과적이다. 이처럼 딥페이크 기술은 다양한 산업에서 새로운 가치를 창출하고 있다.

반면 부정적 측면도 뚜렷하다. 딥페이크는 가짜 뉴스 생성을 통해 허위 정보를 유포하거나 특정인의 명예를 훼손하는 데 악용될 수 있다. 또한 합성 영상·이미지를 이용한 정치적 선동, 동의 없는 음란물 제작, 보이스피싱, 생체인증 우회 등 범죄 사례가 증가하고 있다[8]-[13]. 이러한 악용은 개인의 인권을 침해할 뿐만 아니라 사회 전반의 신뢰를 훼손하고 디지털 안보에 심각한 위협을 초래한다[14][15]. 이에 따라 법·정책적 규제의 마련과 함께, 딥페이크 콘텐츠를 자동 탐지·방어하는 기술 연구의 필요성이 대두되고 있다[14][15].

기존 딥페이크 탐지 기술은 정상 콘텐츠와 조작 콘텐츠를 구분하는 데 초점을 맞춰, 픽셀·텍스처 차원의 미세 왜곡, 인물의 비 자연스러운 움직임, 시공간적 불일치, 영상·음성 간 동기화 오류 등을 단서로 활용해 왔다[16]-[18]. 그러나 특정 데이터셋에 대한 과적합 경향이 있어, 압축률·해상도·촬영 환경이 다른 실제 데이터에서는 일반화 성능이 저하되는 한계가 있다.

본 연구는 이러한 일반화 취약성을 개선하고자, RGB와 주파수 정보를 결합하는 이중 스트림 후기 융합 방식을 제안한다. 제안 모델의 목표는 RGB와 주파수 두 도메인의 상보적 특징을 활용해 다양한 딥페이크 생성 기법과 실제 환경에서도 안정적으로 위조 영상을 탐지하는 것이다. 제안 모델은 두 개의 독립 스트림으로 구성된다. 첫째, RGB 스트림은 이미지의 경계, 얼굴 형상, 텍스처 등 의미론적 특징을 학습한다. 둘째, 주파수 스트림은 웨이블릿 변환과 DCT로 얻은 주파수 정보로부터 GAN·디퓨전 모델의 생성 과정에서 남는 고주파 인공물과 비정상 패턴을 포착한다. 두 스트림에서 추출된 특징은 다층 퍼셉트론(MLP, Multi-Layer Perceptron)을 사용한 후기 융합(Late fusion)으로 결합한다. 이를 통해 단일 도메인으로는 탐지하기 어려운 미세 조작 흔적까지 보완적으로 포착하여 일반화 성능을 극대화한다.

일반화 성능 검증을 위해 상이한 특성을 지닌 공개 데이터셋 Celeb-DF v2[19], DFDC[20], DeepfakeTIMIT[21], WildDeepfake[22]에 대해 교차 평가를 수행하였으며, 정확도, 정밀도, 재현율, F1 스코어를 지표로 사용하였다. 실험 결과, 제안한 MLP 기반 RGB - 주파수 후기 융합 프레임워크는 다양한 실제 환경을 가정한 실험에서 적용 가능성을 입증하였다.

본 논문의 구성은 다음과 같다. 2장에서는 RGB 색상 공간 기반 탐지와 주파수 기반 탐지 관련 연구를 다룬다. 3장에서는 데이터 전처리와 제안 모델의 세부 방법론을 기술한다. 4장에서는 실험 환경과 학습 및 테스트 데이터셋을 소개하고, 제안 모델과 비교 모델의 성능을 보고하며, 추가로 RGB/주파수 베이스 모델 비교를 통해 모델 선정 과정을 보여준다. 마지막으로 5장에서 결론과 향후 과제를 논의한다.

Ⅱ. 관련 연구

딥페이크 생성 기술이 빠르게 고도화됨에 따라, 그 진위를 판별하기 위한 탐지 연구 역시 다양한 방향으로 활발히 진행되고 있다. 탐지 연구의 핵심 전략은 조작 과정에서 발생하는 미세한 시각적·신

호적 불일치, 즉 아티팩트(Artifacts)를 포착하는 것이다. 본 장에서는 RGB 공간 기반 탐지와 주파수 기반 탐지의 주요 방법론과 각각의 기술적 한계를 정리한다.

2.1 RGB 색상 공간 기반 위조 탐지

초기 연구 다수는 단일 프레임의 공간적 왜곡을 포착하기 위해 CNN을 활용하였다. XceptionNet[23], EfficientNet[24] 기반 방법들은 RGB/HSV 등 색상 채널 간 상관관계, 경계부 블렌딩 흔적, 피부 텍스처의 통계적 일탈과 같은 단서를 이용해 여러 벤치마크에서 높은 정확도를 보였다. 또한 얼굴 정합 과정에서 발생하는 기하학적 왜곡, 조명 불일치, 컬러 톤 미스매치 등도 효과적인 신호로 활용되었다. 그러나 이러한 방법은 공간 도메인 단서에 대한 과의존이라는 구조적 한계를 지닌다. 예를 들어 NeuralTextures[25]와 같은 딥러닝 기반 고품질 합성은 경계부를 매끈하게 만들어 단서를 약화시키며, 저해상도·고압축 환경에서는 프레임 노이즈, 압축 블로킹(Blocking), 업샘플링 링잉(Ringing) 등의 영향으로 미세 단서의 신호대잡음비(SNR, Signal-to Noise Ratio)가 저하된다. 실제 환경에서는 카메라와 후처리 파이프라인이 제각각이라 학습 분포와의 도메인 편차가 커지고, 그 결과 압축률·해상도·조명이 달라지면 성능 하락이 빈번히 보고된다. 요컨대 RGB 기반 탐지는 자연 영상 통계에는 강하지만, 합성·압축이 남기는 주기적·스펙트럼 단서를 직접 모델링하지 못해 일반화에 취약하다.

2.2 주파수 기반 위조 탐지

RGB 단계 의존을 보완하기 위해, 합성·디코딩 과정에서 나타나는 비정상 스펙트럼 분포와 디코딩·업샘플링 단계의 주기적 잔향(링잉)을 직접 포착하는 주파수 기반 탐지 방법이 제안되었다. Y. Qian et al.[26]의 F3Net은 주파수 필터를 통해 고주파 대역의 미세 아티팩트를 강조하고, 이를 공간특징과 결합해 저품질 영상에서도 비교적 안정적인 성능을 보였다. H. Li et al.[27]의 FreqBlender는 데이터 생성 관점에서 일반화 취약성을 다루며, 얼굴 주파수 성분을 의미·구조·노이즈로 분해한 뒤 위조 영상의 구조 성분을 실제 영상에 주입하여 주파수 지식이 반영된 pseudo-fake를 생성한다. 이 방식은 전통적 공간 블렌딩 기반 pseudo-fake와 상보적이며, 결합 시 교차 데이터셋 일반화가 개선되는 것으로 보고된다.

그럼에도 주파수 전용 접근은 코덱·해상도 설정에 민감하고, 촬영·업샘플링 파이프 라인이 달라질 때 대역별 에너지 분포가 크게 변해 설정 의존적 튜닝이 필요하며, 주제·포즈·조명에 대한 해석력이 상대적으로 낮아 거짓양성과 거짓음성이 발생하기 쉽다. 즉, 주파수만으로는 내용 정합성을 점검하기 어렵고, RGB만으로는 합성·압축 특이성을 포착하기 어렵다.

본 연구는 RGB 색상 공간 모델과 주파수 기반 모델의 특징 수준 후기 융합(MLP)을 통해 도메인 간 분포 차이를 견고하게 포착하는 것을 목표로 한다. 이를 통해 개별 도메인 단서의 한계를 보완하고, 다양한 품질·압축·촬영 조건에서의 일반화 성능 향상을 지향한다.

III. 제안하는 방법

본 장에서는 제안하는 딥페이크 탐지 모델의 전체적인 구축 과정을 서술한다. 먼저 원본 비디오 데이터로부터 모델 학습에 사용할 데이터를 가공하는 전처리 과정을 설명한 후 RGB 및 주파수 정보를 사용하는 앙상블 모델의 구체적인 구조와 학습 방식을 기술한다. 먼저 원본 비디오로부터 학습용 입력을 구성하는 전처리를 기술하고, 이어서 RGB·주파수 이중 스트림과 특징 수준 후기 융합으로 이루어진 ER-FusNet의 구조와 학습 방식을 서술한다.

3.1 데이터 전처리

본 연구에서는 딥페이크 위조 탐지 연구의 학습 전 성능 향상을 위해 다음과 같이 데이터 전처리를 진행하였다. 학습 효율과 자원 소모를 고려하여 5프레임마다 1장을 균등 샘플링하여 정적 이미지

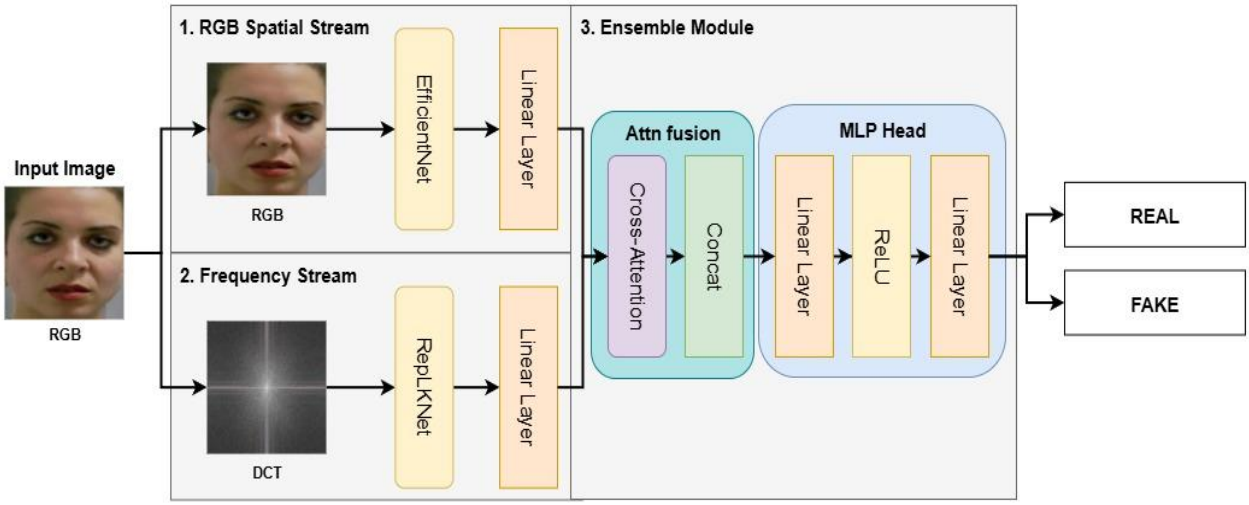


그림 1. ER-FusNet 모델 구조
Fig. 1. ER-FusNet model architecture

로 변환하였다. 이는 전체 프레임을 사용할 때 발생하는 중복 정보와 연산과 메모리에 사용되는 자원 부담을 완하하기 위함이다. 이후 샘플링된 각 이미지에 대해 Multitask Cascaded Convolutional Networks for Face Detection and Alignment[28]를 이용해 얼굴을 검출하였다. 검출한 검출 박스를 기준으로 얼굴 영역을 잘라낸 후 244 × 244 픽셀로 정규화 하였다. 이 전처리를 통해 배경에서 발생하는 노이즈를 줄이고 모델이 얼굴 중심의 위조 단서에 집중하도록

유도하였다. 또한, 제안하는 모델에 주파수 이미지를 입력하기 위해 전처리를 마친 RGB 이미지를 주파수 도메인으로 변환하였다. 본 연구에서는 대표적인 주파수 변환 기법인 FFT(Fast Fourier Transform)과 DCT(Discrete Cosine Transform)을 고려하였다.

FFT는 고주파 성분을 강조하여 조작된 프레임에 나타나는 미세한 인위적 패턴을 분리하는 데 유리하며, DCT는 블록 단위 압축 특성을 활용하여 합성 및 재인코딩 과정에서 발생하는 주파수 왜곡을 잘 드러낸다. 본 연구의 주파수 모델 비교 실험 결과, 본 연구의 데이터셋에서는 DCT가 더 안정적인 성능을 보였으므로, 이를 최종 표현 방식으로 채택하였다.

3.2 딥페이크 탐지 모델

본 연구에서 제안하는 딥페이크 탐지 모델 ER-FusNet은 그림 1과 같이 RGB 스트림과 주파수 스트림을 특징 수준에서 교차 어텐션으로 융합하는 이중 스트림 구조이다. RGB 스트림에는 EfficientNet을 적용하여 깊이와 너비, 해상도를 균형 있게 확장하는 복합 스케일링 전략을 통해 얼굴 이미지의 전역적·저주파 시각 패턴을 효과적으로 추출한다. 주파수 스트림에서는 DCT을 통해 RGB 영상에서 포착하기 어려운 고주파 영역의 위·변조 흔적 및 압축 인공물을 주파수 도메인으로 변환하고, RepLKNet의 대규모 커널 컨볼루션 구조를 활용하여 광범위한 수용영역을 확보함으로써 미세한 주파수 변동과 위조 패턴을 정밀하게 포착한다. 이와 같이 서로 상보적인 통계적 특성을 지닌 두 스트림을 교차 어텐션으로 결합함으로써, RGB 기반 전역 특징과 DCT 기반 국소 주파수 특징 간의 상관 관계를 학습하고 단일 도메인 기반 접근법 대비 위조 탐지의 강인성과 일반화 성능을 동시에 향상시킨다.

첫 번째는 RGB 스트림은 그림 2와 같은 구조의 EfficientNet-B7[24]을 사용하여 얼굴의 색상·형태 왜곡 등 공간·의미적 단서를 학습한다. 두 번째 주파수 스트림은 DCT 변환 후 그림 3과 같은 구조의 RepLKNet[29]를 사용하여 생성·압축 과정에서 남는 비정상 주파수 패턴과 아티팩트를 포착한다.

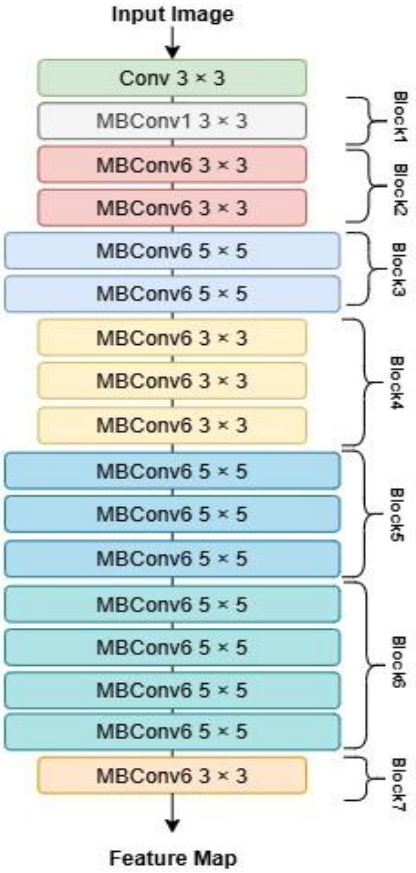


그림 2. EfficientNet-B7 모델 구조
Fig. 2. EfficientNet-B7 model architecture

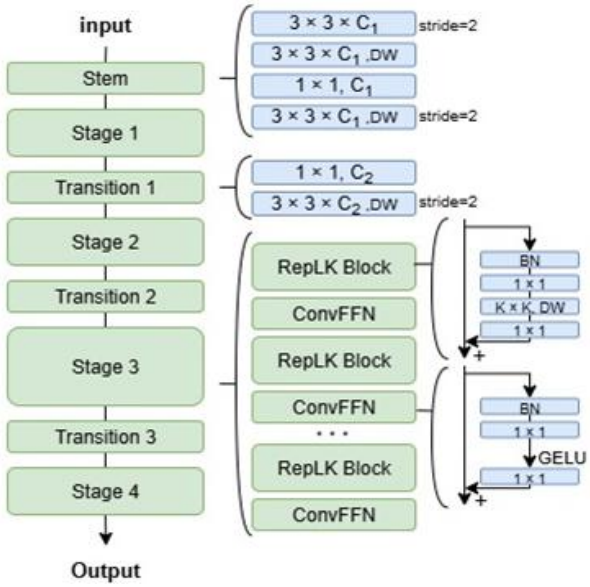


그림 3. RepLKNet 모델 구조
Fig. 3. RepLKNet model architecture

두 스트림의 융합 과정은 다음과 같다. 먼저 각 스트림에서 추출된 t번째 프레임의 RGB 특징 벡터 $f_t^{rgb} \in R^{D_r}$ 주파수 특징 벡터 $f_t^{freq} \in R^{D_f}$를 식 (1)~(3)에 따라 선형 계측으로 공통 차원 d에 투영한다.

$$f_t^{rgb} = W_r f_t^{rgb} \tag{1}$$

$$f_t^{freq} = W_r f_t^{freq} \tag{2}$$

$$f_t^{rgb}, f_t^{freq} \in R^d \tag{3}$$

여기서 $W_r \in R^{d \times D_r}$ $W_f \in R^{d \times D_f}$는 각각 RGB와 주파수 특징을 공통 차원으로 정렬하기 위한 투영 가중치이다. 이어서 식 (4)에서 투영된 RGB 특징을 쿼리로, 식 (5)에서 주파수 특징을 키와 값으로 변환한다. 이후 식 (6)에 제시된 다중 헤드 어텐션을 적용하여 주파수 정보로 보강된 RGB 특징 f_t^{rgb}를 생성한다. 이 과정에서는 잔차 연결과 LN (LayerNorm)이 포함된다.

$$q_t = W_q f_t^{rgb} \tag{4}$$

$$k_t = W_k f_t^{freq} \tag{5}$$

$$v_t = W_v f_t^{freq} \tag{6}$$

쿼리, 키, 값을 생성하는 선형 투영 가중치 $W_q, W_k, W_v \in R^{d \times \mathrm{d}}$ 는 각 스트림 별로 독립적으로 학습되며, 이렇게 얻은 (q_t, k_t, v_t)에 대하여 식 (7)과 같이 멀티 헤드 어텐션(MHA)을 계산한다. 이러한 MHA는 h개의 헤드를 병렬로 적용한 뒤 출력을 연결한다. 이후 식 (8)에서 어텐션 출력 a_t에 원본 RGB 경로 정보를 $W_{res} f_t^{rgb}$와 결합해 잔차를 더함으로써 RGB 고유 단서를 보존하고 학습 안정성을 높인다. 마지막으로 식 (9)에 따라 s_t에 LN을 적용하여 특징 분포를 정규화함으로써 입력 스케일 편차를 줄이고, 후단 MLP의 최적화와 일반화에 유리한 보강 특징 f_t^{rgb}를 제공한다.

$$a_t = MHA(q_t, k_t, v_t) \tag{7}$$

$$s_t = a_t + W_{res} f_t^{rgb} \tag{8}$$

$$f_t^{rgb} = LN(s_t) \in R^d \tag{9}$$

최종적으로, 식 (10)과 같이 보강된 RGB 특징 f_t^{rgb}과 투영된 원본 주파수 특징 f_t^{freq}을 연결해 이

를 프레임마다 가중치가 공유되는 2계층 MLP에 입력되며, 식 (11)에서 은닉 표현을 얻는다. 이후 식 (12)에 따라 프레임 단위의 위조 로짓 y_t를 계산한다.

$$z_t = [f_t^{rgb};\ f_t^{freq}] \in R^{2d} \tag{10}$$

$$h_t = ReLU(W_1 z_t + b_1) \tag{11}$$

$$y_t = W_2 u_t + b_2 \tag{12}$$

이렇게 얻은 프레임별 로짓에 소프트맥스 함수를 적용하여 위조 확률을 계산하고, 식 (13)과 같이 이를 영상 V에 대해 평균하여 영상 수준의 최종 판별 결과를 도출한다.

$$\hat{p}(fakevert\ V) = \frac{1}{T}\sum_{t=1}^{T} softmax(y_t)_{[fake]} \tag{13}$$

이러한 교차 어텐션 기반 융합 구조는 두 모달리티 간의 상호작용을 모델링함으로써, 다양한 외부 조건 변화에도 강건한 일반화 성능을 달성하는 것을 목표로 한다.

Ⅳ. 실험 및 성능 평가

4.1 실험 환경

본 연구의 실험은 Ubuntu 20.04 LTS 운영체제를 사용하는 두 대의 독립 서버에서 수행되었다. 첫 번째 서버는 Intel Core i9-10900X CPU, 188GB RAM, NVIDIA GeForce RTX 4090 GPU 4대로 구성되었으며, 두 번째 서버는 NVIDIA GeForce RTX 3090 GPU 4대를 장착하였다.

4.2 데이터셋

본 연구의 훈련 데이터셋은 대표적인 딥페이크 탐지 분야의 공개 데이터셋인 FaceForensics++[30]를 사용하였다. FaceForensics++은 유튜브에서 수집된 977개의 원본 영상과 이를 FaceSwap, DeepFakes, Face2Face, NeuralTextures라는 네 가지의 대표적인 조작 기법을 이용해 각각 생성한 1,000개의 위조 영상과 28명의 배우가 촬영한 386개의 원본 영상과 이를 DeepFakes 기법으로 변조된 3,068개의 위조 영상으로 구성된다. 본 연구에서는 영상의 품질 저하가 없는 raw 버전을 사용하여 데이터에 내재 된 미세한 조작 흔적까지 모델이 학습하도록 설계하였다.

한편, 모델의 일반화 성능을 다각적으로 검증하기 위해, 동일 데이터셋을 내부적으로 나누어 훈련과 평가하는 대신 훈련에 사용되지 않은 외부 데이터셋으로만 평가하는 교차 데이터셋 설정을 채택하였다. 동일 데이터셋 내 분할은 동일 인물, 카메라, 압축과 같은 식별자 및 촬영 환경 단서가 공유되어 모델의 성능이 과대 추정될 수 있다. 반면, 외부 데이터셋 기반 평가는 촬영 조건과 조작 기법이 달라지는 분포 불일치 상황을 모사함으로써, 실제 배포 환경에서의 도메인 일반화 능력을 보다 정확히 측정할 수 있다.

각 평가 데이터셋은 다른 목적과 특성을 가지도록 의도적으로 선정되었다. Celeb-DF V2는 시각적 결함이 적은 고품질 위조 영상에 대한 모델의 탐지 정교성을 평가한다[19]. DFDC는 통제된 환경에서 제작된 영상에 대한 기본적인 탐지 성능을 평가한다[20]. DeepfakeTIMIT는 의도적으로 화질이 제어된 영상들로 구성되어, 영상의 압축 및 품질 저하 환경에서의 모델 강건성을 평가한다[21]. WildDeepfake는 실제 인터넷 환경에서 수집된 영상들을 통해 모델의 실용성을 검증한다[22]. 이처럼 다각화된 평가 데이터셋은 제안 모델이 다양한 생성 기법과 데이터 분포를 가진 환경에서도 안정적인 성능을 유지하는지를 검증하는 데 사용된다. 본 논문에서 사용한 훈련 및 평가 데이터셋의 상세 구성 및 표본 수는 표 1과 같다.

표 1. 훈련 및 평가 데이터셋의 구성 및 샘플 수
Table 1. Distribution of datasets used in experiments

	Dataset	Label	Count
Train	FF++	Real	1,363
		Fake	7,068
Test	Celeb-DF V2	Real	5,698
		Fake	300
	DFDC	Real	363
		Fake	3,068
	DeepfakeTIMIT	Fake	620
	WildDeepfake	Real	3,805
		Fake	3,509

4.3 제안 모델과 비교 모델의 성능 비교

본 연구에서 제안하는 RGB - 주파수 융합 모델 ER-FusNet의 성능을 입증하기 위해 대표적인 단일 스트림 모델들과 비교하였다. RGB 비교군으로는 Xception과 EfficientNet으로 구성되었다. 두 백본은 딥페이크 탐지 선행 연구에서 널리 사용되는 비교 모델로, 재현성과 대표성이 높다. 주파수 비교군은 주파수 기반 딥페이크 탐지에서 폭넓게 인용되는 F3Net과 FreqBlender를 선정하였다. 한편 ER-FusNet의 주파수 스트림 내부 백본으로는 RepLKNet을 사용하였으나, 이는 전통적인 주파수 전용 모델이 아니므로 비교 모델로는 포함하지 않았다. 표 2에서 보듯이 ER-FusNet은 Celeb-DF v2, DFDC, DeepfakeTIMIT에서 기존 RGB 기반 모델 Xception, EfficientNet과 동등하거나 그 이상의 성능을 보여주며, F1 스코어 0.92에서 0.992라는 최고 수준 성능을 유지하였다. 반면, 실제 환경을 모사한 WildDeepfake에서는 RGB 모델 Xception에서 F1 스코어 0.6363, EfficientNet에서 F1 스코어 0.3857로 급격히 하락하였다. 이에 비해 ER-FusNet의 F1 스코어는 0.6470으로 단일 스트림 평균인 0.53 대비 약 12% 향상되었으며, 재현율은 0.9894로 모든 모델 중 가장 높은 성능을 기록하였다.

주파수 전용 모델인 F3Net과 FreqBlender는 대부분 데이터셋에서 낮은 성능을 보였으나, WildDeepfake에서는 각각 0.6526과 0.4479의 F1 스코어를 기록하여 주파수 단서가 RGB와는 다른 일반화 기여 가능성을 보여주었다. 특히 F3Net은 WlidDeepfake에서

ER-FusNet과 유사한 F1 스코어를 보였지만 Celeb-DF v2의 F1 스코어 0.0663과 같이 다른 데이터셋에서는 심각한 성능 저하가 발생하였다.

종합하면 ER-FusNet은 RGB 모델이 강점을 보이는 정형 데이터셋에서 성능 저하 없이 동등한 수준을 유지하면서, WildDeepfake와 같이 RGB 모델이 취약한 실제 환경에서는 주파수 모델의 강점을 효과적으로 흡수하여 안정적이고 일관된 성능을 달성하였다. 이는 RGB와 주파수 단서를 상보적 특징을 융합하는 본 연구의 접근 방식이 딥페이크 탐지의 일반화 성능 향상에 기여함을 뒷받침한다.

표 2. 제안 모델과 비교 모델의 성능 비교
Table 2. Performance comparison between the proposed model and baseline models

Model		Test Dataset	Accuracy	Precision	Recall	F1-Score
RGB	Xception[23]	Celeb-DF V2	0.8638	0.8637	1.0000	0.9269
		DFDC	0.8935	0.8934	1.0000	0.9437
		DeepfakeTIMIT	1.0000	1.0000	1.0000	1.0000
		WildDeepfake	0.4778	0.4778	0.9521	0.6363
	EfficiententNet[24]	Celeb-DF V2	0.8578	0.8634	0.9923	0.9234
		DFDC	0.8885	0.8936	0.9934	0.9409
		DeepfakeTIMIT	0.9748	1.0000	0.9748	0.9872
		WildDeepfake	0.4274	0.3974	0.3747	0.3857
Frequency	F3Net[26]	Celeb-DF V2	0.1603	0.8394	0.0345	0.0663
		DFDC	0.1100	1.0000	0.0039	0.0078
		DeepfakeTIMIT	0.0566	1.0000	0.0566	0.1071
		WildDeepfake	0.5509	0.5189	0.8791	0.6526
	FreqBlender[27]	Celeb-DF V2	0.4453	0.8592	0.4280	0.5714
		DFDC	0.4293	0.8986	0.4072	0.5604
		DeepfakeTIMIT	0.0063	1.0000	0.0063	0.0125
		WildDeepfake	0.5122	0.4900	0.4124	0.4479
Ensemble	Ours	Celeb-DF V2	0.8568	0.8651	0.9884	0.9226
		DFDC	0.8835	0.8942	0.9862	0.9379
		DeepfakeTIMIT	0.9811	1.0000	0.9811	0.9904
		WildDeepfake	0.4820	0.4806	0.9894	0.6470

4.4 RGB 색상 공간 백본 모델 선정

본 연구에서는 최적의 RGB 색상 공간 기반 모델을 선정하기 위해 CoaTNet[31], HorNet[32], MaxVit[33], Xception[23], EfficientNet-B7[24]에 대한 성능을 비교하여 최적의 아키텍쳐를 선정하였다. 손실함수는 교차 엔트로피와 Focal 손실함수[34]를 각각 적용하여 성능을 비교 검증을 진행하였다. 모델들의 학습은 동일한 하이퍼파라미터 조건에서 진행하였다. 학습률은 1e-4, 옵티마이저는 AdamW, 활성화 함수는 ReLU, 배치 크기는 16, 최대 에포크는 2000으로 설정하고 조기 종료의 patience를 5로 설정하였다. Focal 손실함수의 경우, 가중치 조절 파라미터인 감마는 2.0, 알파는 0.75, 0.25로 설정했다.

표 3. RGB 색상 공간 모델 성능 비교
Table 3. Performence comparison of RGB models

Loss	Model	Test Dataset	Accuracy	Precision	Recall	F1-Score
Cross-Entropy	CoaTNet[31]	Celeb-DF V2	0.6804	0.8583	0.7545	0.8031
		DFDC	0.6775	0.8942	0.7247	0.8006
		DeepfakeTIMIT	1.0000	1.0000	1.0000	1.0000
		WildDeepfake	0.4699	0.4736	0.9432	0.6306
	HorNet[32]	Celeb-DF V2	0.8408	0.8639	0.9683	0.9131
		DFDC	0.8583	0.8924	0.9566	0.9234
		DeepfakeTIMIT	0.9182	1.0000	0.9182	0.9573
		WildDeepfake	0.4834	0.4813	0.9900	0.6477
	MaxVit[33]	Celeb-DF V2	0.8639	0.8639	1.0000	0.9269
		DFDC	0.8926	0.8936	0.9986	0.9432
		DeepfakeTIMIT	1.0000	1.0000	1.0000	1.0000
		WildDeepfake	0.4804	0.4798	0.9886	0.6461
	Xception[23]	Celeb-DF V2	0.8638	0.8637	1.0000	0.9269
		DFDC	0.8935	0.8934	1.0000	0.9437
		DeepfakeTIMIT	1.0000	1.0000	1.0000	1.0000
		WildDeepfake	0.4778	0.4778	0.9521	0.6363
	EfficiententNet[24]	Celeb-DF V2	0.8578	0.8634	0.9923	0.9234
		DFDC	0.8885	0.8936	0.9934	0.9409
		DeepfakeTIMIT	0.9748	1.0000	0.9748	0.9872
		WildDeepfake	0.4274	0.3974	0.3747	0.3857
Focal	CoatNet[31]	Celeb-DF V2	0.8587	0.8638	0.9928	0.9239
		DFDC	0.8477	0.8939	0.9412	0.9169
		DeepfakeTIMIT	1.0000	1.0000	1.0000	1.0000
		WildDeepfake	0.4784	0.4782	0.9561	0.6375
	HorNet[32]	Celeb-DF V2	0.4107	0.8488	0.3867	0.5313
		DFDC	0.4357	0.8912	0.4197	0.5706
		DeepfakeTIMIT	0.9984	1.0000	0.9984	0.9992
		WildDeepfake	0.5099	0.4943	0.9302	0.6455
	MaxVit[33]	Celeb-DF V2	0.8197	0.8617	0.9425	0.9002
		DFDC	0.8923	0.8933	0.9987	0.9430
		DeepfakeTIMIT	0.9685	1.0000	0.9685	0.9840
		WildDeepfake	0.5128	0.4940	0.6344	0.5554
	Xception[23]	Celeb-DF V2	0.8637	0.8638	1.0000	0.9269
		DFDC	0.8935	0.8935	1.0000	0.9437
		DeepfakeTIMIT	1.0000	1.0000	1.0000	1.0000
		WildDeepfake	0.4648	0.4710	0.9384	0.6272
	EfficiententNet[24]	Celeb-DF V2	0.1362	0.0000	0.0000	0.0000
		DFDC	0.1065	0.0000	0.0000	0.0000
		DeepfakeTIMIT	0.0000	0.0000	0.0000	0.0000
		WildDeepfake	0.5202	0.0000	0.0000	0.0000

표 3의 결과에서 확인할 수 있듯이 교차 엔트로피를 사용하면 대부분의 모델이 Celeb-DF v2, DFDC, DeepfakeTIMIT에서 F1 스코어 0.92에서 0.99라는 우수한 성능을 보였다. 그러나 WildDeepfake에서는 CoaTNet은 F1 스코어 0.6306, EfficientNet은 F1 스코어 0.3857 등 일부 모델의 성능이 급격히 저하되었다. 반면 Xception은 F1 스코어 0.6363, HorNet은 F1 스코어 0.6467, MaxVit는 F1 스코어 0.646은 상대적으로 높은 수준을 유지하였다. Focal 손실 함수를 적용한 경우, 일부 모델에서는 개선되기도 했으나 전반적으로 불안정성이 나타났다. 예를 들어 CoaTNet은 Celeb-DF v2에서 F1 스코어 0.9239와 DeepfakeTIMIT에서 F1 스코어 1.0000에서 안정적인 성능을 보였지만, HorNet은 Celeb-DF v2에서 F1 스코어 0.5313, DFDC에서 0.5706으로 성능이 크게 저하되었다. EfficientNet-B7은 모든 데이터셋에서 F1 스코어가 0.0 수준으로 붕괴하여, 작은 배치 크기와 Focal 손실의 결합이 학습 불안정을 초래한 것으로 판단된다. 이러한 결과를 종합하면 교차 엔트로피 손실 함수가 RGB 스트림 학습에 더 안정적임을 알 수 있다. 또한 DeepfakeTIMIT과 같은 통제된 데이터셋에 서는 대부분의 모델이 완벽에 가까운 F1 스코어가 1.0000을 기록하였으나, 실제 환경을 모사한 WildDeepfake에서는 전반적인 성능이 하락하였다, 따라서 본 연구에서는 안정성과 성능을 균형 있게 확보하기 위해 교차 엔트로피 손실 함수를 적용한 EfficientNet-B7을 최종 RGB 백본 모델로 선정하였다.

표 4. 주파수 모델 성능 비교
Table 4. Performence comparison of Frequency models

Loss	Frequency	Model	Test Dataset	Accuracy	Precision	Recall	F1-Score
Cross-Entropy	DCT	ConvNeXt-V2[35]	Celeb-DF V2	0.8612	0.8637	0.9966	0.9253
			DFD	0.8914	0.8934	0.9973	0.9425
			DeepfakeTIMIT	0.9984	1.0000	0.9984	0.9992
			WildDeepfake	0.4798	0.4798	1.0000	0.6484
		RepLKNet[29]	Celeb-DF V2	0.8637	0.8637	1.0000	0.9269
			DFDC	0.8935	0.8935	1.0000	0.9437
			DeepfakeTIMIT	1.0000	1.0000	1.0000	1.0000
			WildDeepfake	0.4798	0.4798	1.0000	0.6484
	FFT	ConvNeXt-V2[35]	Celeb-DF V2	0.1362	0.0000	0.0000	0.0000
			DFDC	0.1065	0.0000	0.0000	0.0000
			DeepfakeTIMIT	0.0000	0.0000	0.0000	0.0000
			WildDeepfake	0.5202	0.0000	0.0000	0.0000
		RepLKNet[29]	Celeb-DF V2	0.1362	0.0000	0.0000	0.0000
			DFDC	0.1065	0.0000	0.0000	0.0000
			DeepfakeTIMIT	0.0000	0.0000	0.0000	0.0000
			WildDeepfake	0.5202	0.0000	0.0000	0.0000
Focal	DCT	ConvNeXt-V2[35]	Celeb-DF V2	0.7631	0.8633	0.8622	0.8627
			DFD	0.7961	0.8940	0.8755	0.8846
			DeepfakeTIMIT	0.9025	1.0000	0.9025	0.9487
			WildDeepfake	0.4807	0.4801	0.9974	0.6482
		RepLKNet[29]	Celeb-DF V2	0.1362	0.0000	0.0000	0.0000
			DFDC	0.1065	0.0000	0.0000	0.0000
			DeepfakeTIMIT	0.0000	0.0000	0.0000	0.0000
			WildDeepfake	0.5202	0.0000	0.0000	0.0000
	FFT	ConvNeXt-V2[35]	Celeb-DF V2	0.1362	0.0000	0.0000	0.0000
			DFDC	0.1065	0.0000	0.0000	0.0000
			DeepfakeTIMIT	0.0000	0.0000	0.0000	0.0000
			WildDeepfake	0.5202	0.0000	0.0000	0.0000
		RepLKNet[29]	Celeb-DF V2	0.1362	0.0000	0.0000	0.0000
			DFDC	0.1065	0.0000	0.0000	0.0000
			DeepfakeTIMIT	0.0000	0.0000	0.0000	0.0000
			WildDeepfake	0.5202	0.0000	0.0000	0.0000

4.5 주파수 백본 모델 선정

주파수 스트림 백본 후보인 RepLKNet과 ConvNeXt-V2에 대해 성능을 비교하여 최적의 아키텍처를 결정하였다. 입력은 각 비디오 프레임을 DCT와 FFT로 변환하여 성능을 검증하였으며, 손실함수는 교차 엔트로피와 Focal을 각각 적용하였다. 학습률은 1e-4, 옵티마이저는 AdamW, 활성화 함수는 ReLU, 배치 크기는 16, 최대 에포크는 2000으로 설정하였고, Focal 손실함수의 경우 감마를 2.0, 알파를 0.75와 0.25로 설정하였다.

표 4의 결과에서 확인할 수 있듯이, FFT 기반 입력은 ConvNeXt-V2와 RepLKNet 모두에서 F1 스코어가 0.0 수준에 머물러 학습이 불안정하였다. 반면 DCT 기반 입력은 안정적인 성능을 보여, Cross-Entropy 손실을 적용한 ConvNeXt-V2(DCT)와 RepLKNet(DCT)는 Celeb-DF v2, DFDC, DeepfakeTIMIT에서 F1 스코어가 0.93에서 0.99를 기록하였고, WildDeepfake에서도 F1 스코어 0.648로 비교적 높은 성능을 유지하였다. Focal 손실은 ConvNeXt-V2(DCT)에서는 일정 수준의 성능을 유지했으나, RepLKNet(DCT)에서는 전반적으로 성능이 붕괴하였다. 따라서 클래스 불균형이 크지 않은 데이터셋 특성을 고려할 때, Cross-Entropy 손실이 더 안정적인 학습을 보장하였다.

최종적으로 ConvNeXt-V2(DCT)와 RepLKNet(DCT)는 유사한 성능을 보였으나, ConvNeXt-V2가 196.4M 파라미터를 사용하는 반면 RepLKNet은 79.8M 파라미터로 훨씬 경량적이었다. 이에 따라 본 연구에서는 Cross-Entropy 손실 함수를 적용한 RepLKNet(DCT)를 최종 주파수 백본 모델로 선정하였다.

V. 결론 및 향후 과제

본 연구는 고도화된 딥페이크 기술에 대응하기 위해 기존 탐지 모델의 일반화 성능 한계를 극복하고자, RGB 공간 정보와 주파수 도메인 정보를 상보적으로 결합하는 이중 스트림 융합 모델 ER-FusNet을 제안하였다. 제안 모델은 비디오 프레임을 전처리하여 얼굴 영역을 정규화한 뒤, EfficientNet-B7과 RepLKNet을 각각 RGB와 주파수 스트림의 백본으로 활용하고, 추출된 특징을 2계층 MLP를 통해 융합하여 위조 여부를 판별한다.

다양한 공개 데이터셋을 활용한 교차 검증 결과, ER-FusNet은 단일 정보에 의존하는 기존 모델 대비 모든 지표에서 일관된 성능 우위를 보였다. 특히 실제 환경을 모사한 WildDeepfake 데이터셋에서 가장 큰 성능 격차를 기록함으로써, 제안된 융합 방식이 딥페이크 탐지의 일반화 성능을 효과적으로 개선할 수 있음을 입증하였다.

본 연구의 기여는 다음과 같다. 첫째, RGB가 제공하는 의미론적 단서와 주파수가 드러내는 저수준 아티팩트를 효과적으로 융합하는 새로운 프레임워크를 제안하였다. 둘째, 체계적인 교차 데이터셋 평가를 통해, 실제 인터넷 환경과 유사한 복잡한 조건에서도 강건한 성능을 확인함으로써 실용적 가능성을 제시하였다. 한편, 본 연구는 GAN 기반 합성 데이터에 초점을 맞추고 있어 확산 모델 등 최신 생성 기술에 대한 추가 검증이 필요하다. 또한 이중 스트림 구조로 인한 계산 비용 문제와 Real과 Fake 클래스 불균형 문제 또한 남아 있다. 향후 연구에서는 데이터 불균형 완화를 위한 SMOTE 기반 증강과 Class-balanced loss, 리샘플링, 임계값 보정등을 병행하고 최신 생성 데이터를 포함한 학습 데이터 확장과 더불어, 지식 증류와 같은 경량화 기법을 적용하여 효율성을 높이는 방향으로 발전시킬 수 있을 것이다.

References

[1] G. M. Haley and B. S. Manjunath, "Rotation-Invariant Texture Classification Using a Complete Space-Frequency Model", IEEE Transactions on Image Processing, Vol. 8, No. 2, pp. 255-269, Feb. 1999. https://doi.org/10.1109/83.743859.

[2] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation", arXiv preprint arXiv:1710.10196, Oct. 2017. https://doi.org/10.48550/arXiv.1710.10196.

[3] A. Brock, J. Donahue, and K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis", arXiv preprint arXiv:1809.11096, Sep. 2018. https://doi.org/10.48550/arXiv.1809.11096.

[4] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation", Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, pp. 8789-8797, Jun. 2018. https://doi.org/10.1109/CVPR.2018.00916.

[5] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks", Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 4401-4410, Jun. 2019. https://doi.org/10.1109/CVPR.2019.00453.

[6] T. Dudeja, S. K. Dubey, and A. K. Bhatt, "Ensembled EfficientNetB3 Architecture for Multi-Class Classification of Tumours in MRI Images", Intelligent Decision Technologies, Vol. 17, No. 2, pp. 395-414, May 2023. https://doi.org/10.3233/IDT-220150.

[7] R. Katarya and A. Lal, "A Study on Combating Emerging Threat of Deepfake Weaponization", Proc. 4th Int. Conf. on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, pp. 485-490, Oct. 2020. https://doi.org/10.1109/I-SMAC49090.2020.9243588.

[8] S. A. Buo, "The Emerging Threats of Deepfake Attacks and Countermeasures", arXiv preprint arXiv:2012.07989, Dec. 2020. https://doi.org/10.48550/arXiv.2012.07989.

[9] J. Ice, "Defamatory Political Deepfakes and the First Amendment", Case Western Reserve Law Review, Vol. 70, No. 2, pp. 417-455, 2019. https://scholarlycommons.law.case.edu/caselrev/vol70/iss2/12.

[10] M. Safiullah and N. Parveen, "Big Data, Artificial Intelligence and Machine Learning: A Paradigm Shift in Election Campaigns", The New Advanced Society, Wiley, Hoboken, NJ, USA, pp. 247-261, 2022. https://doi.org/10.1002/9781119884392.ch11.

[11] V. Karasavva and A. Noorbhai, "The Real Threat of Deepfake Pornography: A Review of Canadian Policy", Cyberpsychology, Behavior, and Social Networking, Vol. 24, No. 3, pp. 203-209, Mar. 2021. https://doi.org/10.1089/cyber.2020.0272.

[12] P. Korshunov and S. Marcel, "Vulnerability Assessment and Detection of Deepfake Videos", Proc. Int. Conf. on Biometrics (ICB), Crete, Greece, pp. 1-6, Jun. 2019. https://doi.org/10.1109/ICB45273.2019.8987375.

[13] K. N. Ramadhani and R. Munir, "A Comparative Study of Deepfake Video Detection Method", Proc. 3rd Int. Conf. on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, pp. 394-399, Nov. 2020. https://doi.org/10.1109/ICOIACT50329.2020.9331963.

[14] F. Coletti, "Revenge Porn: The Concept and Practice of Combatting Nonconsensual Sexual Images in Europe", Master's thesis, University of Latvia (European Master's Degree in Human Rights and Democratisation), 2017. https://repository.gchumanrights.org/items/95d62d70-a749-4d36-9635-c33f191694e7.

[15] K. Mania, "Legal Protection of Revenge and Deepfake Porn Victims in the European Union: Findings From a Comparative Legal Study", Trauma, Violence, & Abuse, Vol. 25, No. 1, pp. 117-129, Jan. 2024. https://doi.org/10.1177/15248380221143772.

[16] K. N. Ramadhani and R. Munir, "A Comparative Study of Deepfake Video Detection Method", Proc. 3rd Int. Conf. on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, pp. 394-399, Nov. 2020. https://doi.org/10.1109/ICOIACT50329.2020.9331963.

[17] F. Matern, C. Riess, and M. Stamminger,

"Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations", Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW), Waikoloa Village, HI, USA, pp. 83-92, Jan. 2019. https://doi.org/10.1109/WACVW.2019.00020.

[18] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-Stream Neural Networks for Tampered Face Detection", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Honolulu, HI, USA, pp. 1831-1839, Jul. 2017. https://doi.org/10.1109/CVPRW.2017.229.

[19] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics", Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), Online, pp. 3207-3216, Jun. 2020. https://doi.org/10.1109/CVPR42600.2020.00327.

[20] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The DeepFake Detection Challenge (DFDC) Dataset", arXiv preprint arXiv:2006.07397, Jun. 2020. https://doi.org/10.48550/arXiv.2006.07397.

[21] P. Korshunov and S. Marcel, "Deepfakes: A New Threat to Face Recognition? Assessment and Detection", arXiv preprint arXiv:1812.08685, Dec. 2018. https://doi.org/10.48550/arXiv.1812.08685.

[22] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection", Proc. 28th ACM Int. Conf. on Multimedia (MM '20), Seattle, WA, USA (Virtual), pp. 2382-2390, Oct. 2020. https://doi.org/10.1145/3394171.3413769.

[23] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions", Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 1251-1258, Jul. 2017. https://doi.org/10.1109/CVPR.2017.195.

[24] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", Proc. 36th Int. Conf. on Machine Learning (ICML), Long Beach, CA, USA, pp. 6105-6114, Jun. 2019. https://proceedings.mlr.press/v97/tan19a.html.

[25] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred Neural Rendering: Image Synthesis Using Neural Textures", ACM Transactions on Graphics, Vol. 38, No. 4, pp. 66:1-66:12, Jul. 2019. https://doi.org/10.1145/3306346.3323035.

[26] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues", Proc. European Conf. on Computer Vision (ECCV), Glasgow, UK (Virtual), pp. 86-103, Aug. 2020. https://doi.org/10.1007/978-3-030-58610-2_6.

[27] H. Li, J. Zhou, Y. Li, B. Wu, B. Li, and J. Dong, "FreqBlender: Enhancing DeepFake Detection by Blending Frequency Knowledge", Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Vancouver, BC, Canada, Dec. 2024. https://openreview.net/forum?id=6QA5eKNl34.

[28] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks", IEEE Signal Processing Letters, Vol. 23, No. 10, pp. 1499-1503, Oct. 2016. https://doi.org/10.1109/LSP.2016.2603342.

[29] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling Up Your Kernels to 31×31: Revisiting Large Kernel Design in CNNs", Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 11963-11975, Jun. 2022. https://doi.org/10.1109/CVPR52688.2022.01166.

[30] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images", Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), Seoul, South Korea, pp. 1-11, Oct. 2019. https://doi.org/10.1109/ICCV.2019.00009.

[31] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying Convolution and Attention for All Data Sizes", Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Online, Vol. 34, Dec. 2021. https://proceedings.neurips.cc/paper/2021/hash/20568692db622456cc42a2e853ca21f8-Abstract.html.

[32] Y. Rao, W. Zhao, Y. Tang, J. Zhou, S.-N. Lim, and J. Lu, "HorNet: Efficient High-Order Spatial Interactions with Recursive Gated Convolutions", Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), New Orleans, USA, Dec. 2022. https://proceedings.neurips.cc/paper_files/paper/2022/hash/436d042b2dd81214d23ae43eb196b146-Abstract-Conference.html.

[33] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. C. Bovik, and Y. Li, "MaxViT: Multi-Axis Vision Transformer", Proc. European Conf. on Computer Vision (ECCV), Tel Aviv, Israel, pp. 459-479, Oct. 2022. https://doi.org/10.1007/978-3-031-20053-3_27.

[34] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection", Proc. IEEE Int. Conf. on Computer Vision (ICCV), Venice, Italy, pp. 2999-3007, Oct. 2017. https://doi.org/10.1109/ICCV.2017.324.

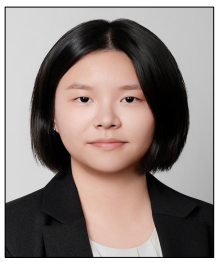
[35] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt V2: Co-Designing and Scaling ConvNets With Masked Autoencoders", Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada, pp. 16133-16142, Jun. 2023. https://doi.org/10.1109/CVPR52729.2023.01548.

저자소개

하 유 진 (Yu-Jin Ha)

2023년 2월 : 경상국립대학교 컴퓨터공학부(공학사)
2023년 3월 ~ 현재 : 경상국립대학교 AI융합공학부 석사과정
관심분야 : 인공지능, 멀티모달, 컴퓨터 비전, 딥페이크

박 수 진 (Su-Jin Park)



2023년 3월 ~ 현재 : 경상국립대학교 경영정보학과 학사과정
관심분야 : 인공지능, Meme Detection, 딥보이스, 딥페이크

박 종 찬 (Jong-Chan Park)

2020년 2월 : 동아대학교 전자공학과(공학사)
2022년 2월 : 동아대학교 전자공학과(공학석사)
2022년 2월 ~ 현재 : 경상국립대학교 AI융합공학과 박사과정
관심분야 : Deepfake detection, defense LLMs, fire/smoke perception on edge devices

김 건 우 (Gun-Woo Kim)

2006년 12월 : 호주뉴캐슬대학교 컴퓨터공학과(공학사)
2007년 9월 : 호주뉴캐슬대학교 정보공학과(공학석사)
2017년 8월 : 한양대학교 컴퓨터공학과(공학박사)
2021년 9월 ~ 현재 : 경상국립대학교 컴퓨터공학과 조교수
관심분야 : 인공지능, 시멘틱 헬스케어, 데이터마이닝