Visual Saliency 기반의 딥페이크 이미지 탐지 기법

[1] 노하림, [2*]유제혁

Deepfake Image Detection based on Visual Saliency

[1]Harim Noh, [2*]Jehyeok Rew

요 약

딥페이크(Deepfake)란 다양한 인공지능 기술을 활용해 진짜와 같은 가짜를 만드는 영상 합성기술로, 가짜 뉴스 생성, 사기, 악의적인 도용 등에 활용되어 개인과 사회에게 심각한 혼란을 유발시키고 있다. 사회적 문제방지를 위해, 딥페이크로 생성된 이미지를 정교하게 분석하고 탐지하는 방법이 필요하다. 따라서, 본 논문에서는 딥페이크로 생성된 가짜 이미지와 진짜 이미지에서 Saliency 특징을 각각 추출하고 분석하여 합성 후보 영역을 검출하며, 추출된 특징들을 중점으로 학습하여 최종적으로 딥페이크 이미지 탐지 모델을 구축하였다. 제안된 Saliency 기반의 딥페이크 탐지 모델은 합성된 이미지, 동영상 등의 딥페이크 검출 상황에서 공통적으로 사용될 수 있으며, 다양한 비교실험을 통해 본 논문의 제안 방법이 효과적임을 보였다.

Abstract

'Deepfake' refers to a video synthesis technique that utilizes various artificial intelligence technologies to create highly realistic fake content, causing serious confusion to individuals and society by being used for generating fake news, fraud, malicious impersonation, and more. To address this issue, there is a need for methods to detect malicious images generated by deepfake accurately. In this paper, we extract and analyze saliency features from deepfake and real images, and detect candidate synthesis regions on the images, and finally construct an automatic deepfake detection model by focusing on the extracted features. The proposed saliency feature-based model can be universally applied in situations where deepfake detection is required, such as synthesized images and videos. To demonstrate the performance of our approach, we conducted several experiments that have shown the effectiveness of the deepfake detection task.

Keywords: *Deepfake, Malicious Manipulation, Visual Saliency, Image Synthesis, Deep Learning*

[1] *덕성여자대학교 컴퓨터공학전공 학사과정 (doris5093@duksung.ac.kr)*
[2*] *교신저자 덕성여자대학교 데이터사이언스학과 교수(jhrew@duksung.ac.kr)*

Received: Jan. 31, 2024, Revised: Feb. 22, 2024, Accepted: Feb. 26, 2024

I. 서론

최근 적대적 생성 신경망(Generative Adversarial Network, GAN)을 포함한 딥 러닝 기술의 발전으로, 실제 현존하는 인물과 유사한 위조 이미지를 생성하는 기술은 놀라운 수준으로 발전하였다. 대표적으로 딥페이크 라고 명명되는 기술은 영상물 또는 사진에 나타나는 인물의 얼굴 영역에 타인의 신체 혹은 가상의 신체를 자연스럽게 합성하는 기술이다. 현재 해당 기술은 실제 인물이 아닌 마치 타인이 행위를 하는 것처럼 합성이 고도화 되었으며, 그 분간이 어려울 정도로 정교한 수준으로 발전되고 있다. 최근, 실제 얼굴과 인공지능으로 합성한 얼굴을 구별하는 실험을 일반인 대상으로 진행하였는데, 실제 얼굴과 인공지능이 합성한 얼굴을 제대로 구별하지 못하며, 되려 합성된 얼굴을 실제처럼 믿는다는 결과를 도출하기도 하였다[1].

딥페이크는 주로 유명 정치인 및 연예인이 대상이 되는 경우가 많으며, 정교하게 제작된 딥페이크 생성물은 미디어 매체 및 소셜 네트워크 서비스 플랫폼을 통해 빠르게 전파되는 경향이 있다. 이는 사회적, 정치적 혼란을 야기하고 개인 혹은 집단의 이미지와 명예를 훼손시키는 등 심각한 피해를 입힐 수 있다. 또한, 시각적으로 표현되는 영상물은 텍스트보다 강력한 영향력을 발휘하는 점에서 딥페이크 기술의 악용은 매우 심각하게 다루어져야 하는 문제이다. 현재, 딥페이크 이미지 및 동영상 합성에 보편적으로 사용되는 방법은 얼굴 교환(FaceSwap) 방법으로, 두 개의 Auto-encoder 를 사용하여 소스 이미지에 포함된 얼굴로 학습한 모델을 타겟 이미지, 즉 합성 대상이 되는 이미지에 교환하고 피부와 같은 세부 텍스처들을 재구성하는 방법을 택한다[2][3]. 이 과정에서, 소스 이미지의 얼굴 특징을 추출하고 타겟 이미지의 얼굴 특징 영역에 전이하는 방법으로 정렬을 수행한다. 그렇기 때문에, 일반적으로 딥페이크로 생성된 이미지는 눈, 코, 입등의 얼굴의 대표적 랜드 마크 근처나 배경과 얼굴이 차분되는 경계선 지점의 텍스처 영역에서 시각적인 부자연스러움이 일어날 수 있다.

따라서 본 논문은 딥페이크 이미지의 텍스처 특성에 기반한 딥페이크 이미지 탐지 기법을 제안한다. 이미지의 텍스처 특성을 추출하기 위해, Superpixels 과 Saliency Map 을 통해 얼굴의 주요 랜드 마크 근처의 합성 후보 영역을 검출한 이후, CNN(Convolution Neural Network) 기반의 ResNet 과 Attention 을 복합 사용하여 딥페이크 이미지를 최종 탐지한다. 제안 방법은 Saliency Map 검출을 통해 딥페이크 이미지에서 합성 의심 영역을 효과적으로 찾아내며, 딥페이크 이미지의 시각적 특징을 더 명확히 파악할 수 있는 효과성을 지닌다. 또한, ResNet 과 Attention 활용을 통해 딥페이크 이미지의 고유 특징을 집중 학습하여, 검출 정확도를 상승시킬 수 있다는 장점이 있다.

본 연구에서 제안하는 기법의 성능을 객관적으로 평가하기 위해, 다양한 딥페이크 이미지 데이터 셋을 대상으로 비교 실험을 진행하였으며, 해당 과정을 통해 제안하는 기법의 성능을 검증하였다. 본 논문의 구성은 다음과 같다. 2 장에서는 딥페이크 데이터 셋 및 딥페이크 탐지 기술에 관련한 기존 연구 사례들에 대해 서술한다. 3 장에서는 제안하는 Saliency 추출과정, 적용된 딥 러닝 모델에 대해 상세하게 설명하며, 4 장에서는 사용된 딥페이크 이미지 데이터 셋, 모델 훈련 상세 내용, 성능 평가 지표를 바탕으로 분석결과를 기술하며, 5 장은 결론, 향후 연구방향성에 대해 제시한다.

II. 관련 연구

컴퓨터 비전 분야에서는 얼굴의 랜드 마크 탐지를 기반으로 하여 다양한 형태의 딥페이크 탐지에 관련된 수많은 연구가 활발히 이루어지고 있다. 딥페이크에 주로 사용되는 FaceSwap 의 경우, 합성의 대상이 되는 소스 이미지와 타겟 이미지의 교환영역을 정확하게 탐지하고 분할해야 하기 때문에, 얼굴 영역의 턱선, 눈썹, 눈, 코, 입등의 랜드 마크 영역를 높은 정확도로 추출하는 연구가 기반이 되었다[4][5]. 딥페이크 생성 과정은 소스 이미지와 타겟 이미지의 주요 랜드 마크 정보를 정확하게 추출한 후, 두 대상간의 얼굴 형태를 비슷하게 정렬하는 작업을 수행하고 마지막으로 VAE(Variational Auto-Encoder)나 GAN 과 같은 생성 모델을 통해 FaceSwap 을 수행하는 형태로 진행되었다. FaceSwap 을 주요 기술로 공개된 데이터 셋은 Celeb-

DF[6], FaceForensic++[7], Real and Fake Face Detection[8]등이 존재한다. 해당 데이터 셋들은 각기 다른 다양한 해상도, 합성 구별 난이도 등의 속성을 가지며 현재 딥페이크 생성 및 탐지 연구분야에서 널리 활용되고 있다.

딥페이크 이미지 탐지 연구분야에서는 CNN(Convolutional Neural Networks) 을 비롯한 VAE, GAN 등이 주로 활용되고 있다. 해당 모델들을 활용한 방법론은 컴퓨터 비전 분야에서 그동안 획기적이고 놀라운 영향력을 보여왔으며, 딥페이크 탐지 및 생성 분야에도 가장 핵심적으로 다루어 지는 모델들이다. 대표적으로 딥페이크 탐지에는 딥 러닝 모델을 활용한 크게 두 가지 방식의 연구가 존재하는 데, 첫 번째로는 딥페이크 이미지 및 딥페이크가 포함된 동영상 구조 자체를 분석하는 연구 방식이 존재하며, 두 번째로는 딥페이크 동영상 내에 포함된 생체 특성의 부자연스러움을 집중하여 탐지하는 연구 방식이 존재한다.

먼저, 이미지 구조 기반의 방식은 딥페이크로 생성된 이미지의 형상적인 구조와 합성 결과의 이상 패턴을 분석하여 이를 학습하고 평가하는 방식으로 연구가 진행되었는데, CNN 기반의 딥페이크 탐지 기술[9], GAN 등의 생성 모델이 가짜 영상을 생성할 때 생기는 이상패턴을 이용하여 딥페이크를 감지하는 기술[10], 딥페이크 이미지의 PRNU(Photo Response Non-Uniformity) 패턴을 바탕으로 GAN 모델이 가지는 특이 구조를 활용하는 검출 기술[11] 등이 존재한다. 같은 맥락으로, 딥 러닝 네트워크 내의 핑거프린트를 활용하는 기술[12], 딥페이크 이미지를 주파수 도메인으로 변환하여 딥페이크 이미지의 스펙트럼을 분석하는 기술 등도 함께 진행되었다[13]. 얼굴 내의 눈, 코, 입 등의 부자연스러운 방향성을 통해 딥페이크 이미지를 검출하는 방법[14], 눈동자의 홍채 색상 특징 변화를 활용한 검출 방법[15], 심박수 및 눈 깜박임등의 생리적 특징을 이용하여 검출하는 방법[16], 영상내 사람의 혈류의 흐름을 관찰하고 광혈류 측정 기반으로 특징을 분석하는 방법[17]등이 존재한다. 이러한 연구들은 기본적으로 딥페이크 영상 내의 부자연스러운 움직임을 감지하는 것에 초점을 둔다.

III. 본론

3.1 제안기법

본 논문에서 제안하는 딥페이크 탐지 모델의 아키텍처는 '그림 1' 과 같다. 생성된 딥페이크 이미지내의 세부적인 텍스처에 기반하여 부자연스럽게 생성된 영역을 찾고, 해당 영역 위주로 딥 러닝 모델을 학습시키는 방법을 택하였다. 먼저 입력된 이미지에서 Superpixels 기반으로 유사 특성을 지니는 픽셀들을 그룹화 한 뒤, Saliency Map 검출을 통해, 영상의 합성 후보 영역을 추출하는 과정을 거친다. 여기서 Saliency Map 은 GAN 등을 통해 생성된 딥페이크 이미지의 텍스처가 부자연스러운 부분을 포착하며, 이는 곧 딥페이크 이미지가 가지고 있는 잠재적인 텍스처 특성과 같다고 할 수 있다. 이후, ResNet 을 활용하여 특징을 추출하고, 주요 랜드 마크에 Attention 을 결합한 결과를 활용하여 최종 딥페이크 이미지 탐지를 수행하는 방식을 취한다.

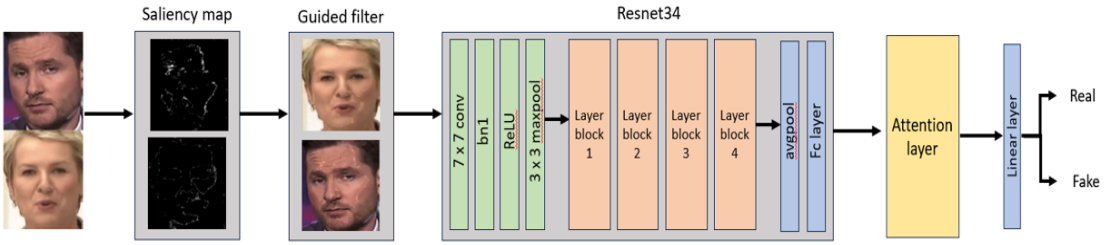


Figure 1. Model Architecture
그림 1. 모델 아키텍처

3.2 딥페이크 탐지를 위한 이미지 전처리

딥페이크 탐지를 위해 딥 러닝 모델을 활용하기 전, 합성 후보영역을 검출하는 과정을

수행한다. 딥페이크 이미지의 경우, 사람이 육안으로 확인할 때, 직관적으로 실제 이미지와는 텍스처 부분에서 미세한 시각적 차이가 발생한다[18]. 이러한 발생차이를 감지하기 위해, SLIC(Simple Linear Iterative Clustering) 알고리즘을 적용하고, 이를 기반으로 하여 Saliency Map 을 구성한다[19]. 먼저, 이미지 처리 과정의 복잡성과 계산량을 줄이면서 텍스처 정보를 효율적으로 구성하기 위해 SLIC 알고리즘을 사용하였다. 주어진 이미지의 색상 및 위치 정보를 이용하여 이미지 영역 분할을 수행하며, 세분화 시키는 과정은 아래와 같다. 식 (1) 은 SLIC 알고리즘에 사용된 거리 계산방법으로, Superpixels 클러스터링을 위한 지표를 제공한다. d_c는 색상 공간에서의 거리, d_{space}는 이미지 내, $\mathrm{x,y}$좌표 공간에서의 거리, S는 클러스터 크기에 대한 정규화 인자를 나타낸다.

$$D = \sqrt{d_c^2 + (\frac{d_{space}}{S})^2 \cdot m^2} \tag{1}$$

m은 SLIC 에서 조밀도(Compactness)를 결정하는 것으로, 색상 유사성과 공간적 근접성 사이의 상대적 중요성을 조절한다. m값이 크면 공간적 근접성이 더 중요해져 고밀도의 Superpixels 가 생성되고, 반대의 경우는 저밀도의 Superpixels 가 생성된다. D 는 클러스터 중심과 개별 픽셀 간의 거리 지표를 나타낸다. D 는 클러스터 중심과 개별 픽셀 간의 최종 유사성을 측정하는 지표로, 색상과 공간적 정보를 모두 고려하여 효율적인 Superpixels 분할을 가능하게 한다. '그림 2'는 조밀도에 따른 Superpixels 분할 예제를 나타낸다. 딥페이크 탐지를 위한 이미지 전처리를 위해 본 연구에서는 분할 개수를 기본 350, 조밀도를 3 으로 설정하였다.

Superpixels 분할 이후, Saliency Map 생성을 위해 전경 기반의 이미지 특징 중요도 맵을 생성한다. 이미지의 전경은 Saliency Map 을 구성할 때 매우 중요한 영향을 미치는데, 사람이 사물을 인지하는 방법이 Pixelwise 하지 않으며, 사람이 사물을 인지할 때 방향성 있게 차례로 바라보지 않고 눈에 띄는 영역을 먼저 인지하여 집중하기 때문이다[20]. 집중하는 영역은 화면에서 가장 눈에 띄는 영역으로, 픽셀 값의 변화가 급격한 부분들을 모아서 매핑하는 방법을 취한다. 따라서, 본 논문에서는 전경 중요 영역인 Foreground-Saliency-Map(FSM)와 지역 중요 영역인 Local-Saliency-Map(LSM)를 각각 추출하여 전처리 과정에 반영한다. 먼저, dlib 라이브러리를 사용하여 얼굴을 감지하고 전경 영역의 분할 시드로 사용한다. 전경 시드 주변의 픽셀에 높은 중요도를 할당하고, 멀어질수록 중요도를 감소시키는 방식으로 식 (2)와 같이 FSM 을 구성한다. $\mathrm{I(x,y)}$는 위치 $\mathrm{x,y}$에서의 픽셀 중요도를 나타낸다. $\mathrm{G(x,y)}$는 그레이 스케일 이미지에서 위치 $\mathrm{x,y}$의 픽셀 밝기 값이다. μ는 전경 픽셀의 밝기 평균값이며, σ는 가우시안 분포의 표준 편차를 나타낸다. σ는 FMS 의 평활화 정도를 조절한다.

$$I(x,y) = exp\,(-\frac{1}{2}(\frac{G(x,y) - \mu}{\sigma})^2) \tag{2}$$

식 (2)와 같이, 주변의 픽셀 밝기 값이 전경 시드의 평균 밝기 값에 가까울수록 중요도가 높아지며, 반대로 전경 시드에서 멀어질수록, 즉 밝기 값의 차이가 클수록 중요도가 낮아진다. 이를 통해 전경에 가까운 픽셀들은 강조되고, 배경에 가까운 픽셀들은 상대적으로 감소된 중요도를 갖게 되는 $\mathrm{FSM}(x,y)$ 를 얻는다. LSM 은 가우시안 블러를 적용해 전경과의 밝기 차이에 따른 중요도를 계산한다. 전경과 유사한 픽셀은 높은 중요도를, 전경과 값 차이가 큰 픽셀은 낮은 중요도를 갖게 된다. 전경 시드와 각 픽셀 위치 $\mathrm{x,y}$를 활용하고 각 픽셀 별로 가우시안 블러를 적용하여 재구성된 $LSM(x,y)$ 을 얻는다.

최종 산출된 Saliency Map(SM)은 식 (3)과 같이, FSM 과 LSM 의 가중 평균을 계산하여 생성된다. 각각의 맵의 특성에 따라 차등 반영하며, 이미지에서 시각적으로 포착되는 부분을 포괄적으로 강조할 수 있으며 여기서 α는 Saliency Map 생성을 위한 0 과 1 사이의 가중치 상수를 의미한다.

$$SM(x,y) = \frac{\alpha \times FSM(x,y) + (1-\alpha) \times LSM(x,y)}{2} \tag{3}$$

Saliency Map 을 통해 최종 전처리 이미지인 Guided Filter 를 생성한다. Guided Filter 는 원본 이미지의 중요한 특징이나 경계선을 강조하는 역할을 한다. 식 (4)는 최종 생성된 Guided Filter 이미지의 픽셀 $G_{x,y}$로, 원본 이미지의 각 픽셀 값인$P(x,y)$, Saliency Map 의 해당 픽셀 값인 $SM(x,y)$의 가중 합산으로 표현할 수 있다. 여기서 β는 Guided Filter 를 생성을 위한, 0 과 1 사이의 가중치 상수를 의미한다.

$$G(x,y) = \beta \times P(x,y) + (1-\beta) \times SM(x,y) \tag{4}$$

Figure 2. SLIC Superpixels Segmentation on Sample Images
그림 2. SLIC 알고리즘을 이용한 이미지의 슈퍼픽셀 분할

3.3 딥페이크 탐지를 위한 Attention 기반 딥 러닝 모델

딥페이크 탐지를 보다 정밀하게 수행하기 위해, ResNet 과 Attention Layer 를 복합 사용하였다. Attention Layer 는 모델이 입력 이미지 내 중요한 특성에 우선 순위를 두고 집중할 수 있도록 하며, 입력 특징 맵에서 중요도에 따른 가중치 분포를 조정한다. '그림 1'과 같이, ResNet34 기본 구조에 Attention Layer 을 추가하여 각 채널 별로 중요한 특징을 강조하는 가중치를 계산하고 적용하였다. 가중치 계산을 위해, dlib 라이브러리의 검출된 랜드 마크 정보를 사용하였으며 주요 특징이 있는 영역에 높은 가중치를 부여한다. 가중치의 집합을 의미하는 가중치 마스크는 랜드 마크 중심에서 멀어질수록 가중치가 감소하는 방식으로 계산되며, 식 (5)를 이용해 각 랜드 마크 $\boldsymbol{l}$에 대한 가중치를 얻는다. 여기서 $\boldsymbol{i,j}$는 이미지 상의 픽셀 위치를 나타낸다[21].

$$\boldsymbol{M_l(i,j) = \exp\left(-\frac{(i-i_l)^2 + (j-j_l)^2}{2\sigma_l^2}\right)} \tag{5}$$

최종 가중치 마스크 $\boldsymbol{M_{landmark}}$는 dlib[22]을 통해 검출된 모든 랜드 마크에 대해 계산된 가중치 $\boldsymbol{M_l}$의 합으로 정의된다. $\boldsymbol{M_{landmark}}$는 Attention Map $\boldsymbol{A}$와 사용되어 출력 특징 맵 $\boldsymbol{O}$를 계산하는데 사용된다. Attention Map $\boldsymbol{A}$는 Query Map $\boldsymbol{Q}$와 Key Map $\boldsymbol{K^T}$ 의 행렬 곱으로 계산되며, 각 채널의 중요도를 나타내는 스코어가 된다. 출력 $\boldsymbol{O}$는 최종 가중치 마스크 $\boldsymbol{M_{landmark}}$를 적용하여 중요 특징이 강조된 특징 맵을 생성한다. 학습을 위해 모델의 배치 크기를 64, 최적화는 Adam(Adaptive Moment Estimation)을 사용하고, Epoch 은 기본 250 을 사용하였으며, 학습률은 0.001 로 설정하였다. 딥페이크 이미지 탐지는 결과가 Real 과 Fake 두 가지 라벨로 분류되기 때문에, 손실 함수로는 교차 엔트로피를 사용하였다.

3.4 딥페이크 탐지 영역 활성화

추가적으로, 딥페이크 탐지 근거에 대한 시각적 활성화를 위해 GradCAM[23]을 사용하였다.

GradCAM 은 CNN 등의 딥 러닝 모델에서 관심 대상 클래스 예측을 위해 추론된 중요한 영역을 시각적으로 나타내기 위한 기술이며, Gradient 를 활용하여 예측결과 에 대한 원인을 직관적으로 해석해 볼 수 있다. 관찰 대상이 되는 CNN 층의 특징 맵의 채널에 대한 평균을 구하고, 각 특징 맵 채널의 중요도를 산출하는 방식으로 진행되며, 본 연구에서는 ResNet 의 마지막 Layer 블록에서의 컨볼루션 레이어의 특징 맵을 대상으로 하였으며, 특징 맵들의 가중 평균 합을 이용하고 ReLU 를 적용하여 Real/Fake 라벨에 대한 영향 정도를 판단하였다. '그림 3'은 활성화 영역을 나타내기 위해 구성된 GradCAM 산출 과정을 나타낸다.

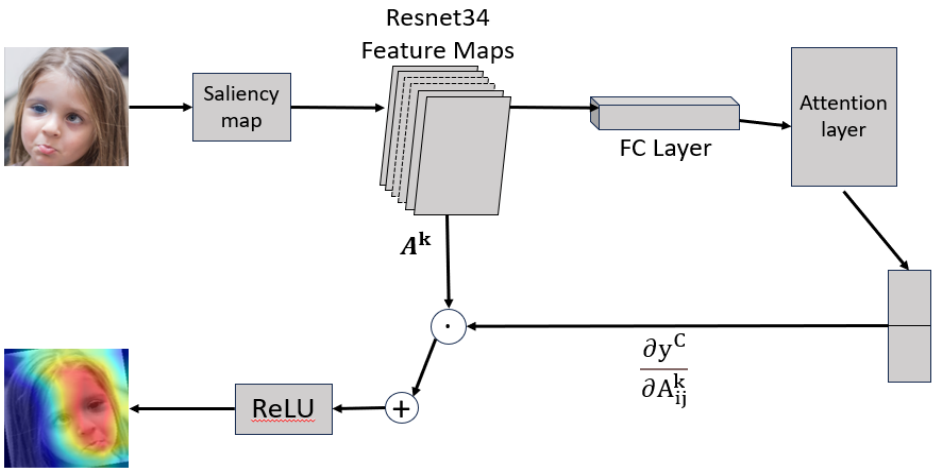


Figure 3. Application Process of GradCAM
그림 3. GradCAM 적용 과정

IV. 실험 결과 및 평가

4.1 딥페이크 탐지를 위한 데이터 셋

본 연구에서 제안한 딥페이크 탐지 모델의 성능 평가를 위해서, 가장 많이 활용 되고 있는 대표적인 두 가지 딥페이크 데이터 셋인 'Real and Fake Face Detection' 데이터 셋(RFFD)와 'FaceForensics++' 데이터 셋(FF++)을 사용하였다. RFFD 는 총 8,000 개의 이미지로 구성되어 있으며, 전문가가 변조,합성한 얼굴 이미지와 변조,합성하지 않은 얼굴 이미지로 구성되어 있다. 특히, 합성 이미지의 경우 난이도가 크게 3 가지로 구성되며, 'easy', 'mid', 'hard'로 나뉘어져 있다. '그림 4'의 우측과 같이, 'mid', 'hard' 그룹으로 구성된 이미지들은 일반적으로 육안 구분이 힘든 것이 특징이다. 두 번째로 사용된 FF++ 데이터 셋은 977 개의 유튜브 비디오에서 추출된 1000 개의 원본 시퀀스와 해당 이미지의 조작 버전을 포함하고 있다. FF++ 데이터 셋은 Face2Face, FaceSwap 을 사용해서 생성되었고, DeepFakes 와 NerualTexture 를 사용하여 학습된 결과이다. 해당 학습 및 생성 기법이 적용된 약 3,000 개 이상의 데이터 셋으로 성능 평가를 진행하였다. 기준 데이터 셋 기반으로, 데이터 증강은 RandomHorizontalFlip, RandomRotation, ColorJitter 를 사용하여 각도, 회전, 밝기, 대비, 채도의 변동성을 추가하여 모델이 강건한 학습을 할 수 있도록 하였다. 또한, 딥페이크 이미지 탐지 모델의 학습을 위해, Intel Xeon Sliver 4125 8 Core, Quadro RTX 8000 를 사용하였다.

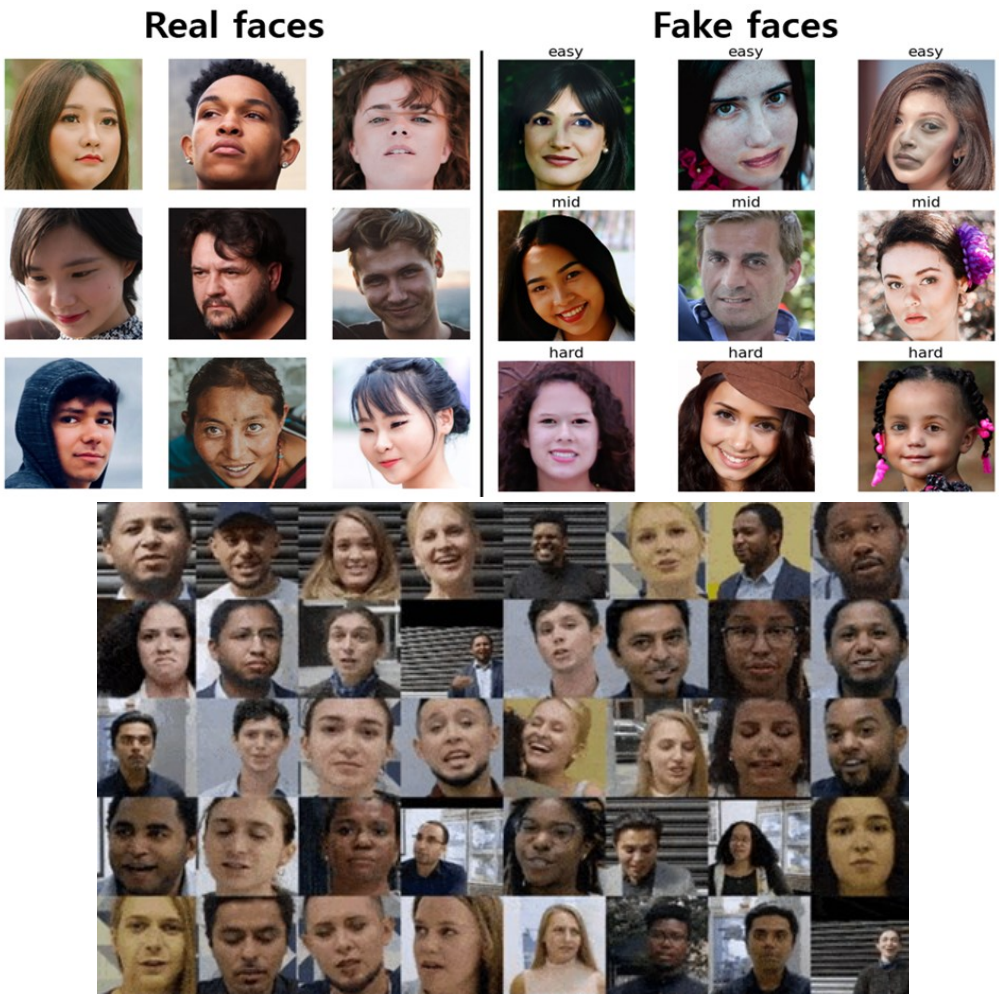


Figure 4. Examples of RFFD and FF++ Dataset
그림 4. RFFD 와 FF++ 데이터 셋 예시

4.2 이미지 전처리 결과

SLIC 알고리즘과 Saliency Map 을 통해 원본 이미지와 딥페이크를 통해 생성된 조작된 이미지 간의 차이를 시각화하였다. 상단의 원본 이미지의 경우, FSM 과 LSM 그리고 Saliency Map 에서 얼굴의 광대, 안가 주변, 볼 영역 등에서 잠재적 시각 특징이 활성화 된 것을 확인 할 수 있다. 중간 이미지와 아래의 이미지의 경우, RFFD 에 포함된 딥페이크 이미지와 Face2Face 로 생성된 이미지를 나타낸다. 딥페이크의 경우, Saliency Map 이 턱과 귀 사이의 경계부근, 하관에 집중되어 나타난 것을 확인할 수 있었고, Face2Face 의 경우는 얼굴의 하부에서 매우 국소적으로 특징이 검출되는 것을 '그림 5' 와 같이 확인 할 수 있다. 이는 Saliency Map 을 통해 진짜와 가짜 이미지 사이에서 발견할 수 있는 세부적인 텍스처 차이가 존재함을 확인할 수 있다.

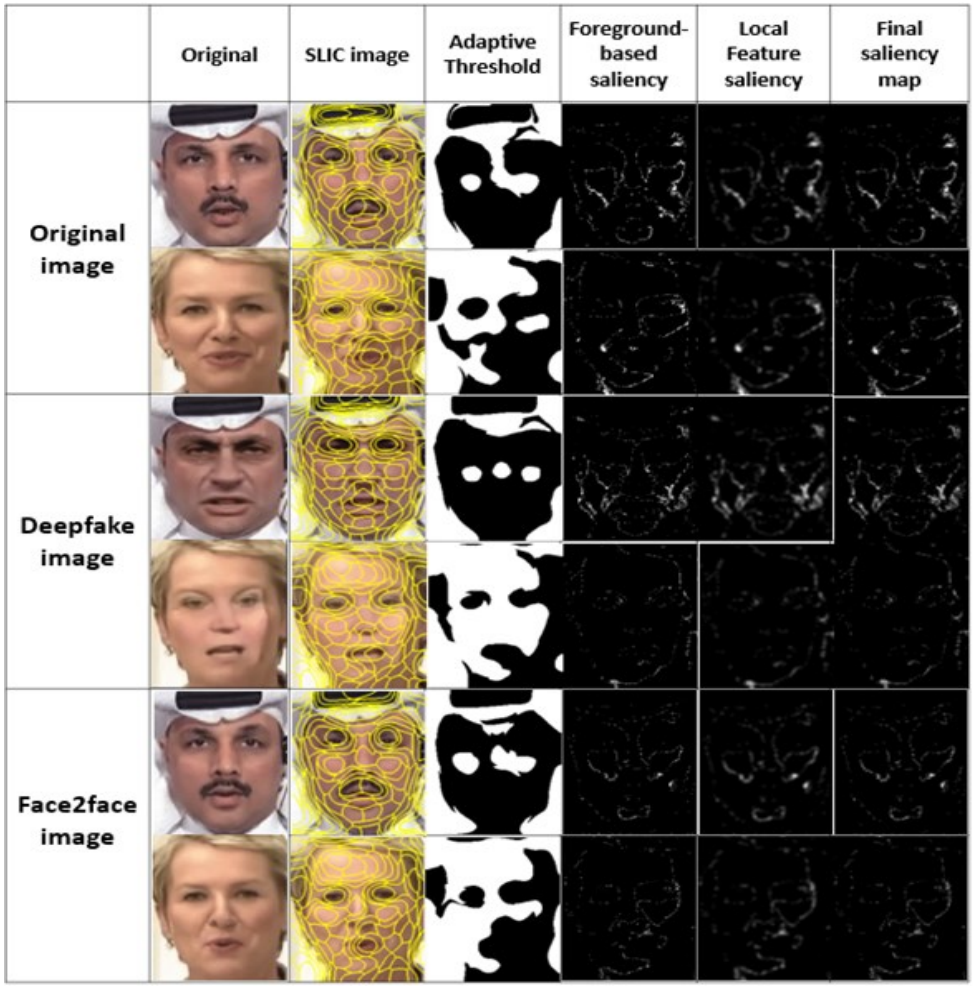
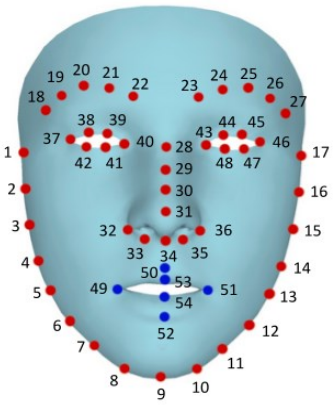


Figure 5. Results of Extracted Saliency Map
그림 5. 추출된 Saliency Map 결과

4.3 Saliency Map 분석 및 합성 후보 영역 검출

4.2 에서 추출한 Saliency Map 정보와 dlib 라이브러리를 통해 검출된 랜드 마크와의 관계성을 확인하기 위해, 얼굴 영역별 활성화 정도에 대한 분석을 수행하였다. '그림 6' 은 dlib 라이브러리에서 추출 가능한 얼굴의 랜드 마크 위치, 위치에 따라 매핑 되는 얼굴영역을 나타낸다. '그림 7'은 전체 딥페이크 데이터 세트 상에서 추출된 Saliency Map 특징 영역이 얼마나 랜드 마크 위에서 활성화 되었는지에 대한 통계적인 분포를 의미한다. '그림 7' 에서 가로 축은 1~68 번의 랜드 마크를 의미하고 세로 축은 전체 중 해당 랜드 마크의 활성화 백분율을 나타낸다. 랜드 마크 활성화의 통계적 분포를 분석한 결과, 턱 라인(1-17)과 입 주변 랜드 마크(48-67)가 타 랜드 마크에 비해 상대적으로 높은 활성화 비율을 나타내었다. 이는 턱 라인과 입 주변의 랜드 마크가 딥페이크 이미지 탐지에 상대적으로 중요한 합성 후보 영역이 될 수 있음을 나타낸다.



Face Landmark	Landmark Number
Right Eye	37 to 42
Left Eye	43 to 48
Mouth	49 to 67
Nose	28 to 36
Eyebrows	18 to 27
Jawline	1 to 17
Eyes (both)	37 to 48

Figure 6. Dlib Landmark
그림 6. Dlib 랜드 마크

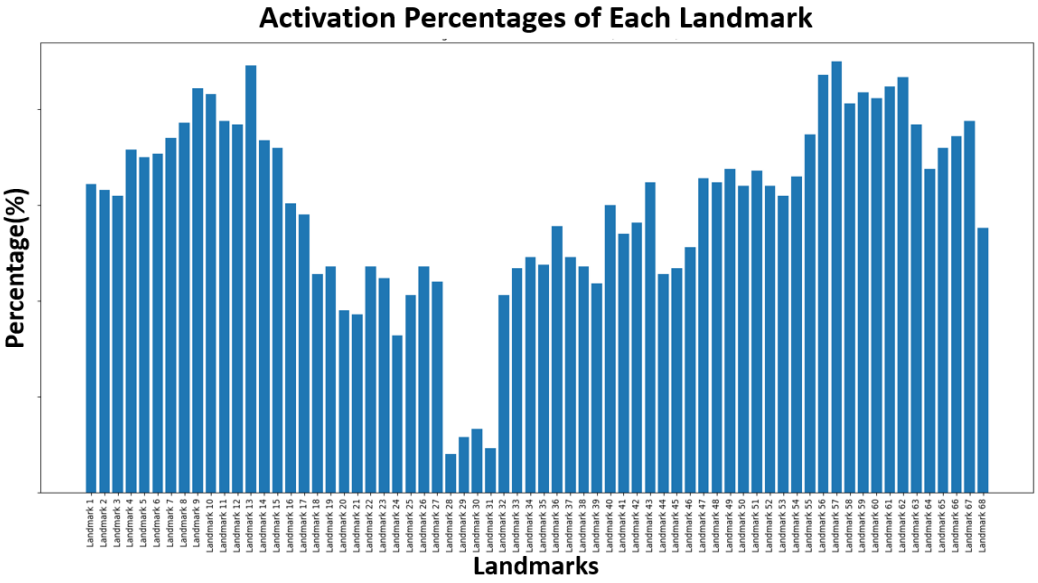


Figure 7. Percentages of Landmark Activation in Saliency Map.
그림 7. Saliency Map 에서 랜드 마크 활성화 정도.

4.4 딥페이크 탐지 성능 및 GradCAM 시각화

본 연구의 제안된 모델을 통해 딥페이크 탐지 성능에 대한 평가를 진행하였다. '표 1'은 테스트 셋에 대한 탐지 예시이다. Real/Fake 클래스를 정확하게 예측한 경우는 초록색으로, 그렇지 못한 경우는 빨간색으로 나타내었다. Saliency Map 이 적용되지 않은 Resnet18 과 VGG16 에서 Real 클래스 구분하지 못하는 경우가 발생하였음을 확인하였고, Fake 클래스의 경우, 제안된 모델 이외의 비교 모델들은 정확한 탐지가 대부분 실패한 경우를 보여주었다. '표 2'은 비교 모델인 ResNet18, ResNet34, MobileNetV2, VGG16 을 Saliency Map 과정을 제외한 경우와 포함한 경우를 포함하여 최종 결과를 나타내었다. 제안하는 모델은 RFFD 와 FF++ 두 가지 데이터 셋에서 상대적으로 높은 딥페이크 감지 성능을 보여주었다. 특히, RFFD 데이터 셋의 mid, hard 그룹에서 각각 약 94%와 95%로 높은 탐지 성능을 보여주었으며, 이는 육안으로 감지 하기 힘든 딥페이크 이미지의 경우도, Saliency Map 이 효과적으로 적용 될 수 있음을 나타낸다. 이에 대한 탐지 예시는 '표 3'에도 나타내었다.

Table 1. Deepfake Detection of Test Images
표 1. 딥페이크 이미지 탐지 결과

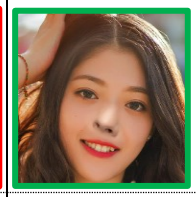
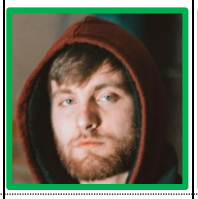
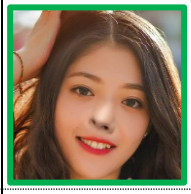
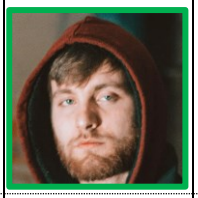
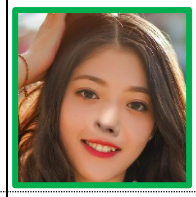
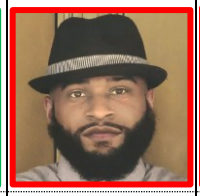
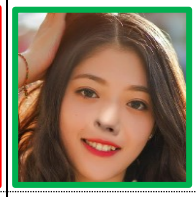
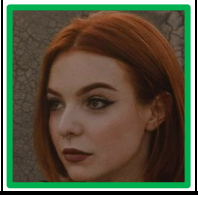
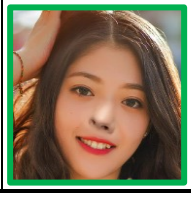
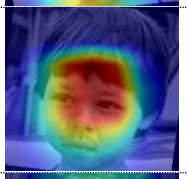
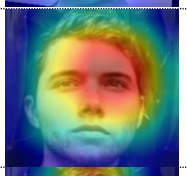
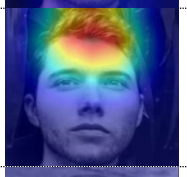
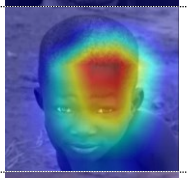
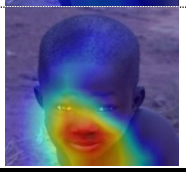
	Real Image	**Real Image**	**Fake Image (Easy)**	**Fake Image (Mid)**	**Fake Image (Hard)**
Resnet18 (W/O Saliency Map)					
Resnet34 (W/O Saliency Map)					
MobileNetV2 (W/O Saliency Map)					
VGG16 (W/O Saliency Map)					
Proposed Model (Saliency Map + Attention)					

Table 2. Performance of Deepfake Detection
표 2. 딥페이크 탐지 성능

Model	**RFFD**	**FF++**	**Total**
Resnet18 (W/O Saliency Map)	0.8552	0.8343	0.8447
Resnet34 (W/O Saliency Map)	0.9085	0.9315	0.9200
MobileNetV2 (W/O Saliency Map)	0.9284	0.9222	0.9253
VGG16 (W/O Saliency Map)	0.9247	0.9113	0.918
Proposed Model (W/O Saliency Map + Attention)	0.9328	0.9418	0.9373
Proposed Model (Saliency Map + Attention)	**0. 9438**	**0.9633**	**0.9535**

Table 3. Visualization of GradCAM
표 3. GradCAM 의 시각화 결과

	Lable (Hardness)	Input Image	GradCAM
Proposed Model (Saliency Map + Attention)	Fake (Easy)		
Resnet34 (W/O Saliency Map W/O Attention)			
Proposed Model (Saliency Map + Attention)	Fake (Mid)		
Resnet34 (W/O Saliency Map W/O Attention)			
Proposed Model (Saliency Map + Attention)	Fake (Hard)		
Resnet34 (W/O Saliency Map W/O Attention)			

‘표 3’은 GradCAM 을 통해 특징 맵을 시각화한 결과를 나타낸다. 상단의 ‘easy’ 의 경우, 아이의 얼굴에 부자연스러운 수염이 합성된 모습을 볼 수 있는데 GradCAM 을 이용해 활성화 영역을 추론해본 결과, 제안한 모델이 합성 영역을 정확하게 포괄한 형태를 관찰 할 수 있다. ‘mid’의 경우, 입력 이미지의 양쪽 안구가 다른점이 주요 특징인데 제안한 모델은 이러한 영역을 잘 포착하고 있음을 보여준다. 마지막으로 ‘hard’의 경우, 육안으로는 확인이 불가능하지만, 제안한 모델이 이마 우측 상단을 활성화 한 것으로 보아, 얼굴의 상단영역에서 합성이 일어날 것으로 추정하고 있었다. 예시를 통해, 제안하는 모델은 시각적인 부자연스러움이나 텍스처의 변화가 있는 부분에 더욱 포착을 잘 하는 모습을 확인 할 수 있었으며, 실험 비교 대상인 다른 딥 러닝 모델보다 향상된 추론 능력을 가지고 있음을 확인 할 수 있다.

IV. 결론

본 연구를 통해, 최근 가장 사회적으로 문제가 되는 이슈인 딥페이크에 대한 대안책인 이미지 탐지 모델을 제안할 수 있었으며, 다양한 실험을 통해 우수한 감지 능력을 확인 할 수 있었다. Visual Saliency 는 인간의 시각적 능력 모방을 참고한 것으로, SLIC 기반의 Superpixels 분할과 전경, 후경의 시각적 인지 중요도를 토대로 구현이 가능하였다. 최종 도출된 Saliency Map 은 딥페이크 이미지의 합성 특성을 추출하거나 분류하는데 효과적으로 적용 될 수 있을 것으로

사료된다. 또한, 제안된 모델은 이러한 합성 특성에 대해 더욱 효과적으로 작용 될 수 있는 추가적인 Attention Layer 을 적용하였고, 특성 맵 활성화 방법인 GradCAM 을 통해 추론 능력을 시각화 하였다. 시각화 결과, 제안한 논문의 구조를 통해 추출된 활성화 영역은 얼굴 형태 전이 및 FaceSwap 등의 방법론으로 이루어진 딥페이크 이미지 감지에 특화된 추론 능력을 보여주었다.

현재의 딥페이크 생성 기술은 매우 빠르게 발전하고 있으며, 새로운 방법과 기술이 계속해서 등장하고 있다. 제안한 방법은 새로운 합성 기술로 의해 생성된 이미지에 대해 동일한 수준의 탐지 성능을 보장하기 어려울 수 있다. 따라서, 향후 연구에서는 제안된 방법을 기반으로 하여, 더 다양한 형태의 딥페이크와 그 생성 기술에 대응할 수 있도록 모델의 적용 범위를 확대하고, 동영상 내에서의 딥페이크 탐지, Diffusion 기반의 딥페이크 이미지 탐지 등 보다 고도화된 연구를 진행할 예정이다.

VI. 참고문헌

[1] S. J. Nightingale, H. Farid, “AI-synthesized faces are indistinguishable from real faces and more trustworthy,” Proceedings of the National Academy of Sciences, Vol. 119, No. 8, e2120481119, Feb, 2022.

[2] I. Korshunova, W. Shi, J. Dambre, and L. Theis, “Fast face-swap using convolutional neural networks,” in Proc. of the IEEE International Conference on Computer Vision(ICCV), Venice, Italy, 2017, pp. 3677-3685.

[3] A. A. Maksutov, V. O. Morozov, A. A. Lavrenov, and A. S. Smirnov, “Methods of deepfake detection based on machine learning,” in 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), St. Petersburg and Moscow, Russia, 2020, pp. 408-411.

[4] H. Kim, H. Kim, J. Rew and E. Hwang, “FLSNet: Robust facial landmark semantic segmentation,” IEEE Access, Vol. 8, pp. 116163-116175, June, 2020.

[5] H. Kim, H. Kim, S. Rho, and E. Hwang, “Augmented EMTCNN: A fast and accurate facial landmark detection network,” Applied Sciences, Vol. 10, No. 7, pp. 2253, March, 2020.

[6] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-df: A large-scale challenging dataset for deepfake forensics,” in Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Seattle, WA, 2020, pp. 3207-3216.

[7] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in Proc. of the IEEE/CVF International Conference on Computer Vision(ICCV), Seoul, Korea, 2019, pp. 1-11.

[8] S. Nam, S. Oh, J. Kang, C. Shin, Y. Jo, Y. Kim, K. Kim, M. Shim, S. Lee, Y. Kim, S. Han, G. Nam, D. Lee, S. Jeon, I. Cho, W. Cho, S. Yang, D. Kim, H. Kang, S. Hwang, and S. Kim, (2019, Jan.). Real and Fake Face Detection [Online], Available: https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection.

[9] B. Zi, M.Chang, J. Chen, X. Ma, and Y. G. Jiang, “Wilddeepfake: A challenging real-world dataset for deepfake detection,” in Proc. of the 28th ACM International Conference on Multimedia, Seattle, USA, 2020, pp. 2382-2390.

[10] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, “Face x-ray for more general face forgery detection,” in Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Seattle, WA, 2020, pp. 5001-5010.

[11] F. Lugstein, S. Baier, G. Bachinger, and A. Uhl, “PRNU-based deepfake detection,” in Proc. of the 2021 ACM Workshop on Information Hiding and Multimedia Security(IH&MMSec), New York, USA, 2021, pp. 7-12.

[12] T. Yang, Z. Huang, J. Cao, L. Li, and X. Li, “Deepfake network architecture attribution,” in Proc. of the AAAI Conference on Artificial Intelligence(AAAI), Vol. 36, No. 4, pp. 4662-4670, June, 2022.

[13] R. Durall, M. Keuper, F. J. Pfreundt, and J. Keuper, “Unmasking deepfakes with simple features,” arXiv preprint arXiv:1911.00686, 2019.

[14] X. Yang, Y. Li, and S. Lyu. “Exposing deep fakes using inconsistent head poses,” In ICASSP

2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 8261-8265.
[15] F. Matern, C. Riess, and M. Stamminger, “Exploiting visual artifacts to expose deepfakes and face manipulations,” In 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 2019, pp. 83-92.
[16] C. M. Liy, and L. Y. U. S. InIctuOculi, “Exposing ai created fakevideos by detecting eye blinking,” in Proc. of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 2018, pp. 11-13.
[17] U. A. Ciftci, I. Demir, and L. Yin, (2020, September). “How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals,” In 2020 IEEE International Joint Conference on Biometrics (IJCB), Online, 2020, pp. 1-10.
[18] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “Deepfakes and beyond: A survey of face manipulation and fake detection,” Information Fusion, Vol. 64, pp. 131-148, Dec, 2020.
[19] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, “SLIC superpixels compared to state-of-the-art superpixel methods,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 34, No. 11, pp. 2274-2282, 2012.
[20] G. Li, and Y. Yu, “Visual saliency based on multiscale deep features,” in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, 2015, pp. 5455-5463.
[21] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” in Proc. of the European Conference on Computer Vision(ECCV), Zurich, Switzerland, 2014, pp. 818-833.
[22] S. Suwarno, and K. Kevin, “Analysis of face recognition algorithm: Dlib and opencv,” Journal of Informatics and Telecommunication Engineering, Vol. 4, No. 1, pp. 173-184, 2020.
[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in Proc. of the IEEE International Conference on Computer Vision(ICCV), Venice, Italy, 2017, pp. 618-626.

저자소개

노하림(*Harim Noh*)

2021 년 3 월~현재 덕성여자대학교 컴퓨터공학전공 학사과정

관심분야 : 컴퓨터비전, 패턴인식, 인공지능

유제혁(*Jehyeok Rew*)

2020 년 8 월 고려대학교 대학원 전기전자컴퓨터공학과 (공학박사)
2021 년 3 월 (주)기아 책임매니저
2023 년 3 월~현재 덕성여자대학교 데이터사이언스학과 조교수

관심분야 : 빅데이터, 패턴인식, 인공지능