

Grad-CAM++ 기반 보조 히트맵 입력을 활용한 경량 얼굴 인식 모델의 강건성 향상

김영언*¹, 변경태*², 김건우**

Enhancing Robustness of Lightweight Face Recognition Models via Auxiliary Heatmaps based on Grad-CAM++

Young-Eon Kim*¹, Gyeong-Tae Byeon*², 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

요약

본 연구는 경량화 얼굴 인식 모델의 성능 보완을 위해, Grad-CAM++ 기반 Heatmap을 보조 입력 채널로 통합하는 방법을 제안한다. 제안된 구조는 모델의 구조적 변경 없이 Grad-CAM++으로 자동 생성된 Heatmap을 통해 시각적 주의 영역을 명시적으로 제공함으로써 학습 초기 수렴 속도를 향상하고, 특히 가림 및 밝기 변화 조건에서 F1-score가 유의하게 향상되었다. 나아가 Zero/Random/Shuffle ablation을 통해 Heatmap 채널이 성능 향상의 인과적 요인임을 검증하였다. 결론적으로 Heatmap 입력은 모델의 해석 가능성과 시각적 정보 집중에 긍정적인 역할을 하였으며, 경량 모델의 실사용 가능성을 높이는 실용적 구조로서의 가능성을 입증하였다.

Abstract

This study proposes a method to enhance lightweight face recognition models by integrating Grad-CAM++ - based Heatmaps as auxiliary input channels. Without modifying the backbone architecture, the proposed structure explicitly guides visual attention through automatically generated Grad-CAM++ Heatmaps. This guidance accelerates early convergence and improves macro F1 under occlusion and brightness variations. A Zero/Random/Shuffle ablation experiment verified that the Heatmap channel is a causal factor in the performance improvement, confirming the robustness of the proposed approach. Overall, the Heatmap input enhances model interpretability and visual focus, demonstrating its practicality for deployment in lightweight face recognition systems.

Keywords

lightweight face recognition, Grad-CAM++, auxiliary input, heatmap integration, attention guidance, mobile deployment

* 경상국립대학교 컴퓨터공학과

- ORCID1: <https://orcid.org/0009-0002-6877-2392>

- ORCID2: <https://orcid.org/0009-0007-2467-9301>

** 경상국립대학교 컴퓨터공학과 교수(교신저자)

- ORCID: <https://orcid.org/0000-0001-5643-4797>

• Received: Sep. 24, 2025, Revised: Nov. 03, 2025, Accepted: Nov. 06, 2025

• Corresponding Author: Gun-Woo Kim

Dept. of Computer Science 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

경량화된 얼굴 인식 모델(MobileNet, ShuffleNet, EfficientNet 등)은 모바일 환경과 같은 자원 제약 조건에서도 빠른 연산 속도와 높은 연산 효율성을 제공하는 장점이 있다[1]-[3]. 특히 실시간 응답이 요구되는 현장 모니터링 시스템이나 모바일 기기에서 이러한 경량 모델의 활용 사례가 크게 늘고 있으며, 실제 환경에 적용 가능한 얼굴 인식 솔루션 개발에 있어 핵심 기술로 자리잡고 있다.

그러나 실제로 사용할 때는 조명 변화, 얼굴 부위의 부분적 가려짐 및 이미지 품질 저하 등 여러 요인이 성능에 부정적인 영향을 미치곤 한다. 이런 상황에서 모델의 안정성 문제에 대한 의문이 제기되고 있다. 특히 계산 속도를 향상하기 위해 구조를 단순하게 만든 경량 모델들의 경우에는 세밀한(Fine-grained) 시각 정보를 충분히 학습하지 못할 수 있다. 그렇기에 특정 환경 조건에서는 성능이 크게 떨어질 위험이 존재하게 된다.

이러한 문제에 대응하기 위해 활용되는 대표적인 해석 기법이 Grad-CAM(Gradient-weighted Class Activation Mapping)이다[4]-[7]. 이는 이미지 내에서 모델이 어느 영역을 중요하게 보는지 시각적으로 강조해 주는 특징이 있다. Grad-CAM++는 이를 더욱 정밀하게 발전시킨 것으로 세밀한 인식(Fine-grained recognition) 작업에서도 효과적으로 작동한다[5]. 보통 이러한 시각화 기법은 주로 사후 해석(Post-hoc visualization) 목적으로 사용되었으나, 최근 연구들은 이를 학습 과정에 직접 통합하여 모델의 주의 집중 영역을 유도하는 보조 정보로 활용하는 방법도 연구되고 있다[6]. 한편, 기존 주의(Attention) 기반 접근들은 주로 모델 내부에 별도의 모듈을 삽입하거나 추가 파라미터를 학습하는 방식으로 진행 되어왔다. 이러한 연구들은 세밀한 인식에서는 유리하지만, 연산 비용이 증가하여 경량 모델에서의 효용성이 떨어지는 한계를 지닌다.

이에 본 논문에서는 구조적 복잡성을 최소화하고 경량 모델에 즉시 적용 가능한 실용적인 대안을 제시하고자 한다. 구체적으로 Grad-CAM++을 통해 사전에 생성된 Heatmap을 기존 RGB 이미지와 함께

사용하는 4채널 입력 구조를 설계하였다. 제안된 방식은 기존 모델 아키텍처를 크게 수정하지 않으면서도 시각적으로 중요한 영역에 대한 정보를 모델에게 미리 제공하여, 보다 효율적이고 안정적인 학습이 가능하게 하였다. 또한, 모바일 환경에 즉시 적용할 수 있는 경량 모델(MobileNetV3, ShuffleNetV2 등)을 기반으로 실험을 진행함으로써 제안한 방식이 실제로 적용 가능한지도 함께 검증하였다.

II. 실험 설계 및 구성

2.1 실험 개요 및 목적

실제 모바일 환경에서 얼굴 인식 시스템을 사용할 때는 여러 문제가 발생하게 된다. 빛의 밝기가 달라짐에 따라, 얼굴의 일부가 가려짐에 따라, 사진 화질이 좋지 않음에 따라 발생한 시각적 변화가 모델이 인식해야 할 핵심 정보들을 잃게 만든다는 것이다. 특히 파라미터 수가 제한된 경량화 모델들은 표현력이 제한되어 있어서, 이런 어려운 상황에서 학습 안정성과 정확도가 더욱 저하될 수 있다.

이와 같은 현실적 제약 조건을 고려하여, 본 연구는 모델이 주목해야 할 시각적 영역을 사전에 명시적으로 제공함으로써 학습 안정성과 시각 정보 집중 효과를 향상하고자 하였다. 구체적으로는 Grad-CAM++ 기법을 활용하여 생성한 Heatmap을 원본 RGB 이미지와 채널 차원에서 결합한 4채널 입력 구조를 설계하였다. 이러한 접근법이 학습 초기 수렴 속도를 개선하는 데, 다양한 조건 변화 환경에서도 안정적인 성능을 유지하는 데 어떤 도움을 주는지 실험을 통해 검증하고자 한다.

2.2 입력 구조 설계

본 연구에서는 모바일 환경에서 널리 활용되는 대표적인 경량화 모델 3개를 선택하였다. 선택한 MobileNetV3-Large, ShuffleNetV2-1.0x, EfficientNet-B0 모델들은 구조가 간결하고 빠른 추론을 가능하게 하여 자원이 제한된 환경에 적합하다. 실험은 두 가

지 방식으로 나누어 진행했다. 첫 번째는 기본(Base) 방식으로, 일반적인 RGB 3채널 이미지만을 입력으로 사용한다. 두 번째는 본 연구에서 제안하는 Heatmap 방식으로, Grad-CAM++로 만든 Heatmap을 네 번째 추가 채널로 추가해 총 4채널 입력을 사용하였다. Grad-CAM++ 기반 Heatmap을 만드는 방법은 다음과 같다. 먼저 모델의 마지막 주요 합성곱 층에서 특징 맵을 추출한다. 이후 클래스별 가중치를 계산해 중요 영역을 강조한다. 이렇게 생성된 Heatmap은 224×224 크기로 조정하고, 0~1 범위로 정규화한 후 npy 포맷으로 저장한다. 학습 데이터를 구성할 때는 원본 RGB 이미지와 저장된 Heatmap을 채널 차원에서 결합하여 4채널 입력을 만든다. 기존 모델 아키텍처들은 입력 채널 수가 3개(RGB)로 고정되어 있기 때문에, 본 연구에서는 입력층의 첫 번째 Convolution layer를 수정하여 4채널 입력을 받을 수 있도록 조정하였다. 이때 학습 안정성을 위해 기존 가중치 값을 활용해 초기화하였다.

주목할 점은 Heatmap이 학습 중 별도의 추가 연산을 요구하지 않는다는 것이다. 학습 전에 모든 검증 이미지에 대해 Grad-CAM++ Heatmap을 사전 생성하여 활용함으로써, 실제 추론 속도에 영향을 주지 않으면서도 시각적 주의 정보를 제공할 수 있다. 이 방식은 기존 모델의 핵심 구조를 유지하면서, 다양한 환경에서의 특징 학습 능력을 높일 수 있다.

2.3 학습 환경 및 설정

실제 모바일 환경에서 발생할 수 있는 다양한 시각 조건을 반영하기 위해, AIHub에서 제공하는 서로 다른 조명에서 촬영한 얼굴 이미지를 활용하여 실험 조건에 맞는 데이터셋을 구성하였다. 전체 데이터셋에는 총 900명의 사람이 포함되어 있으며, 각 인물 파일에는 90장의 얼굴 이미지가 있다. 이 중 학습용 데이터에는 밝은 조명과 중간 조명에서 촬영된 사진이 각각 30장씩, 총 60장이 할당되었다. 검증용 데이터에는 어두운 조명에서 촬영된 사진 30장을 배치하였다. 최종적으로 학습용 54,000장과 검증용 27,000장으로 구성된 데이터셋이 완성되었다.

모든 얼굴 이미지는 전처리 과정을 거쳤다. 데이터셋에 포함된 JSON 메타데이터의 얼굴 위치 정보로 자동 크롭(Crop) 처리를 진행하였다. 이후, 모든 이미지를 224×224 크기로 정규화하여 모델 입력에 사용하였다.

학습 데이터에는 여러 증강 기법을 적용하였다. 여기에는 랜덤 리사이즈 크롭(Random resized crop), 좌우 반전(Horizontal flip), 회전(Rotation), 색상 변화(Color jittering), 블러(Blur), 랜덤 소거(Random erasing) 등이 포함되었으며, 이러한 변형은 모델이 실제 환경의 여러 시각적 변형에 대해 더 일반화된 특징을 학습할 수 있도록 하였다.

검증 데이터는 네 가지 조건으로 구성하였다. 첫째, 원본(Original) 조건은 입력 이미지를 어떠한 변형도 가하지 않은 상태로 유지한 경우이다. 둘째, 가우시안 잡음(Gaussian noise) 조건은 이미지에 무작위 노이즈를 추가하여 시각적 왜곡을 유도한 경우이다. 셋째, 부분 가림(Occlusion) 조건은 얼굴의 일부 영역을 사각형 형태로 가려 정보의 손실을 발생시킨 경우이다. 마지막으로, 밝기 조절(Brightness) 조건은 전체 이미지의 밝기를 인위적으로 조절한 경우이다. 이러한 네 가지 독립 변형 조건은 실제 시각적 외란을 시뮬레이션하고, 특정 외란이 모델 성능에 미치는 영향을 객관적으로 평가하도록 설계되었다.

제안된 Heatmap 구조에서는 Grad-CAM++ 알고리즘을 통해 사전 생성된 Heatmap을 원본 이미지와 통합하여 4채널 입력을 구성하였다. 해당 Heatmap은 원본 RGB 이미지에서 모델이 주목해야 할 중요 특징 영역을 강조하며, 이를 추가 채널로 제공함으로써 모델이 얼굴 인식에 핵심적인 영역에 더욱 집중할 수 있도록 유도하였다. npy 포맷으로 저장된 Heatmap은 학습 시에 블러와 RGB 이미지와 채널 차원에서 결합하여 활용하였다.

모델 학습에는 ArcFaceLoss라는 손실 함수를 사용하였다[8][9]. ArcFaceLoss는 입력된 로짓 벡터(Logit vector)를 정규화하고, 정답 클래스에 대해서만 각도 기반의 마진(Margin)을 부여한 후, 스케일링 과정을 거쳐 최종 손실을 계산하는 방식이다. 즉, 동일 클래스 내 임베딩은 서로 가깝게, 서로 다

른 클래스 간 임베딩은 더 멀리 떨어지도록 하였다. 이는 얼굴 인식처럼 미세한 차이를 구분해야 하는 과제에서는 단순 분류보다 유사도 기반 임베딩 학습이 훨씬 효과적이다. 본 실험에서도 ArcFaceLoss를 통해 학습 안정성과 인식 정확도를 동시에 향상하고자 하였다.

이러한 데이터 구성, 전처리 방법론, 그리고 손실 함수 선택은 다양한 환경 조건에서 모델 성능을 객관적으로 비교한다. 최종적으로는 제안하는 Heatmap 기반 입력 구조가 경량 모델의 성능에 미치는 실질적 효과를 체계적으로 분석하는 데 중점을 두었다.

2.4 평가 지표 및 결과 분석

Grad-CAM++ 기반 Heatmap을 4채널 입력으로 사용하는 구조가 얼굴 인식 성능에 미치는 영향을 확인하고자 기존 Base 모델과 비교 검증하였다. Accuracy와 F1-score를 성능 평가 도구로 사용하였으며, 네 가지 환경 조건을 설정해 모델 성능을 비교하였다. 조건은 Original, Noise, Occlusion, Brightness와 같이 설정하였다.

모든 실험은 동일한 학습 환경과 데이터셋을 기반으로 진행하였다. 그 결과 ShuffleNetV2와 MobileNetV3 모델에서의 Heatmap 구조가 Base 구조보다 전반적으로 우수한 성능을 가진다는 것을 발견하였다. 특히 Occlusion이나 Brightness의 환경 조건에서 F1-score가 평균 5~10%p 이상 향상된다는 결과를 확인할 수 있었다. 이는 시각 정보가 부분적으로 손실되거나 왜곡된 환경에서 Heatmap 사용이 효과적이었다는 것이다. 반면, EfficientNet-B0 모델의 경우 기본 성능은 상대적으로 높게 유지되었지만, Heatmap 입력에 따른 성능 향상은 제한적임을 확인할 수 있었다. 이는 모델이 이미 충분한 표현력을 가지고 있기에 추가적인 시각 정보 입력의 효과가 상대적으로 작게 나타난 것으로 해석할 수 있다. 이러한 실험 결과는 파라미터가 제한된 경량 모델일수록 Heatmap이 제공하는 주의 집중 효과가 얼굴의 핵심 부위(눈, 코, 입 등)에 더 효과적으로 작용함을 시사한다. 표 1에 각 모델별 성능 비교 수치를 정리하였다.

표 1. 모델별 Condition별 F1-score 비교 (Base vs. Heatmap)

Table 1. F1-Score comparison of Base vs. Heatmap models across conditions

Model	Original	Noise	Occlusion	Brightness
MobileNetV3 (Base)	0.7859	0.2666	0.6692	0.7737
MobileNetV3 (Heatmap)	0.8163	0.212	0.6852	0.7998
ShuffleNetV2 (Base)	0.7001	0.0567	0.5089	0.6853
ShuffleNetV2 (Heatmap)	0.7851	0.069	0.6213	0.7661
EfficientNet-B0 (Base)	0.817	0.2735	0.6915	0.8072
EfficientNet-B0 (Heatmap)	0.8218	0.2715	0.6863	0.8046

그림 1은 ShuffleNetV2와 MobileNetV3의 조건별 (Base vs. Heatmap) F1-score 비교 결과를 보여 준다. ShuffleNetV2의 경우 모든 환경(Original, Noise, Occlusion, Brightness)에서 Heatmap 구조가 Base 구조보다 일관되게 높은 성능을 나타냈으며, 특히 Occlusion 조건에서는 F1-score가 0.5089에서 0.6213으로 약 11.2%p 상승하는 현저한 개선을 보였다. MobileNetV3 역시 Original과 Brightness 조건에서 각각 약 3%p 이상의 성능 향상을 기록했으며, Occlusion 조건에서도 소폭의 성능 개선이 관찰되었다. 다만, Noise 환경에서는 MobileNetV3의 Heatmap 구조가 Base보다 다소 낮은 성능을 보였으나, 학습 초기 단계에서는 오히려 더 빠른 수렴 양상을 나타내어 주목할 만한 결과를 보여 주었다.

이러한 성능 향상의 근거는 그림 2의 Grad-CAM++ 시각화 결과를 통해 확인할 수 있다. 시각화 결과, 모델이 얼굴 인식에 필수적인 눈, 코, 입 등 핵심 부위에 명확히 집중하고 있음이 드러난다. 본 연구에서는 시각화의 선명도 자체보다는 경량 모델 환경에서 Heatmap 보조 입력이 성능 향상과 어떻게 연결되는지를 정성적으로 확인하는 데 초점을 두었으며, 이에 따라 연산 효율 대비 성능 개선 효과가 가장 뚜렷하게 나타난 ShuffleNetV2를

기준 모델로 선택하였다. 해당 Heatmap은 학습 이전 단계에서 사전 생성되어 RGB 이미지와 결합된 4채널 입력으로 사용되며, 이를 통해 외란 환경에서도 모델이 얼굴의 핵심 영역을 참고할 수 있도록 한다.

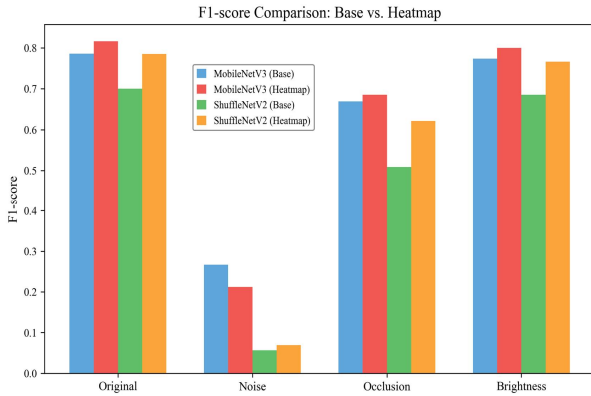


그림 1. ShuffleNetV2 및 MobileNetV3의 Condition별 F1-score 비교 (Base vs. Heatmap)

Fig. 1. F1-score comparison of ShuffleNetV2 and MobileNetV3 under each condition (Base vs. Heatmap)

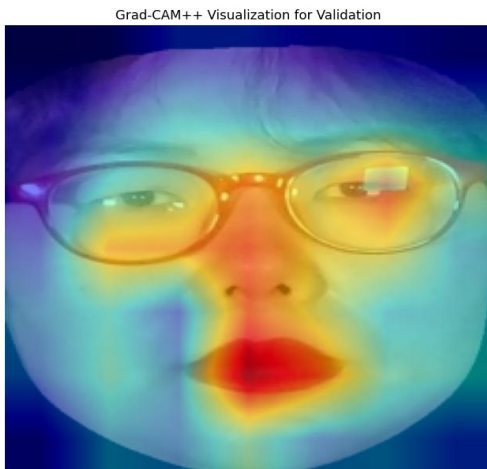


그림 2. Grad-CAM++ 기반 ShuffleNetV2 시각화 결과
Fig. 2. Grad-CAM++ visualization result of ShuffleNetV2 model

단일 지점 비교만으로는 외란 강도 증가에 따른 성능 변화를 입증하기 어렵다. 이에 본 연구에서는 외란 강도(Severity)를 단계적으로 변화시키고, 이 과정에서 변화하는 F1-score를 분석하는 Severity sweep을 수행하였다. 구체적으로, Occlusion에서는 0.1/0.2/0.3/0.4 순서로 가리는 면적 비율을 변화하였으며, Brightness에서는 0.7/0.85/1.15/1.3 순서로 조도 계수를 변화시켰다.

이어서 Noise에서는 가우시안 잡음의 표준편차를 0.05/0.1/0.15/0.2로 증가시켜 평가하였다. 이때 Occlusion 실험에서는 가리는 위치가 성능에 영향을 미칠 수 있어, 위치를 무작위로 변경하며 반복 측정 후 평균값을 사용하였다. 본 분석은 곡선의 기울기와 위치를 기준으로 “동일한 외란 강도일 때 어느 모델이 더 높은 성능을 유지하는지”와 “성능 저하가 얼마나 완만한지”에 중점을 두었다.

Brightness 실험에서는 세 모델 모두 성능이 비교적 완만하게 하락하였다. 대체로 Heatmap 모델의 성능 곡선이 Base 모델보다 일관되게 높게 위치하였다. 이는 밝기가 어두워지거나 과도하게 밝아지는 극단적인 조명 상황에서도 Heatmap 구조가 F1-score를 더 오래, 더 높게 유지한다는 것을 의미한다. 이러한 결과는 Grad-CAM++을 통해 강조된 얼굴의 핵심 부위 신호가 조명 변화에도 상대적으로 안정적이라는 해석이 가능하다. Occlusion 실험에서도 유사한 패턴을 관찰할 수 있었다. 가리는 비율을 0.1에서 0.4로 증가할수록 모든 모델의 F1-score가 감소함을 확인할 수 있었다. 그러나 ShuffleNetV2-Heatmap과 MobileNetV3-Heatmap의 경우 모든 강도에서 항상 Base 모델보다 일관되게 높은 F1-score를 기록하였다. 또한 성능 감소 곡선의 기울기도 더 완만하게 나타났다. 이는 얼굴의 일부가 가려지더라도 남아 있는 핵심 단서(예: 양쪽 눈 중 하나, 코 윤곽 등)를 Heatmap 구조가 효과적으로 활용할 수 있음을 나타낸다.

반면 Noise 실험에서는 세 모델 모두 성능이 급격하게 하락하였다. Heatmap의 이점도 위 환경에서는 상대적으로 제한적이었다. 무작위 잡음의 경우 전체 화질을 고르게 망가뜨리기에 Heatmap이 제공하는 "어디를 볼지"에 대한 안내 효과가 줄어들었다고 해석할 수 있다. 즉, 무작위 잡음이 얼굴 전체를 흐리게 만들어, Heatmap이 집중해야 할 눈·코·입 등 주요 부위조차 명확하게 구분되지 않았기 때문으로 보인다. 결과적으로 Heatmap이 위치 정보를 제공하더라도 해당 영역의 시각 정보 자체가 손상되어 유효한 구별 신호로 작동하지 못한 것이다.

위를 바탕으로 실무적 관점에서 대응할 방법은 다음과 같다. 먼저, 조도 변화나 마스크·모자·그늘과 같은 구조적 변화가 빈번한 환경일 때는 Heatmap이

유리하다. 그러나 무작위 잡음이 심한 환경에서는 사전 노이즈 억제나 데이터 증강과 같은 추가적인 대책을 먼저 사용한 후 병행되어야 한다. 이러한 강도별 성능 변화 곡선은 그림 3에 제시되어 있다. 각 그림은 조건별 F1-score와 강도 간의 관계를 모델별로 비교하여, Heatmap 곡선이 Base 곡선보다 상단에 위치하면서 더 완만하게 감소하는 패턴을 명확히 보여 준다.

학습 초기 수렴 속도 측면에서도 Heatmap 구조가 도움을 준다는 것을 확인할 수 있었다. 세 모델 모두 Epoch 3 이내에 Heatmap 구조가 Base보다 더 높은 F1-score를 기록하며 빠르게 성능이 안정화되는 경향을 보였다. 예를 들어, MobileNetV3 모델의 Original F1-score는 Epoch 3 시점에서 Base가 0.6471, Heatmap이 0.7830으로 나타났다. 이는 Grad-CAM++ 기반 Heatmap 입력이 시각적 핵심 정보에 대해 집중을 유도함으로써 학습 효율성을 향상하였다고 해석할 수 있다.

정성적 분석에서도 모델의 주의 유도가 효과적으로 이루어짐을 확인하였다. 예를 들어, ShuffleNetV2

모델의 Grad-CAM++ 시각화 결과는 눈, 코, 입 등 얼굴의 주요 부위에 명확하게 집중되는 패턴을 보였다. 이는 Occlusion이나 Brightness와 같은 변화 환경에서도 중요한 영역 인식에 도움을 주었음을 시사한다.

종합하면, 본 연구에서 제안한 Grad-CAM++ 기반 Heatmap 입력 구조는 ShuffleNetV2와 MobileNetV3과 같은 경량 모델의 경우 Occlusion 및 Brightness와 같은 환경 조건에 대한 학습 효율성을 향상하는 경향을 보여 주었다. 특히 MobileNetV3에서는 대부분의 외란 조건에서 일관된 성능 개선이 관찰되었으며, ShuffleNetV2에서도 Brightness와 같은 특정 조건에서 뚜렷한 이점이 확인되었다. EfficientNet-B0의 경우엔 기본 표현력이 높아 추가적인 이득이 상대적으로 제한적이었으나, 전체적인 경향(특히 Brightness에 대한 안정성)은 유사하다는 것을 확인할 수 있었다. 본 절의 결과는 후속 절(2.5 Ablation Study)에서 히트맵 채널을 제거하거나 교란을 줄 때 모델과 조건에 따라 성능에 어떤 영향을 미치는지에 대한 분석으로 이어진다.

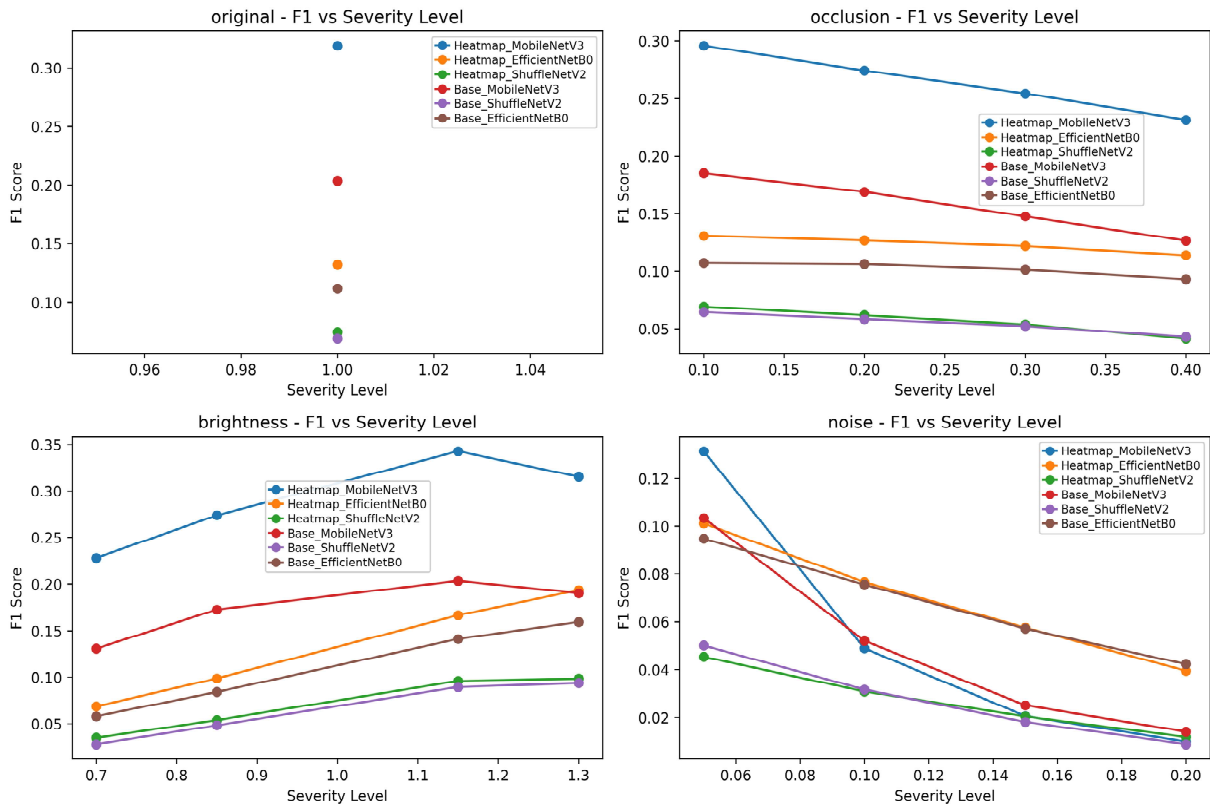


그림 3. 조건별 Severity 수준 변화에 따른 F1-score 비교 (Base vs. Heatmap)

Fig. 3. F1-score vs. severity levels for each condition (Base vs. Heatmap)

2.5 Heatmap 채널의 기여도 검증 (Ablation)

앞선 2.4절에서는 Grad-CAM++ 기반 Heatmap을 입력에 추가했을 때, 다양한 외란(Noise, Occlusion, Brightness) 환경에서 경량 모델의 성능이 더 오래 유지되는 것을 확인하였다. 본 절에서는 이러한 강건성 향상이 Heatmap 신호 자체의 기여에서 비롯된 것인지, 단순한 채널 수 증가 효과인지 명확히 구분하기 위해 Ablation 분석을 수행한다.

실험은 모델별로 실험을 진행하며 사전에 생성된 Heatmap 채널(4번째 채널)을 네 가지 방식으로 바꿔서 비교하였다. 조건은 원래 생성한 Heatmap을 사용하는 Normal, Heatmap 채널을 완전히 제거하고 전부 0으로 대체한 Zero, Heatmap을 무작위 값으로 채운 Random, 배치 내 다른 샘플의 Heatmap으로 교체한 Shuffle이 있다. 각 설정 목적은 다음과 같다. Zero는 Heatmap이 없을 때의 기준선 역할을 하며, Random과 Shuffle은 존재하지만 정확하지 않은 신호가 들어올 때 모델이 어떻게 반응하는지를 보기 위한 것이다. 실험 데이터셋과 전처리 과정은 2.4절과 똑같이 진행하였다. 그러나 외란 강도는 실험 목적에 맞게 Occlusion 0.30, Brightness 1.15, Noise 0.10으로 고정하였다. 이때 Occlusion은 위치가 성능에 영향을 줄 수 있어 반복 측정 후 평균값을 사용하였다. 최종적으로 Accuracy와 macro F1-score를 사용하여 성능 평가를 진행하였다.

실험 결과는 모델과 조건에 따라 다양한 패턴을 보였다(표 2 참조). 첫 번째, MobileNetV3 모델에서

는 Normal 조건이 Zero보다 항상 F1-score가 높은 것을 확인하였다(Occlusion: 0.2542 vs 0.2511, Brightness: 0.3426 vs 0.3401, Noise: 0.0494 vs 0.0478). 이는 Heatmap이 단순한 채널 증가 효과를 넘어 실제로 유효한 정보를 제공함을 보여 준다. 반면 EfficientNet-B0에서는 일부 조건(Occlusion: 0.1220 vs 0.1239, Brightness: 0.1664 vs 0.1683)에서 오히려 Zero가 Normal보다 약간 높은 성능이 보였다. 이는 모델의 기본 성능이 높아 추가 채널의 정보가 덜 중요할 수 있음을 의미한다. ShuffleNetV2에서는 다시 Normal이 Zero보다 항상 높은 성능을 보여 주었다(Occlusion: 0.0533 vs 0.0503, Brightness: 0.0958 vs 0.0896, Noise: 0.0301 vs 0.0283). 따라서 경량 모델에서는 Heatmap이 확실히 효과적임을 알 수 있다.

두 번째, Random과 Shuffle 조건의 효과는 예상과는 달리 모델과 조건에 따라 다양하게 나타났다. MobileNetV3의 Occlusion 조건에서는 Normal(0.2542) > Shuffle(0.2526) > Zero(0.2511) > Random(0.2455) 순으로, Random이 가장 낮은 성능을 보였지만 Shuffle은 오히려 Zero보다 높았다. ShuffleNetV2에서는 더욱 흥미로운 패턴이 관찰되었다. Occlusion 조건에서 Shuffle(0.0537) > Random(0.0536) > Normal(0.0533) > Zero(0.0503) 순이었으며, 일부 무작위성이 오히려 성능을 높여주는 효과를 보였다. 이는 특정 조건에서 Heatmap의 일부 교란이 모델의 과적합을 막아주어 일종의 정규화 효과를 제공할 가능성을 보여 준 것으로 해석된다.

표 2. Ablation 실험에서의 모델별 F1-score 비교

Table 2. Comparison of F1-score across models in Ablation experiments

Model	Condition	Normal	Zero	Random	Shuffle
MobileNetV3	Occlusion (0.3)	0.2542	0.2511	0.2455	0.2526
MobileNetV3	Brightness (1.15)	0.3426	0.3401	0.3369	0.3426
MobileNetV3	Noise (0.1)	0.0494	0.0478	0.0493	0.0487
EfficientNet-B0	Occlusion (0.3)	0.1220	0.1239	0.1164	0.1217
EfficientNet-B0	Brightness (1.15)	0.1664	0.1683	0.1590	0.1658
EfficientNet-B0	Noise (0.1)	0.0772	0.0783	0.0761	0.0772
ShuffleNetV2	Occlusion (0.3)	0.0533	0.0503	0.0536	0.0537
ShuffleNetV2	Brightness (1.15)	0.0958	0.0896	0.0967	0.0954
ShuffleNetV2	Noise (0.1)	0.0301	0.0283	0.0316	0.0303

세 번째, 외란 조건별로도 다른 패턴이 관찰되었다. Brightness 조건에서는 MobileNetV3의 경우 Normal과 Shuffle이 비슷한 성능을 보였다(0.3426 vs 0.3426). 또한 ShuffleNetV2에서는 Random이 Normal보다 약간 더 높은 성능을 기록하였다(0.0967 vs 0.0958). Noise 조건에서는 전반적으로 성능이 낮았다. 특히 MobileNetV3에서는 Random(0.0493)이 Normal(0.0494)과 거의 성능이 동일하였다. 이는 이미 노이즈가 심한 환경에서는 추가적인 Heatmap 품질이 성능에 큰 영향을 미치지 못한다는 것을 나타낸다.

본 Ablation 연구 결과를 종합하면, Heatmap의 효과는 모델의 복잡도와 외란 조건에 따라 달라져 최적의 Heatmap 전략이 달라질 수 있음을 보여 준다. 특히 Normal이 Zero보다 일관되게 높은 결과가 다수 관찰되었으며, 이는 성능 향상이 단순한 채널 수 증가가 아니라 Heatmap 신호 자체의 기여에서 비롯됨을 보여 준다. 대체로 MobileNetV3와 같은 중간 규모 모델에서는 Normal Heatmap이 항상 안정적인 성능 개선에 도움을 주었다. 그러나 ShuffleNetV2와 같은 더 경량화된 모델에서는 일부 조건에서 Random이나 Shuffle이 오히려 도움이 되는 경우도 존재하였다. 반면 EfficientNet-B0와 같이 기본 성능이 높은 모델은 Heatmap 변형에 상대적으로 둔감했다. 즉, 모델이 단순할수록 Heatmap 품질의 영향을 크게 받지만, 모델이 크면 그 영향이 작을 수 있다는 것을 짐작할 수 있다. 이번 2.5절은 ablation으로 "어떤 조건에서 왜 강건한지"에 대한 이해를 제공하였다.

2.6 LFW 벤치마크 기반 교차 데이터셋 평가

제안한 Grad-CAM++ 기반 Heatmap 입력 구조의 일반화 성능을 검증하기 위해, 널리 사용되는 얼굴 인식 벤치마크인 LFW(Labeled Faces in the Wild) 데이터셋을 활용하였다.

LFW에는 5,749명의 인물로 구성되어 있으며, 각 인물당 평균 1장부터 최대 4장 정도의 얼굴 이미지가 포함되어 있다. 본 연구에서는 이들 이미지를 무작위로 조합하여 6,000개의 비교 쌍을 구성하였다. 이때 같은 사람의 사진을 두 장 짝지은 양성 쌍과, 서로 다른 인물의 이미지를 조합한 음성 쌍을 각각 절반씩 포함해 모델이 동일 인물과 타인을 알

마나 잘 구분하는지 확인하도록 하였다. 각 이미지 쌍은 입력 전처리 과정(Resize 및 Center Crop)을 거쳐 임베딩 벡터(Embedding vector)로 변환되었다. 특히 Heatmap 입력 구조를 사용하는 경우, RGB 이미지(3채널)와 Grad-CAM++을 통해 사전 생성된 히트맵(1채널)을 결합하여 총 4채널 입력으로 구성하였다. 임베딩 추출 시에는 오인식률을 낮추고 안정성을 확보하기 위해 원본 이미지와 좌우 반전 이미지에 대한 Flip TTA(Test-Time Augmentation)를 적용하여 얻은 두 임베딩의 평균을 최종 임베딩 벡터로 사용하였다. 이후 두 벡터 간의 유사도는 코사인 유사도(Cosine similarity)로 계산되었다. 유사도 값은 클수록 동일 인물일 가능성이 크다고 판단하였으며, 모든 쌍에 대해 계산된 점수를 기반으로 ROC 곡선(Receiver Operating Characteristic curve)을 그렸다. 최종적으로 이 곡선으로부터 AUC(Area Under Curve)와 EER(Equal Error Rate) 지표를 산출하였다. 이때, AUC와 EER은 각각 모델의 판별력과 오류율을 나타내는 지표로, AUC가 높고 EER이 낮을수록 인식 성능이 우수함을 의미한다.

평가는 이전 절에서 사용한 세 가지 백본(MobileNetV3, ShuffleNetV2, EfficientNet-B0)을 동일하게 적용하여 수행하였다. 그림 4와 그림 5는 각각 AUC와 EER 비교 결과를 시각화한 그래프이다. 모든 모델에서 Heatmap 구조가 Base보다 우수한 성능을 보였으며, 특히 ShuffleNetV2은 AUC 0.8372, EER 0.2435로 가장 뛰어난 결과를 기록하였다(그림 6 참조). MobileNetV3와 EfficientNet-B0 역시 각각 AUC 0.804와 0.809를 달성하며 기존 대비 약 3~7%p의 성능 향상을 보였다. 또한, ROC 곡선 분석 결과, Heatmap 구조는 동일 임계값 범위 내에서 Base보다 더 높은 TPR(True Positive Rate)을 유지하는 것으로 확인되었다.

이는 Grad-CAM++ 기반의 보조 입력이 임베딩 공간에서 동일 인물 간 유사도는 높이고 타인 간 유사도는 낮춤으로써 특징 벡터의 분리도를 명확히 개선했음을 보여 주는 것이다. 결과적으로, 외부 벤치마크 데이터셋에서 확인된 이러한 일관된 성능 향상은 제안된 방법론이 모델 종류나 학습 데이터와 무관하게 실제 일반화 능력을 갖추었음을 입증한다.

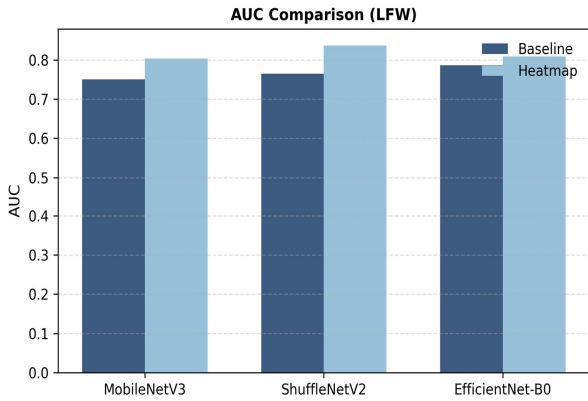


그림 4. LFW 데이터셋에서 모델별 AUC 비교 결과 (Base vs. Heatmap)

Fig. 4. Comparison of AUC across models on the LFW dataset (Base vs. Heatmap)

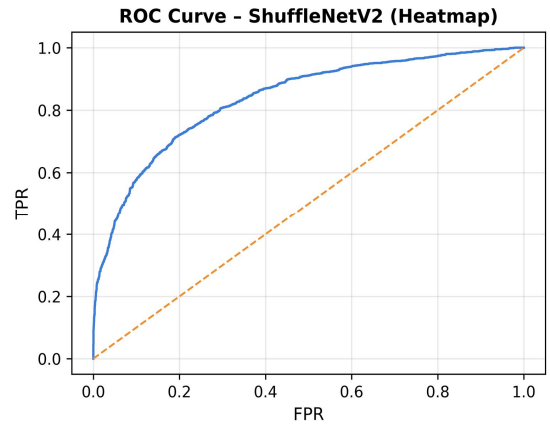


그림 6. ShuffleNetV2 Heatmap 모델의 ROC 곡선

Fig. 6. ROC curve of the ShuffleNetV2 Heatmap model

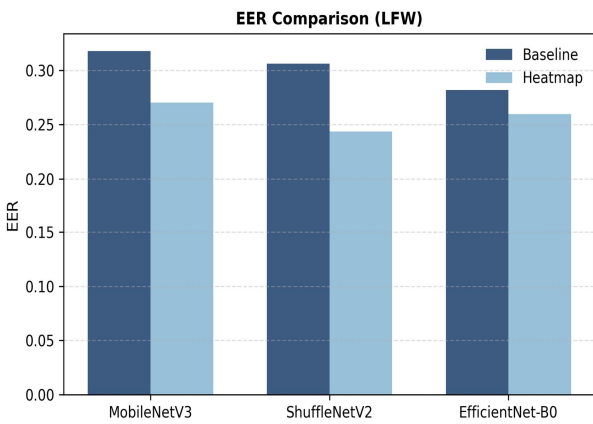


그림 5. LFW 데이터셋에서 모델별 EER 비교 결과 (Base vs. Heatmap)

Fig. 5. Comparison of EER across models on the LFW dataset (Base vs. Heatmap)

2.7 전체 시스템 구성 및 동작 흐름

본 연구에서 제안하는 얼굴 유사도 기반 매칭 시스템은 모바일 환경에서도 원활한 실시간 처리가 가능하도록 설계하였다. 전체 구조는 모바일 애플리케이션(React native), 비즈니스 로직 서버(Spring), AI 추론 서버(Flask)로 구성되며, 각 컴포넌트는 비동기 방식으로 연결된다[10].

사용자는 모바일 애플리케이션을 통해 실종자 사진 또는 유사 인물 사진을 업로드할 수 있으며, 해당 이미지 파일은 Spring 서버로 전송된다. Spring 서버는 이를 중계하여 AI 서버에 전달하고, 추론 결과를 수신한 뒤 사용자에게 다시 반환한다. 모든 요청과 응답은 RESTful API 기반의 비동기 HTTP 통신으로 이루어지며, 네트워크 지연에 따른 블로킹 없이 실시간 처리가 가능하다.

System Architecture

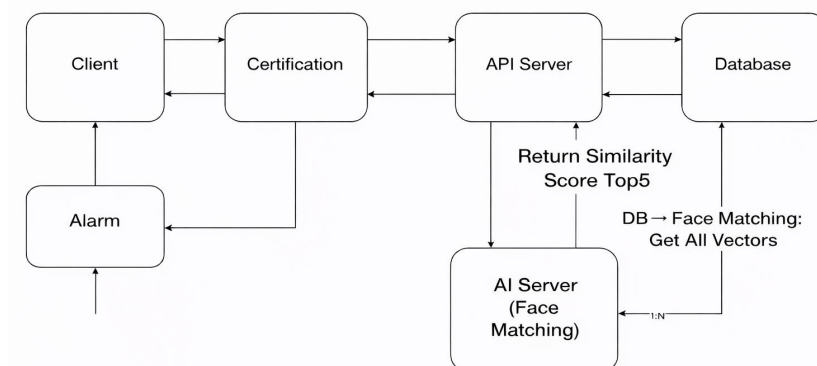


그림 7. 전체 시스템 아키텍처 흐름도

Fig. 7. Overall system architecture and processing flow

시스템은 입력 이미지와 등록 이미지 간의 유사도를 계산하여 상위 5개의 결과를 추출하고, 이에 대한 라벨 및 유사도 점수를 사용자에게 즉시 제공한다. 이미지와 결과는 서버 측 저장소에 자동 저장되어 추후 분석 및 검증 자료로 활용할 수 있다. 그림 7, React Native 기반 클라이언트가 Java Spring 비즈니스 로직 서버와 통신하며, Spring 기반 API 서버는 Flask AI 추론 서버와 비동기 방식으로 연동된다. 이미지 전송부터 유사도 결과 반환까지의 흐름과 각 구성 요소의 역할이 포함되어 있다.

2.8 AI 추론 서버의 Grad-CAM++ 기반 유사도 벡터 생성 방식

AI 추론 서버는 PyTorch 프레임워크를 기반으로 구현된 두 가지 ShuffleNetV2 계열 모델 (ShuffleNetV2-Basic, ShuffleNetV2-Heatmap)을 활용하여 입력 이미지의 시각적 주목도와 얼굴 특징을 효과적으로 추출한다. Base 모델은 RGB 3채널을 그대로 입력받고, Heatmap 모델은 Grad-CAM++으로 생성한 Heatmap을 4번째 채널로 결합하여 4채널 (RGB+H) 입력을 사용한다. AI 서버의 추론 파이프라인은 다음과 같은 단계로 진행된다. 먼저 입력 이미지를 정규화하고 224×224 크기로 조정된 후에 모델 추론을 수행하고, 마지막 합성곱 블록을 기준으로 Grad-CAM++ Heatmap을 계산한다.

생성된 Heatmap은 128×128 크기로 조정된 후 1차원 벡터(16,384차원)로 변환한다. 추출된 특징 벡터는 L2 정규화를 거친 후, 데이터베이스에 저장된 기존 벡터들과 코사인 유사도를 계산한다. 계산된 유사도 점수를 기준으로 내림차순 정렬하여 상위 5개의 가장 유사한 결과를 선별한다. 각 결과에는 해당 인물의 식별자, 유사도 점수, 그리고 신뢰도 지표가 포함되어 클라이언트에 JSON 형식으로 반환된다.

그림 8와 그림 9은 동일한 쿼리 이미지를 사용하여 모바일 애플리케이션에서 실시간으로 얼굴 인식 결과를 확인한 화면을 보여 준다. 사용자가 업로드한 사진과 함께 서버에서 반환된 상위 5개 유사 인물 결과가 표시되며, 각 결과마다 유사도 점수와 실시간 피드백 기능이 제공된다.

인식 결과






	1.	99.26
	2.	97.58
	3.	94.83
	4.	94.30
	5.	92.39

그림 8. 모바일 연동 시스템 인식 결과 (Base ShuffleNetV2, Top-5 유사도 점수)

Fig. 8. Mobile integrated system recognition results for the same input (Base ShuffleNetV2; top-5 similarity scores)

인식 결과






	1.	99.63
	2.	98.52
	3.	98.28
	4.	88.36
	5.	87.61

그림 9. 모바일 연동 시스템 인식 결과 (Heatmap ShuffleNetV2, Top-5 유사도 점수)

Fig. 9. Mobile integrated system recognition results for the same input (Heatmap ShuffleNetV2; top-5 similarity scores)

주목할 점은 Heatmap 모델이 Base 모델에 비해 상위 순위의 유사도 점수가 일관적으로 높게 나타났다는 것이다. 이러한 결과는 2.4절에서 진행한 강도(Severity sweep) 실험과도 동일한 방향의 결과를

보여 주며, 실제 서비스 환경에서도 Grad-CAM++ 기반 Heatmap 입력 구조가 체감 성능 향상에 기여함을 확인한 것이다.

추론 과정 중 발생할 수 있는 다양한 예외 상황(입력 오류, 벡터 생성 실패 등)은 Flask 서버 내에서 처리되며, 명확한 JSON 에러 메시지를 클라이언트에 반환하도록 구성되어 있다.

2.9 비즈니스 로직 서버 및 전체 시스템 연동

비즈니스 로직 서버는 Java Spring 프레임워크 기반으로 구축되어 있으며, 클라이언트와 AI 추론 서버 간의 인터페이스 역할을 담당한다. 클라이언트로부터 수신한 이미지와 메타데이터는 이 서버에서 검증된 뒤, 적절한 포맷으로 가공되어 Flask 서버에 비동기 전송된다.

클라이언트로부터 Multipart/Form-data 형식으로 전송된 이미지는 Spring 서버에서 내부 검증 및 저장 과정을 거친다. 이후 Flask 서버로 POST 요청이 비동기 방식으로 전달되며, 응답으로 받은 유사도 결과는 JSON 형태로 파싱되어 클라이언트에 즉시 반환된다.

Spring 서버는 병렬 요청에 대응하기 위해 비동기 처리를 기반으로 구현되어 있으며, 이미지 저장, 요청 실패 로그 기록, 결과 캐싱 등 다양한 예외 처리 로직이 포함되어 있다.

III. 결 론

본 연구에서는 경량화 얼굴 인식 모델에 Grad-CAM++ 기반 Heatmap을 보조 입력 채널로 추가하는 구조를 제안하였다. 기존 모델 구조를 변경하지 않고 입력 단계에서 주목 영역을 명시적으로 제공함으로써, 학습 초기 수렴 속도 및 일부 조건 변화 환경에서의 성능 향상을 확인하였다.

실험 결과, 제안된 Heatmap 구조는 모든 모델에서 Epoch 3 이내에 빠른 성능 수렴을 달성하였으며, 특히 Occlusion 및 Brightness 환경에서는 Base 모델 대비 F1-score가 최대 10%p 이상 향상되었다. 이는 주의 정보를 보조 입력으로 제공함으로써, 경량 모

델에서도 효과적인 특징 학습을 유도할 수 있음을 시사한다.

한편, Noise 환경에서는 모든 모델에서 성능 향상이 상대적으로 제한적으로 나타났다. 이는 잡음이 입력 영상 전역의 통계적 분포를 불안정하게 만들어, Heatmap이 주의해야 할 영역과 비주요 영역의 구분을 모호하게 만들기 때문이다. 특히 Gaussian 노이즈와 같은 확률적 잡음은 영상의 저주파 및 고주파 성분 모두에 영향을 주어, 모델이 학습한 시각적 패턴의 일관성을 훼손시킨다. 그 결과 Heatmap이 유효한 주의 영역을 정확히 식별하더라도 해당 정보가 분류나 임베딩 단계에서 안정적으로 활용되지 못하는 현상이 나타나는 것이다. 따라서 Noise 환경에서의 제한적 성능은 Heatmap 입력 구조의 일반화 능력뿐 아니라 모델 자체의 잡음 민감도(Noise sensitivity) 문제와도 밀접한 관련이 있다. 이러한 점은 본 연구의 주요 한계 중 하나로, 향후 노이즈 억제 또는 노이즈 적응형 구조 탐색을 통해 개선할 필요가 있음을 확인하였다.

종합적으로 제안한 Heatmap 기반 보조 입력 구조는 경량 모델의 표현력 한계를 보완하면서도, 연산 복잡도를 최소화하고 학습 효율성 및 해석 가능성을 동시에 강화할 수 있는 효과적인 방법임을 실험적으로 검증하였다. 또한, 본 구조를 포함한 전체 시스템이 실제 모바일 환경에서 실시간 얼굴 인식 및 유사도 비교에 성공적으로 적용되었으며, 비동기 처리 기반의 안정적인 서비스 운영이 가능함을 확인하였다.

향후 연구에서는 Layer-CAM, Score-CAM 등 다양한 Heatmap 기반 시각 정보를 입력 채널로 확장하거나, 모바일 환경에 최적화된 추가 경량화 모델 및 기법과의 결합 가능성을 탐구할 계획이다. 아울러, 본 연구에서 확인된 노이즈 환경의 한계를 보완하기 위해, 노이즈 특성을 반영한 데이터 증강 기법과 사전 노이즈 억제(Pre-denoising), 강도별 노이즈 시뮬레이션, 그리고 적응형 학습(Adaptive learning) 전략을 결합하는 등 더욱 강건한 학습 구조를 구축하는 방향으로 연구를 확장하고자 한다. 이를 통해 다양한 환경적 변동성에도 안정적인 성능을 유지할 수 있는 모델로 발전시켜 나갈 예정이다.

마지막으로, 본 연구의 성과는 모바일 기반 보안, 실종자 탐색, 스마트 감시 등 다양한 산업 분야에서 경량 얼굴 인식 모델의 신뢰성과 효율성을 향상하는 데 도움이 될 것이다.

References

- [1] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3", Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, pp. 1314-1324, Oct. 2019. <https://doi.org/10.1109/ICCV.2019.00140>.
- [2] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design", Proc. of the European Conference on Computer Vision (ECCV), Munich, Germany, pp. 122-138, Sep. 2018. https://doi.org/10.1007/978-3-030-01264-9_8.
- [3] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, California, USA, Vol. 97, pp. 6105-6114, Jun. 2019. <https://doi.org/10.48550/arXiv.1905.11946>.
- [4] 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", Proc. IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 618-626, Oct. 2017. <https://doi.org/10.1109/ICCV.2017.74>.
- [5] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li, "Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs", arXiv preprint, Vol. 2008, No. 02312, pp. 1-9, Aug. 2020. <https://doi.org/10.48550/arXiv.2008.02312>.
- [6] F. M. Talaat, S. A. Gamel, R. M. El-Balka, M. Shehata, and H. ZainEldin, "Grad-CAM Enabled Breast Cancer Classification with a 3D Inception-ResNet V2: Empowering Radiologists with Explainable Insights", Cancers, Vol. 16, No. 21, pp. 3668, Oct. 2024. <https://doi.org/10.3390/cancers16213668>.
- [7] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks", Proc. IEEE Winter Conf. on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, pp. 839-847, Mar. 2018. <https://doi.org/10.1109/WACV.2018.00097>.
- [8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition", Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, California, USA, pp. 4690-4699, Jun. 2019. <https://doi.org/10.1109/CVPR.2019.00482>.
- [9] F. Liu, D. Chen, F. Wang, Z. Li, and F. Xu, "Deep learning based single sample face recognition: a survey", Artificial Intelligence Review, Vol. 56, No. 3, pp. 2723-2748, 2023. <https://doi.org/10.1007/s10462-022-10240-2>.
- [10] G.-T. Byeon and C. Kang, "Design and Implementation of a Missing Pet Search System using Deep Learning-based Image Analysis", Journal of Korean Institute of Information Technology, Vol. 23, No. 5, pp. 203-209, May 2025. <https://doi.org/10.14801/jkiit.2025.23.5.203>.

저자소개

김 영 언 (Young-Eon Kim)



2022년 3월 ~ 현재 :
경상국립대학교 컴퓨터공학부
학사과정
관심분야 : 인공지능, 컴퓨터 비전,
자연어 처리

변 경 태 (Gyeong-Tae Byeon)



2024년 3월 ~ 현재 :
경상국립대학교 컴퓨터공학과
학사과정
관심분야 : 웹, 플랫폼, 인공지능

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 부교수
관심분야 : 인공지능, 시멘틱 헬스케어, 데이터마이닝