Journal of the Korean Institute of Industrial Engineers
Vol. 51, No. 3, pp. 266-275, June 2025.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2025.51.3.266

설명가능 인공지능 기반 중요 얼굴 영역 탐색을 통한 효율적인 FAS(Face Anti-Spoofing) 모델 구축

한태혁[1,2] · 정준각[1†]

[1]한양대학교 산업융합학부 / [2]성균관대학교 데이터사이언스융합학부

Building an Efficient Face Anti-Spoofing Model with the Exploration of Important Face Areas Based on Explainable AI

Taehyeok Han[1,2] · Junegak Joung[1]

[1]School of Interdisciplinary Industrial Studies, Hanyang University
[2]Department of Applied Data Science, Sungkyunkwan University

Compared to other biometric methods, facial recognition is relatively slow and vulnerable. To address this issue, the development of fast and accurate Face Anti-Spoofing (FAS) models is essential. In this study, we propose an explainable neural network-based approach that leverages important facial areas to construct an efficient FAS model. These important areas are quantitatively identified by analyzing the prediction mechanism of the FAS model. To validate the proposed approach, we train a new model using images that include only the identified important areas and compare its performance to that of the traditional model. The results demonstrate that the performance of the two models is comparable, indicating the feasibility of replacing existing models. Additionally, the proposed method reduces computational overhead by pre-removing irrelevant areas, enabling the construction of an efficient FAS model that focuses on learning from the key facial areas.

Keywords: Grad-CAM++, Face Recognition, Facial Landmark, CNN, AI-Hub, XAI

1. 서 론

오늘날 생체인식은 여러 분야에서 활용되고 있다. 그중에서도 얼굴 인식은 휴대기기의 암호화 수단이나 인터넷 뱅킹의 개인 인증수단으로 활용되는 등 기술의 시장규모가 커지고 있으며, 그에 따라 정확하면서 쉽고 빠르게 학습 가능한 시스템의 개발이 요구되고 있다. 그러나 사용자의 얼굴을 다른 매체에 복사하여 본인인증을 시도할 시 인증에 성공하는 등 얼굴 인식 기술이 보안에 상대적으로 취약하다는 문제점이 최근 몇 년간 지속적으로 제기되고 있다. 이렇게 위조된 데이터를 사용하여 다른 사람이나 시스템을 속이는 행위를 스푸핑(Spoofing)이라고 하며, 최근 페이스 스푸핑 공격을 방지하기 위한 다양한 연구가 이루어져 왔다. 하지만 기술이 고도화됨에 따라 스푸핑 공격의 방법과 범위도 예측할 수 없을 정도로 정교하게 발전하고 있으며, 이러한 위조 데이터를 사람이 전부 추출하기에는 어려움이 있다. 이러한 문제를 해결하기 위해 본 연구에서는 설명 가능한 인공지능(eXplainable Artificial Intelligence, XAI) 기법을 활용한 효율적인 FAS(Face Anti-Spoofing, FAS) 기술을 제시하고자 한다.

지금까지 보고된 FAS 기술 관련 연구에서의 한계는 다음과 같다. 얼굴 인식에 사용되는 이미지를 영역별 구분 없이 전체 이미지를 사용했다는 점이다. 얼굴 인식에 중요한 영향을 주

이 논문은 한양대학교 교내연구지원사업과 한국연구재단의 지원을 받아 수행되었음 (HY-202300000003614, NRF-RS-2024-00344286).
† 연락저자 : 정준각 조교수, 서울특별시 성동구 왕십리로 222, 제2공학관 503-2호 한양대학교 산업융합학부, Tel : 02-2220-2363, Fax : 02-2220-2363, E-mail : june30@hanyang.ac.kr
2025년 1월 22일 접수; 2025년 3월 14일 수정본 접수; 2025년 3월 20일 게재 확정.

는 영역이 어디인지 고려하지 않고 연구를 진행했기 때문에 시스템 구현에 필요한 이미지 데이터의 용량이 굉장히 커지며, 속도도 상대적으로 느리다는 한계가 있다.

기존 연구 대비 본 연구의 기여는 FAS 모델에 중요한 얼굴 영역을 탐색하는 방법에 있다. Dlib 라이브러리에서 제공하는 68개의 특징점을 추출하는 Face Landmarks를 통해 얼굴 영역을 검출한다. 사람마다 얼굴의 크기와 생김새가 다르므로 이를 정형화된 데이터로 변환하기에 어려움이 있지만, Face Landmarks를 통해 각각의 얼굴 이미지를 수작업으로 확인하지 않고도 일관된 데이터로 구축할 수 있다. 그 후, 얼굴 중요 영역은 CNN(Convolutional Neural Network)과 Grad-CAM++(Gradient-weighted Class Activation Mapping++) 설명에 기반하여 추정한다. CNN은 이미지 분류에 보편적으로 사용되는 AI 모델로서 복잡한 작동방식으로 인해 설명 불가능한 블랙박스 모델로 간주하지만, 본 연구에서는 XAI 기법의 한 종류인 Grad-CAM++을 사용하여 결과를 설명한다.

본 연구는 (주)딥핑소스에서 제공하는 RGB 기반 안면 위변조 감지(Anti-Spoofing) AI 모델을 활용하여 진행한다. 데이터는 AI-Hub에서 수집했으며, 성별과 나이가 관계없이 구축된 3천여 개의 데이터 중 일부를 연구에 사용한다. 얼굴 이미지 데이터를 학습한 후, 위조 데이터를 해당 모델에 대입했을 때 출력되는 결과를 Grad-CAM++로 분석한다. 분석 결과는 점수로 환산하여 얼굴 영역별 중요도를 확인할 수 있으며, 중요 영역만을 사용하여 학습한 모델의 성능을 확인한다. 그 후 기존 모델과의 성능을 비교함으로써 검증을 거치고 얼굴 중요 영역만을 사용하는 효율적인 FAS 모델을 구축하는 방법을 제시한다.

2. 문헌조사

2.1 FAS 관련 연구의 흐름

초기의 FAS 연구는 주로 통계적 방법에 기반하여 얼굴 스푸핑을 탐지하는 데 초점을 맞췄다. Yan *et al.*(2012)은 Gaussian Mixture Model(GMM)을 활용하여 얼굴과 배경의 일관성을 계산하고, 행렬 기반 연산으로 눈 깜빡임과 같은 비정형 움직임을 분석하며, 웨이블릿 변환을 통해 이미지 밴딩을 측정하는 방식으로 print 스푸핑 공격을 100%의 정확도로 탐지한다. Kim *et al.*(2013)은 카메라의 초점 기능을 활용하여 가변 초점에 따른 픽셀값 변화를 분석함으로써 2D 스푸핑 공격을 탐지하는 방법을 제시한다.

딥러닝 기술이 발전하면서 FAS 연구는 점차 CNN 기반의 블랙박스 모델을 활용한 탐지 방법으로 전환되었다. 이는 통계적 접근법에 비해 더 높은 예측 성능을 가능하게 하며, 복잡한 패턴 분석에 적합하다. Chen *et al.*(2019)은 CNN을 사용하여 global/deep 특징을 추출하고, LBP(Local Binary Patterns)를 사용해 local/color texture 특징을 추출한 후, SVM(Support Vector Machine)을 활용하여 스푸핑 공격을 탐지하는 방식을 제안한다. Hadiprakoso *et al.*(2020)은 눈 깜빡임과 입술 움직임 같은 동작 특징을 분석하고, 이를 텍스처 특징과 결합한 CNN 기반 모듈을 제안한다. 이를 통해 2D 사진 기반의 스푸핑 공격에 대한 탐지 정확도를 크게 향상시킨다. Li *et al.*(2016)은 얼굴 피부의 맥박을 감지하는 방식으로 3D 마스크 스푸핑 공격을 탐지하며, 기존의 텍스처 분석 기법과 결합하여 robust cascade system 방식을 제안한다.

최근에는 XAI를 활용하여 모델의 설명 가능성을 높이려는 시도가 진행되고 있다. Liu *et al.*(2020)은 적대적 학습 프레임워크를 활용해 스푸핑 흔적을 분리하고, 다양한 스푸핑 유형에서 모델의 설명 가능성을 향상시키는 접근 방식을 제안한다. Sequeira *et al.*(2021)은 CNN 기반 PAD(Presentation Attack Detection) 모델의 해석 가능성을 탐구하며 전통적인 성능 평가와 설명 가능성 도구를 결합하여 모델의 안정성과 신뢰성을 보완한다. Prasad *et al.*(2023)은 RGB 이미지를 활용해 깊이 지도를 생성하고, 실제 얼굴과 스푸핑된 얼굴 간의 물리적 차이를 명확히 구분함으로써 다양한 환경에서도 높은 일반화 성능을 달성한다. Zhang *et al.*(2015)은 SPED(Spoofing Evidence Discovery)라는 새로운 접근 방식을 제안하여 활성화 맵과 스푸핑 증거를 시각화함으로써 모델의 판단 근거를 명확히 이해할 수 있도록 돕는다.

2.2 모델 압축과 효율을 높이기 위한 연구

고성능 모델의 효율성을 높이기 위해 가지치기(Pruning), 지식 증류(Knowledge Distillation), 양자화(Quantization)와 같은 모델 압축 기법이 널리 활용되고 있다(Yesuf and Assefa, 2023). 이들은 자원이 제한된 환경에서도 딥러닝 모델을 적용 가능하게 만드는 중요한 기술로 자리 잡고 있다. 특히, Yesuf and Assefa(2023)는 이러한 방법들이 단순히 모델 크기를 줄이는 것을 넘어 모델의 설명 가능성(Explainability)과 공정성(Fairness)에 미치는 영향을 탐구할 필요가 있음을 강조하였다. Vision Transformer(ViT)와 지식 증류를 결합한 경량화된 FAS 모델(HaTFAS, Zhang *et al.*, 2024)은 17배 적은 메모리와 9배 빠른 추론 속도로 자원 제한 환경에서도 높은 활용 가능성을 보여준다. Grad-CAM을 활용한 분석에서는 모델이 스푸핑 공격에서 나타나는 빛 반사와 같은 주요 특징에 주목했음을 확인할 수 있었다. 이를 기반으로 t-SNE(t-distributed Stochastic Neighbor Embedding) 시각화를 통해 교사 모델과 학생 모델 간의 특징 분포를 비교한 결과, 경량화된 학생 모델이 주요 특징을 효과적으로 학습하면서도 데이터 간 구별 능력을 유지했음을 확인했다.

XAI를 활용한 모델 압축 사례도 점차 확대되고 있다. Li and Song(2024)은 LRP(Layer-wise Relevance Propagation)를 활용하여 CNN의 가지치기 과정에서 각 필터의 중요도를 정량화하고, 이를 기반으로 불필요한 필터를 제거하며 성능과 해석 가능성을 동시에 유지하는 방법을 연구한다. 이 연구는 기존 가

지치기 기법의 한계를 극복하는 새로운 방안을 제시한다. Banerjee *et al.*(2024)은 LRP를 활용한 가지치기 기법을 Vision Transformer(ViT)에 적용하여 과도한 매개변수를 제거하면서도 ImageNet과 같은 대규모 데이터셋에서 높은 성능을 유지하도록 했다. 특히, ViT의 구조적 특성을 반영한 LRP 최적화를 통해 불필요한 요소를 정량적으로 제거하며 기존 접근법 대비 더 높은 압축률을 달성했다. 또한, XAI 기반 설명 가능성을 활용해 모델 신뢰성과 실질적 활용 가능성을 크게 강화했다. Becking *et al.*(2024)은 Entropy-Constrained and XAI-adjusted Quantization(ECQ^X) 기법을 제안하여 LRP를 기반으로 중요도를 평가하고, 이를 활용해 양자화를 수행한다. 이 기법은 성능 손실을 최소화하면서도 모델의 효율성과 설명 가능성을 극대화한다.

2.3 기존 연구의 한계 및 본 연구의 기여

기존 FAS 연구는 주로 정확도 향상에 초점을 맞췄으며, 블랙박스 모델을 활용한 방식은 높은 예측 성능에도 불구하고 해석 가능성과 신뢰성을 보완할 필요가 있었다. 특히, XAI를 활용한 연구가 부족하여 FAS 모델의 판단 근거를 명확히 제시하지 못한 한계가 존재했다. 또한, XAI를 활용한 모델의 효율성을 높이기 위한 접근 방식이 제한적이었다.

본 연구는 이러한 한계를 보완하기 위해 Grad-CAM++와 같은 XAI 도구를 활용하여 모델의 주요 탐지 영역을 정량적으로 분석하고, 중요 영역 이미지만 남기는 방식을 제안한다. 이를 통해 모델 학습, 추론, 데이터 저장 효율성을 모두 증대시키며, 유지보수 측면에서도 실질적인 개선을 제공한다. 또한, Grad-CAM++를 활용하여 생성된 활성화 맵은 모델의 결정 과정을 시각적으로 이해할 수 있게 하며, 개발자와 사용자가 모델의 신뢰도를 더욱 높일 수 있도록 돕는다. 이 연구는 FAS 분야에서 설명 가능성과 효율성의 균형을 맞춘 새로운 패러다임을 제시하며, FAS 모델의 신뢰성과 적용 가능성을 크게 확장하는 데 기여한다.

3. 연구 방법

효율적인 FAS 모델 구축을 위한 과정은 <Figure 1>과 같다. 먼저 전통적인 방법으로 FAS 모델을 구축하고, Face Landmarks와 Grad-CAM++을 활용하여 영역별 중요도를 추출한다. 그 후 중요 영역만을 사용하여 FAS 모델을 구축하는 순서로 연구를 진행한다.

3.1 데이터 수집 및 전처리

실제 Face Spoofing은 여러 조건(조도, 각도, 연령대, 성별 등)에서 시도되고 있으므로 다양한 환경에서 촬영하여 수집된 데이터가 필요하다. 이에 본 연구에서는 AI-Hub의 '안면 위변조 감지를 위한 데이터'를 사용하고자 한다. AI-Hub는 한국지능정보사회진흥원에서 운영하는 AI 통합 플랫폼으로, 국내외 기관/기업에서 보유한 인공지능 학습용 데이터를 공개하고 있다. 본 연구에서는 Dlib 라이브러리를 사용하여 전처리를 수행

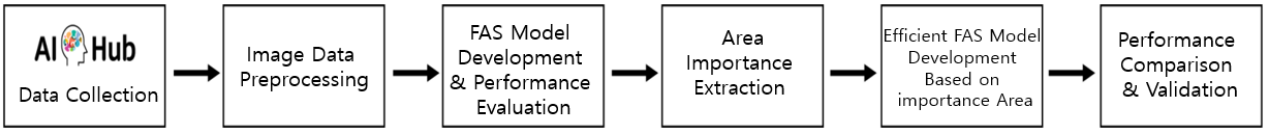


Figure 1. The Overall Process of the Proposed Approach

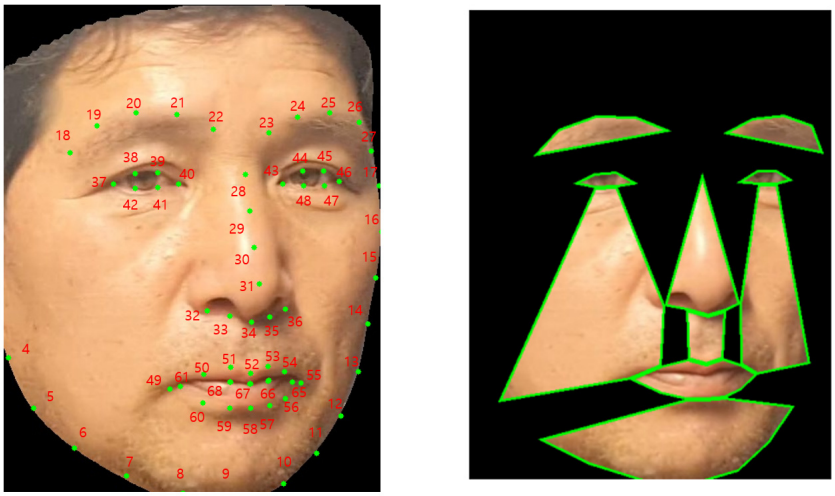


Figure 2. Facial Feature Points and Face Areas in the Dlib Library

한다. Dlib 라이브러리는 얼굴 이미지에서 68개 특징점을 추출하는 Face Landmarks 기술을 제공한다. 또한, <Figure 2>의 오른쪽 그림처럼 특징점을 조합해서 눈, 볼, 코 등과 같이 얼굴의 특정 영역을 추출한다.

3.2 전통적인 FAS 모델 구축

본 연구에서는 ㈜딥핑소스에서 제공하는 AI 모델을 활용한다. ㈜딥핑소스 기업은 2018년에 설립되어 AI 기술을 활용하여 공간과 행동 분석을 통해 공간 운영의 최적화, 사용자 경험 개선, 안전 및 보안 강화를 목표로 하는 솔루션을 제공한다. 또한 독자적인 데이터 익명화 기술을 통해 AI가 대규모 영상을 분석할 때 개인정보를 보호한다.

본 연구에 Real(실제 얼굴)/Spoof(위조된 얼굴)를 분류하기 위해 사용되는 모델은 Feng *et al.*(2020)에서 제안한 LGSC (Learning Generalized Spoof Cues for Face Anti-spoofing, LGSC) 구조로, <Figure 3>처럼 Spoof Cue Generator와 Aux Classifier로 나누어져 있다. U-Net 기반의 Spoof Cue Generator에서 생성된 Spoof Cue를 입력 이미지와 합친 후, Aux Classifier의 잔차 학습을 통해 분류된다. 이 모델의 손실 함수는 Triplet Loss(논문 인용 추가 필요)와 Regression Loss, Classification Loss 3가지를 각 1:5:5의 비율로 가중 합하여 계산된다. <Equation 1>에서 정의된 Triplet Loss를 통해 Real/Spoof 집단의 분리성을 높이고, Real 집단의 밀집성을 높인다. 또한 Regression Loss를 통해 Spoof Cue의 성능을 일반화하고, Classification Loss를 통해 Spoof Cue와 입력 이미지를 Real/Spoof 집단을 분류할 수 있게 한다.

Equation 1. Compute Triplet Loss

$$L_t = \frac{1}{T}\sum_{i=1}^{T}(\max(d(a_i,p_i) - d(a_i,n_i) + m, 0),$$

$$d(i,j) = \| \frac{v_i}{\|v_i\|_2} - \frac{v_j}{\|v_j\|_2} \|_2$$

a_i : anchor (real)
p_i : positive (real)
n_i : negative (spoof)
T : triplet의 개수

3.3 중요 영역 기반 효율적인 FAS 모델 구축

효율적인 FAS 모델을 구축하기 위해 XAI 기법과 Dlib 라이브러리를 사용하여 중요한 얼굴 영역을 추출한다. 이를 반영

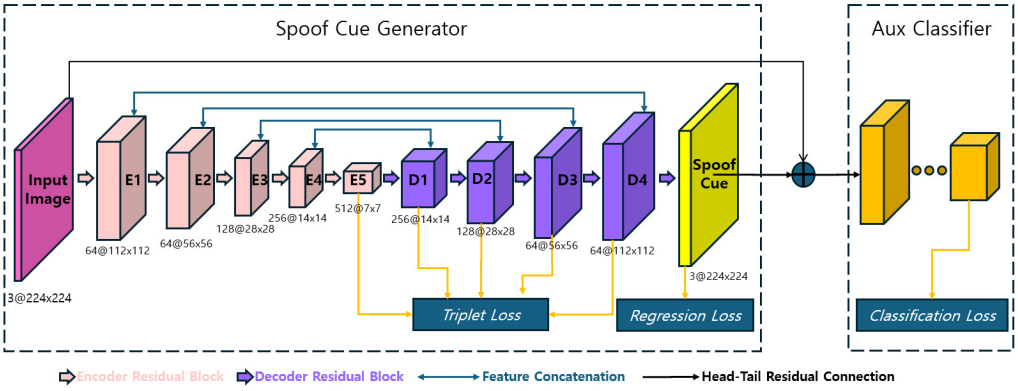


Figure 3. Structure of a Traditional FAS Model

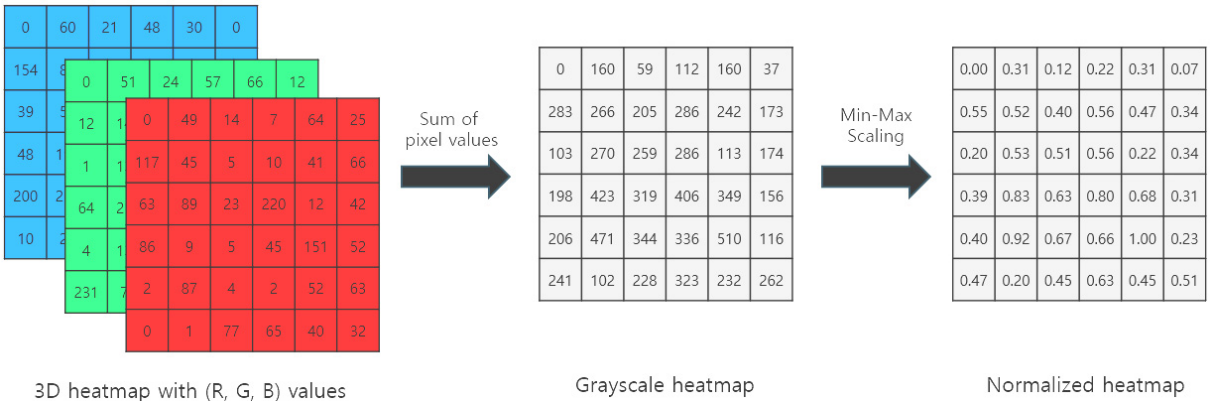


Figure 4. Conversion Process of a Heatmap Obtained with Grad-CAM++

한 데이터 셋을 재구성하여 학습한 모델과 기존 모델의 성능을 비교한다.

(1) Grad-CAM++과 Face Landmarks를 활용한 중요 영역 계산

모델을 분석하기 위한 XAI 기법으로 Grad-CAM++을 사용한다(Chatto and Sarkear, 2018) Yang *et al.*(2021)**의 연구에 따르면, Grad-CAM++는 Grad-CAM보다 객체를 더 포괄적으로 커버하며, 특히 모델이 판단에 활용한 주요 영역을 더욱 정확하게 시각화할 수 있다. 본 연구에서 Grad-CAM++의 target layer는 Auxiliary Classifier의 마지막 Convolution layer를 선택한다. 최종적으로 Grad-CAM++의 결과는 모델이 얼굴의 어느 영역을 통해 판단했는지를 알려준다.

얼굴 중요 영역 후보는 Dlib의 Face Landmarks를 사용하여 도출하고, 중요 영역을 정량적으로 계산하기 위해 Area Importance를 정의한다. Area Importance는 0과 1 사이의 값으로 구성되고 1에 가까울수록 중요한 영역으로 평가되며 <Figure 4>의 과정을 통해 도출된다. Area Importance를 계산하기 위해서는 Grad-CAM++을 통해 얻은 (R, G, B) 3차원 Heatmap을 1차원이면서 0~1사이의 값으로 정규화 된 Grayscale Heatmap으로 변환한다.

이미지 얼굴 내에서 특정한 얼굴 영역(눈, 볼, 코, 입 등)의 내부 영역은 1, 특정 영역 외부 영역은 0으로 구성된 mask를 생성한다. 생성한 mask와 Grayscale Heatmap을 요소별로 곱한 결과 값을 masked_Heatmap으로 정의한다. 각 이미지마다 masked_Heatmap 합계를 mask 합계로 나누어 Area Importance를 <Equation 2>와 같이 계산한 후, 모든 이미지의 Area Importance를 평균 내어 중요 영역을 도출한다.

Equation 2. Area_Importance formula for determining important areas

$$Area\_Importance(A) = \sum_{i}^{m}\left(\sum(masked\_Heatmap_i)/\sum(mask_i)\right)$$

A : 얼굴 영역

m : 데이터 개수

$masked\_Heatmap_i$: i 데이터의 마스크 처리된 영역의 중요도 합

$mask_i$: i 데이터의 마스크 영역의 크기

(2) 중요 영역 기반 데이터 셋 재구성 및 검증

효율적인 FAS 모델 구축을 위해 훈련 이미지에서 중요한 영역만 사전에 추출하여 학습하는 방식을 도입한다. 이를 위해서 Area Importance가 높은 일부 얼굴 영역만 추출하여 검증 데이터 셋을 재구성한다. 재구성한 데이터 셋으로 학습한 FAS 모델과 전통적인 방식으로 학습한 FAS 모델의 성능이 차이가 없고 훈련/예측 시간이 줄어든다면, 본 연구에서 도출한 Area Importance가 중요 영역을 탐색하는 데에 의의가 있음을 의미한다.

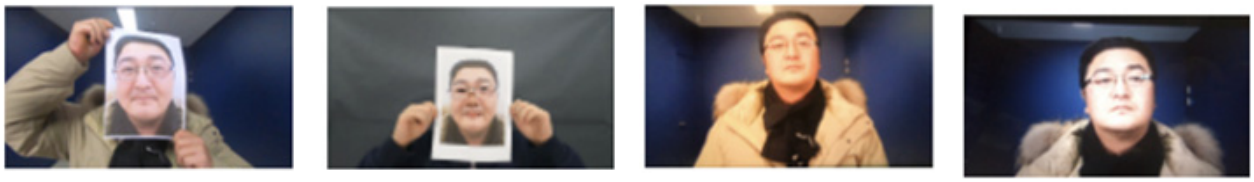
4. 사례 연구

4.1 데이터 수집 및 전처리

AI-Hub 데이터는 스마트폰, 태블릿 등 총 20여 종의 스마트 디바이스를 이용한 RGB 데이터에 셋으로 구성되어 있고, 촬영 대상자는 3,000명 이상으로 총 54,150개의 이미지로 구성되어 있다. 실제 사람의 얼굴로는 <Figure 5>와 같이 구성되어 있고, 안면 위변조로 <Figure 6>과 같이 Print, Media Replay, 3D mask 등을 시도하며 고조도, 일반조도, 저조도의 3가지의 조도 환경과 30가지의 각도로 촬영한 사진에 대하여 두 Class(real:

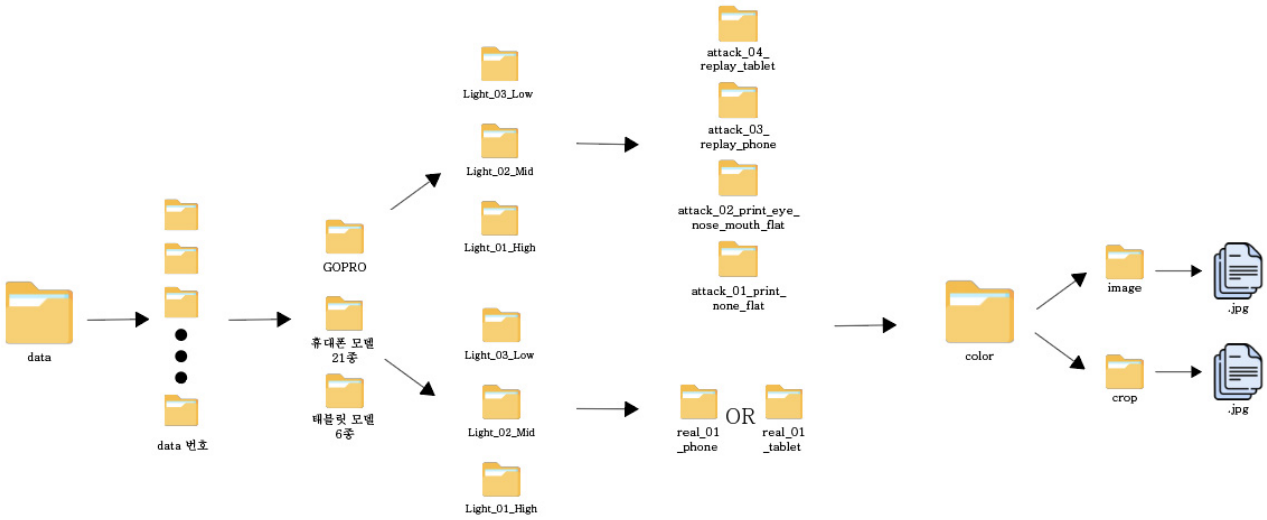
<Taken with Smartphone> <Taken with Tablet>

Figure 5. 2 Types of Real



<Printed Photo> <Printed Photo Eyes, Nose, and Mouth Flattened> <Replayed video on Smartphone> <Replayed Video on Tablet>

Figure 6. 4 Types of Attack



* Real samples were taken with both a smartphone and a tablet, whereas attack samples (photos) were taken exclusively with a GoPro.

Figure 7. File Structure of Collected Data

실제 사람의 얼굴, attack: Face Spoofing)로 나뉜다. 추가로 각 이미지마다 얼굴 영역만 추출된 이미지도 함께 구성되어 있다. 최종적으로 본 연구에서는 <Figure 7>과 같은 구조로 구성된 150명의 데이터 셋을 사용한다.

추후 효율적인 FAS 모델 구축을 위한 얼굴의 중요 영역을 추출하는 전처리를 진행한다. 중요 얼굴 영역 후보는 Landmarks를 조합하여 <Table 1>과 같이 구성한다.

Table 1. Important Face Area Candidates

Face Area	Landmark Combination Numbers
Left Eyebrow	18, 19, 20, 21, 22
Right Eyebrow	23, 24, 25, 26, 27
Nose	28, 32, 33, 34, 35, 36
Lips	49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59,60
Left Eye	37, 38, 39, 40, 41, 42
Right Eye	43, 44, 45, 46, 47, 48
Left Cheek	42, 41, 32, 50, 49, 5
Right Cheek	47, 48, 36, 54, 55, 13
Chin	57, 58, 59, 6, 7, 8, 9, 10, 11, 12
Philtrum	33, 34, 35, 53, 52, 51

4.2 전통적인 FAS 모델 구축 및 성능 평가

본 연구에서는 FAS 모델을 다양한 측면에서 균형 있게 학습할 수 있도록 한다. 최종적으로, 모델은 Precision, Recall, F1-score 등 다양한 평가 지표에서 <Table 2>와 같이 우수한 성능을 보인다.

데이터 셋 구성 시, 단순 데이터의 개수가 아닌 전체 피험자를 6:2:2 비율로 분할하여, Train, Validation, Test Set을 구성하여 데이터 셋 간 피험자 중복을 배제하였다. 이런 과정을 통해 모델의 같은 피험자가 다른 데이터 셋에 섞이는 것을 방지하였다. 각 데이터 셋은 다음과 같은 역할을 수행한다. Train Set은 모델의 패턴 학습 및 가중치 최적화에 사용되며, Validation Set은 학습 과정에서 모델의 성능 평가와 하이퍼파라미터 튜닝을 수행한다. Test Set은 학습에 사용되지 않은 독립적인 데이터로, 최종 모델의 일반화 성능을 객관적으로 평가한다.

Table 2. Performance of the Traditional FAS Model

Precision	Recall	F1-score
0.9983	0.9952	0.9929

4.3 효율적인 FAS 모델 구축

(1) Grad-CAM++과 Face Landmarks를 활용한 중요 영역 계산

모델 분석을 위한 Grad-CAM++의 target layer는 모델의 마지막 Convolution layer를 선택했다. Grad-CAM++의 결과는 <Figure 8>과 같이 Class가 1(real)인 경우는 얼굴의 모든 부분이 중요하게 나왔으며, Class가 0(attack)인 경우는 얼굴의 특정 영역이 localizing 되는 것을 확인했다.

얼굴의 중요 영역 후보는 총 10개로 왼쪽 눈썹, 오른쪽 눈썹, 코 전체, 입술 전체, 왼쪽 눈, 오른쪽 눈, 왼쪽 볼, 오른쪽 볼, 턱, 인중으로 구성했다.

Grad-CAM++과 Dlib의 Face Landmarks를 활용한 영역별 Area Importance의 결과는 <Table 3>에 나타난 대로 도출되었다. <Figure 9>에서 확인되는 것처럼 오른쪽 볼, 오른쪽 눈, 왼쪽 눈, 왼쪽 눈썹 순서대로 높게 계산되었으며 인중 영역은 거의 0에 가까운 결과였다.

얼굴 영역별 Area Importance는 좌우 비대칭한 결과가 나타난다. 특히, 왼쪽 볼과 오른쪽 볼의 Area Importance는 각

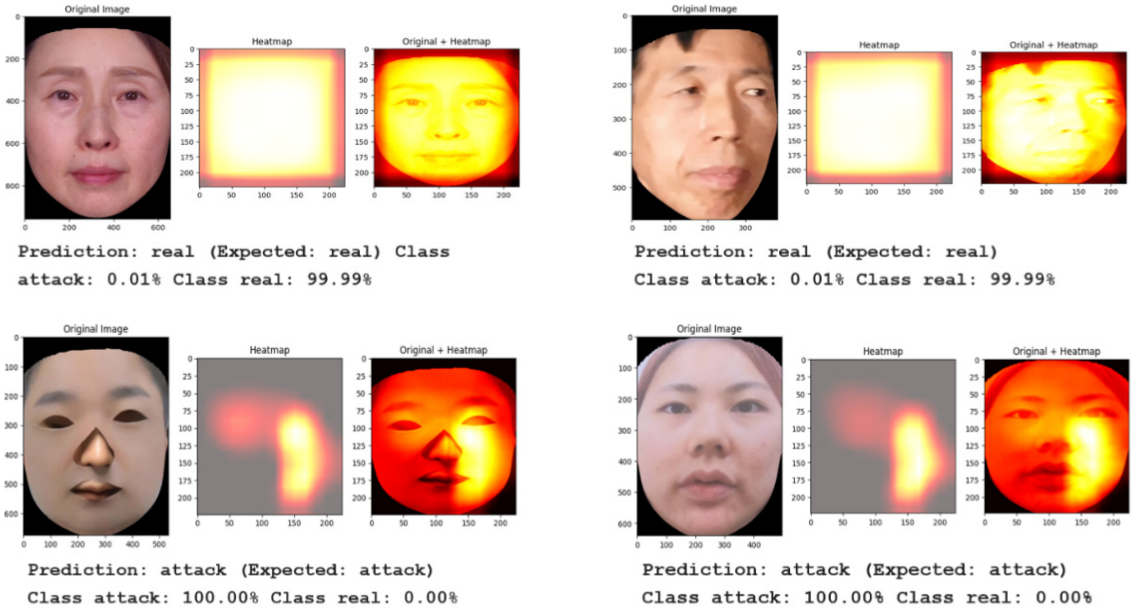


Figure 8. Visual Explanations of Model Predictions Using Grad-CAM++

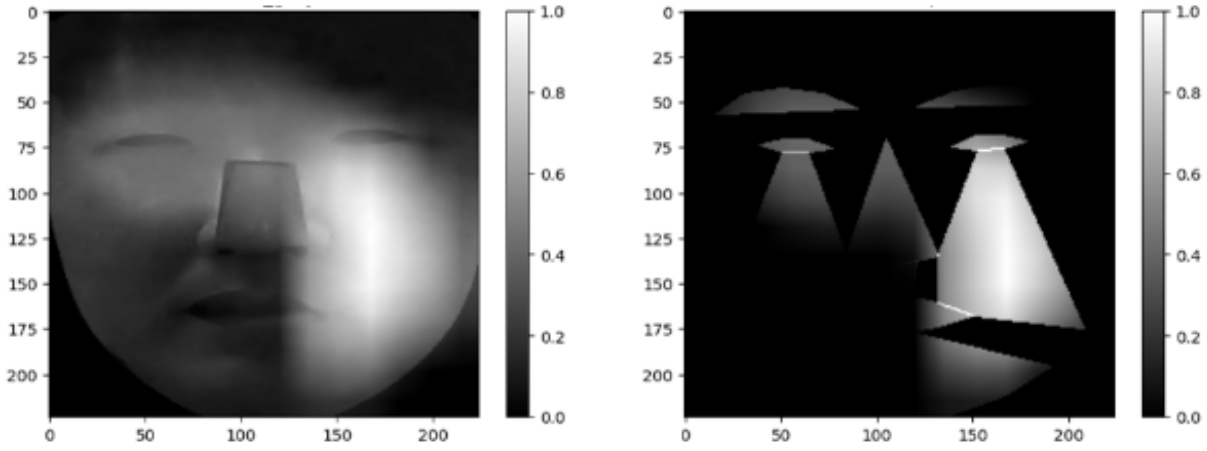


Figure 9. Visual Explanation of Face Area Candidates Using Grad-CAM++

Table 3. Area Importance Values for Selected Face Area Candidates

Face Area	Area Importance
Left Eyebrow	0.2930
Right Eyebrow	0.1086
Nose	0.1811
Lips	0.1005
Left Eye	0.3875
Right Eye	0.5748
Left Cheek	0.0784
Right Cheek	0.7831
Chin	0.0868
Philtrum	0.0082

0.0784, 0.7831로 가장 비대칭한 결과가 나온다. 비대칭한 결과의 이유는 데이터 수집 단계에서의 편향으로 예상된다. <Figure 10>과 같이 데이터의 수집 단계에서 조명의 위치가 왼쪽으로 치우쳐져 있고, 두 번째 그림에서는 왼쪽 볼의 일부분이 조명에 반사되는 것을 확인할 수 있다. 따라서, 모델이 전반적으로 왼쪽 볼보다, 일관된 데이터의 품질의 오른쪽 볼을 더 활용하며 일반화되었다고 예상된다. 또한, 이러한 결과는 Chan *et al.*(2017)의 연구에서 조명 조건에 민감하다는 결론과 유사하다.

(2) 중요 영역 기반 데이터 셋 재구성 및 검증

검증 데이터 셋은 기존 데이터 셋의 Train, Validation, Test와 같은 분포에서 상위 4개의 영역(오른쪽 볼, 오른쪽 눈, 왼쪽 눈, 왼쪽 눈썹)을 <Figure 11>과 같이 추출하여 재구성하였다. 추

Figure 10. Examples of Facial Images Showing the Impact of Lighting Asymmetry

Figure 11. Example Image Extracted with Important Face Areas

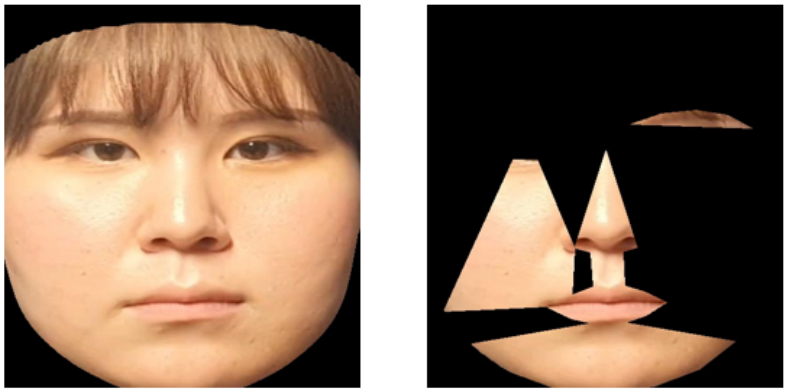


Figure 12. Example Image Extracted with Unimportant Face Areas

Table 4. Performance of the Efficient FAS Model

	Precision	Recall	F1-score
Model trained with important face areas	0.9759	0.9966	0.9855
Model trained with unimportant face areas	0.9404	0.9991	0.9688

가로 검증 데이터 셋과 비교를 위한 하위 6개의 영역(오른쪽 눈썹, 코, 입술, 왼쪽 볼, 턱, 인중)을 <Figure 12>와 같이 추출하여 데이터 셋을 재구성했다.

기존 데이터 셋과 동일한 분포를 사용하여 Train, Validation, Test 각 39,377개, 12,972개, 13,651개로 구성하였다. 검증 데이터 셋을 사용하여 학습한 검증 모델의 성능은 <Table 4>와 같이 기존 모델의 성능과 거의 유사했다. 검증 모델과 기존 모델의 hyperparameter는 동일하게 적용했다. 추가로, 상위 4개의 영역으로 학습한 모델은 하위 6개의 영역으로 학습한 모델에 비해 얼굴 픽셀이 더 적음에도 불구하고 Precision과 Recall의

균형이 더 우수하고, F1-score 또한 더 높게 나왔다. 특히 Face Anti-Spoofing 모델에서 중요한 Precision에서 더 큰 차이를 보였다. 훈련에 걸리는 시간을 확인하기 위해서 사용한 컴퓨터 환경은 윈도우 11 프로 운영체제에서 GPU는 NVIDA RTX A6000 1개, CPU는 Intel사의 제온 Gold 6248R이다. 비교 실험은 Dlib 라이브러리로 검출에 성공한 데이터만 사용하여 두 방식의 학습 데이터 수를 동일하게 하였다. 상위 4개 영역만 추출된 이미지는 전통적인 방식의 이미지와 같은 픽셀 수를 갖는 이미지지만, JPG 기준 용량은 평균적으로 25% 감소했으며, 훈련 시간은 <Figure 13>과 같이 약 9% 감소했다.

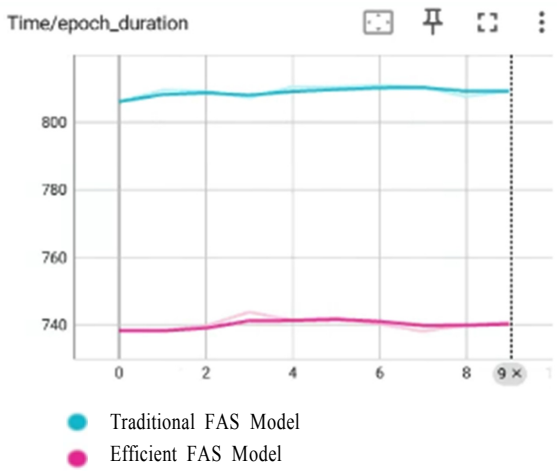


Figure 13. Training Time Per Epoch

5. 결론 및 향후 연구

5.1 연구 결과 요약 및 결과에 따른 시사점

본 연구에서는 효율적인 FAS 모델 구축을 위한 XAI 기반 접근법을 제시한다. 첫째, 전통적인 방법으로 FAS 모델을 설계 및 구축한다. 둘째, Grad-CAM++과 Face Landmarks를 활용하여 얼굴 중요 영역의 값을 정량적으로 도출한다. 마지막으로, 데이터 셋을 얼굴 중요 영역으로 재구성하여 학습한다. 재구성한 데이터 셋으로 학습한 모델은 전통적인 모델의 성능을 유지하고, 훈련 이미지 영역 내에서 약 80%의 불필요한 영역을 검은색으로 변환하여 계산 영역이 획기적으로 감소한다. 이를 통해 모델이 학습 과정에서 중요한 영역에 집중하도록 유도하며, 컴퓨팅 자원의 효율성을 높일 수 있다.

다만, 초기에 중요 얼굴 영역을 탐색하는 과정에서 더 많은 시간과 자원이 필요할 수 있다. 하지만, 최초로 만들어진 모델은 완벽한 일반화 성능을 갖추지 못하기 때문에, 스푸핑 방법과 환경의 변화에 따라 추가적인 학습이 필수적이다. 따라서 모델 유지보수를 위한 데이터 저장 및 추가 학습의 관점에서 XAI 기법을 이용한 중요 영역 탐색은 자원 관리의 효율성을 높이는 효과적인 방법이라 할 수 있다. 추가로 효율적인 FAS 모델을 설계하기 위해 Area Importance를 분석하면, 조명과 같은 데이터 수집 과정에서의 편향을 예측할 수 있는 가능성을 제공한다.

5.2 연구의 한계점 및 향후 과제

본 연구에는 몇 가지 한계점이 있으며, 이는 향후 연구의 방향을 제시한다. 첫째, Dlib 라이브러리의 Face Landmarks는 정면 얼굴에 특화되어 있어 전체 이미지의 약 20%를 연구에서 사용하지 못했다. 향후 연구에서는 다양한 각도에서 얼굴 특징점을 분류할 수 있는 모델을 사용하여 해당 접근법을 개선할 수 있다. 둘째, 향후 연구에서는 다양한 얼굴 영역 조합을 실험하여 해당 방법론의 한계를 파악할 수 있다. 예를 들어, Area Importance가 가장 높게 나온 오른쪽 볼 영역으로만 학습하고, 가장 낮게 나온 인중 영역으로만 학습했을 때의 차이를 비교하는 등의 여러 조합을 시도할 수 있다. 셋째, 본 연구에서는 Grad-CAM++을 통해 Area Importance를 도출했으나, 향후 연구에서는 AblationCAM, ScoreCAM 등의 XAI 기법을 사용하여 Area Importance의 안정성을 검증할 수 있다. 넷째, 중요 영역을 활용한 효율적인 FAS 모델 구축법은 스푸핑 공격 탐지에서 성능과 신뢰성을 보장하지만, 얼굴 인식 성능을 보장하지 못한다. 본 연구에서 중요 영역은 FAS 모델의 Grad-CAM++ 결과를 바탕으로 도출했기 때문에, 스푸핑 공격 탐지와 얼굴 인식의 연관성에 대해 추가적인 조사와 실험이 필요하다. 다섯째, 본 연구에서 사용한 데이터에는 조명 편향이 있는 것으로 확인되었다. 조명은 얼굴 인식에서 중요한 환경 요소로, 다양한 조명 환경에 따라 FAS 모델의 성능이 저하될 수 있으므로 추가적인 검증이 필요하다. 다양한 조명 환경의 데이터가 수집되면, 본 연구에서 제시한 방법론이 유효할 것으로 예상된다.

참고문헌

Becking, D., Dreyer, M., Samek, W., Müller, K., and Lapuschkin, S. (2021), ECQx: Explainability-Driven Quantization for Low-Bit and Sparse DNNs, *Proc. International Conference on Machine Learning*, Springer, Cham, pp. 14-28.

Chan, P. P. K., Liu, W., Chen, D., Yeung, D. S., Zhang, F., Wang, X., and Hsu, C.-C. (2017), Face Liveness Detection Using a Flash Against 2D Spoofing Attack, *IEEE Transactions on Information Forensics and Security*, **13**(2), 521-534.

Chattopadhyay, A. and Sarkar, A. (2018), Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks, arXiv:1710.11063v3.

Chen, F.-M., Wen, C., Xie, K., Sheng, G.-Q., and Tang, X.-G. (2019), Face Liveness Detection: Fusing Colour Texture Feature and Deep Feature, IET Biometrics.

Feng, H., Hong, Z., Yue, H., Chen, Y., Wang, K., Han, J., Liu, J., and Ding, E. (2020), Learning Generalized Spoof Cues for Face Anti-spoofing,

arXiv:2005.03922.
Hadiprakoso, R. B., Setiawan, H., and Girinoto (2020), Face Anti-Spoofing Using CNN Classifier & Face Liveness Detection, *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, Indonesia.
Hatefi, S. M. V., Dreyer, M., Achtibat, R., Wiegand, T., Samek, W., and Lapuschkin, S. (2024), Pruning by Explaining Revisited: Optimizing Attribution Methods to Prune CNNs and Transformers, arXiv preprint arXiv:2408.12568.
He, K., Zhang, X., Ren, S., and Sun, J. (2015), Deep Residual Learning for Image Recognition. arXiv:1512.03385v1.
Kim, S., Yu, S., Kim, K., Ban, Y., and Lee, S. (2013), Face Liveness Detection Using Variable Focusing, *Proc. International Conference on Biometrics (ICB)*, IEEE, Madrid, Spain, 04-07.
Li, X., Komulainen, J., Zhao, G., Yuen, P.-C., and Pietikäinen, M. (2016), Generalized Face Anti-Spoofing by Detecting Pulse from Face Videos, *Proc. 23rd International Conference on Pattern Recognition (ICPR)*, 04-08 December 2016, Cancun, Mexico, 4321-4326.
Liu, Y., Stehouwer, J., and Liu, X. (2020), On Disentangling Spoof Trace for Generic Face Anti-Spoofing, *To appear in Proc. European Conference on Computer Vision (ECCV) 2020*.
Ronneberger, O., Fischer, P., and Brox, T. (2015), U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597v1.
Sequeira, A. F., Gonçalves, T., Silva, W., Pinto, J. R., and Cardoso, J. S. (2021), An Exploratory Study of Interpretability for Face Presentation Attack Detection, *IET Biometrics*, **10**(2), 181-191.
Yan, J., Zhang, Z., Lei, Z., Yi, D., and Li, S. Z. (2012), Face Liveness Detection by Exploring Multiple Scenic Clues, *Proc. 12th International Conference on Control Automation Robotics & Vision (ICARCV)*, IEEE, Guangzhou, China, 05-07.
Yang, Y., Zhang, L., Du, M., Bo, J., Liu, H., Ren, L., Li, X., and Deen, M. J. (2021), A comparative analysis of eleven neural networks architectures for small datasets of lung images of COVID-19 patients toward improved clinical decisions, *Computers in Biology and Medicine*, **139**, 104887.
Yesuf, M. M. and Assefa, B. G. (2023), Model Compression Techniques in Deep Neural Networks, Pan-African Conference on Artificial Intelligence, *Communications in Computer and Information Science*, **169**, Springer, Cham, 169-190.
Zhang, H., Zhu, X., Gao, L., Zhao, G., and Lei, Z. (2023), A Self-explainable Face Anti-spoofing Solution Based on Depth Estimation, *Proc. International Conference on Artificial Intelligence and Robotics*, Springer, Singapore, 120-134.
Zhang, H., Zhu, X., Gao, L., Zhao, G., and Lei, Z. (2024), Concept Discovery in Deep Neural Networks for Explainable Face Anti-Spoofing, arXiv preprint arXiv:2412.17541.
Zhang, J., Zhang, Y., Shao, F., Ma, X., Feng, S., Wu, Y., and Zhou, D. (2024), Efficient Face Anti-Spoofing via Head-Aware Transformer Based Knowledge Distillation with 5 MB Model Parameters, *Applied Soft Computing*, **136**, Article 110011.

저자소개

한태혁 : 논문 투고 시점에 한양대학교 산업융합학부 정보공학전공 학사과정 재학 중이었고, 2025년 2월에 졸업이후 현재 성균관대학교 데이터사이언스 융합학과 석사과정에 재학 중이다. 관심 연구분야는 컴퓨터 비전 분야 및 설명가능 인공지능 응용이다.

정준각 : 포항공과대학교 산업경영공학과에서 2013년 학사, 2019년 석박사 통합 학위를 취득하였다. 일리노이 대학교 어바나-샴페인과 울산과학기술원 산업공학과에서 박사후연구원으로 근무했으며, 현재 한양대학교 산업융합학부에서 조교수로 재직 중이다. 관심 연구분야는 텍스트 마이닝, 품질 데이터 분석, 설명가능 인공지능 응용이다.