Transactions of KSAE, Vol. 34, No. 7, pp.833-841 (July, 2026)

pISSN 1225-6382 / eISSN 2234-0149
DOI http://dx.doi.org/10.7467/KSAE.2026.34.7.833

SDV 적용을 위한 단일 RGB 기반 3차원 얼굴 좌표 추정 네트워크

손 민 기[1] · 이 성 잔[2] · 박 성 근[*3]

순천향대학교 모빌리티융합보안학과[1] · 순천향대학교 스마트자동차학과[2] · 한국공학대학교 기계설계공학부[3]

3D Face Coordinate Estimation from a Single RGB Image for SDV

Minki Son[1] · Sungjin Lee[2] · Seongkeun Park[*3]

[1]*Convergence Security for Automobile, Soonchunhyang University, Chungnam 31538, Korea*
[2]*Department of Smart Automobile, Soonchunhyang University, Chungnam 31538, Korea*
[3]*Department of Mechanical Design Engineering, Korea Polytechnic University, Gyeonggi 15073, Korea*
(Received 6 January 2026 / Revised 1 March 2026 / Accepted 3 March 2026)

Abstract : With the advancement of software-defined vehicles, in-cabin occupant monitoring has become increasingly important for safety, comfort, and intelligent vehicle functions. Nonetheless, conventional approaches for three-dimensional facial landmark estimation rely on stereo or depth sensors, increasing hardware cost and system complexity. Thus, this paper proposed a lightweight RGB-only network that could estimate 3D facial landmarks of multiple occupants using a single monocular camera. The framework extended the Retina Face architecture with a 3D coordinate regression head and an auxiliary geometric fusion module, enabling direct spatial landmark regression without explicit depth estimation. A paired RGB-D dataset was collected in a real vehicle environment. The experimental results exhibit a mean 2D landmark error of 3.7 pixels and a mean 3D coordinate error of 105.8 mm. Computational analysis indicated that the model achieved approximately 24 FPS on a CPU-based platform, suggesting practical feasibility for camera-only occupant monitoring in SDVs.

Key words : In-cabin monitoring(차내 모니터링), Facial landmark detection(얼굴 랜드마크 검출), 3D coordinate estimation (3차원 좌표 추정), Monocular RGB camera(단일 RGB 카메라), Software-defined vehicle(소프트웨어 정의 차량), Deep learning(딥러닝)

Nomenclature

F_i	: i-th feature map
P_i	: feature map at the i-th level of FPN
$Conv_{i\times i}$	: i × i convolution layer
Up	: upsampling
F_{fpn}	: multi-scale feature set
$\hat{p}$	: face classification score
$\hat{b}$	: bounding box offset
$\hat{L}$	: 2D landmark output
N	: number of anchor positions
X	: predicted 3D landmark vector
X^{3D}	: predicted 3D landmark vector at level
C	: channel
W	: weight
W_{xyz}	: regression weight for 3D landmark prediction
b	: bias
b_{xyz}	: bias term for 3D landmark regression
(u, v)	: spatial location on feature map
$\hat{x}, \hat{y}, \hat{z}$	: predicted 3D coordinates of facial landmarks
r	: residual
L_{xyz}	: Charbonnier loss
A_{aux}	: auxiliary feature vector
s	: normalized scale factor of bounding box
a	: aspect ratio of bounding box

*Corresponding author, E-mail: skpark@tukorea.ac.kr

σ : standard deviation of inter-landmark distances
w, h : width and height of bounding box
p_i, p_j : positions of facial landmarks
$\bar{d}$: mean inter-landmark distance
Z_{aux} : transformed auxiliary feature vector
Φ : nonlinear transformation function
$ReLU(\cdot)$: rectified linear unit activation function
GT : ground truth
FPN : feature pyramid network
SSH : single stage headless
MLP : multi-layer perceptron
Concat : concatenate
NMS : non-maximum suppression

Subscripts

3D : three-dimensional coordinate
xyz : three-dimensional coordinate components
aux : auxiliary feature

1. 서 론

자동차 산업은 전통적인 기계 중심 구조에서 벗어나 소프트웨어 중심으로 산업 전반이 점차 재편되고 있다. SDV(Software-Defined Vehicle)는 하드웨어에 종속되지 않고 소프트웨어를 통해 차량 기능을 지속적으로 확장할 수 있는 플랫폼으로, 향후 자동차 산업 경쟁력의 핵심으로 자리 잡을 것으로 전망된다.[1)]

이러한 변화 속에서 탑승자와 차량 내부 상황을 정밀하게 인식하는 능력은 안전 확보를 넘어 개인화 서비스, 인포테인먼트 제어, 에너지 관리 등 다양한 기능과 직결되며 그 중요성이 더욱 부각되고 있다. 특히 운전자 및 동승자의 상태를 실시간으로 파악하는 인캐빈 모니터링 기술은 SDV의 지능형 서비스를 가능하게 하는 핵심 기반 기술로, 최근 연구와 산업적 관심이 집중되고 있다.

탑승자 모니터링 시스템[2)]을 구축하기 위해서는 얼굴 검출과 랜드마크 인식 등 컴퓨터 비전 기반 기술이 필수적으로 요구된다. 기존 연구는 2차원 영상에서 얼굴 위치와 특징점을 검출하는 데 초점을 맞추었으나, 공간적 위치 정보를 충분히 제공하지 못해 한계가 존재한다. 이를 보완하기 위해 깊이 카메라나 스테레오 비전을 활용하는 시도가 이루어졌지만, 고가의 센서와 복잡한 장치 구성은 대량 생산 차량에 적용하기 어렵다는 제약을 남겼다. 단일 RGB 영상을 이용한 깊이 추정 기법도 제안되었지만, 대부분 장면 전체의 깊이 맵을 생성하는 방식에 집중되어 있어 연산 효율이 떨어지고, 얼굴 랜드마크 수준에서 필요한 정밀한 3차원 좌표를 직접 제공하지 못하였다.

본 연구는 이러한 한계를 극복하기 위해, 단일 RGB 카메라만을 활용하여 차량 내 다중 탑승자의 얼굴 주요 랜드마크에 대한 3차원 좌표를 직접 회귀하는 딥러닝 기반 모델을 제안한다. (Fig. 1 참조) 제안된 모델은 얼굴 검출과 랜드마크 검출을 동시에 수행하면서 각 랜드마크의 좌표를 3차원 공간으로 회귀하는 다중 헤드 구조를 채택하였다. 불필요한 전역 깊이 맵 생성을 배제하고 필요한 지점만을 학습 대상으로 설정하여 연산 효율을 확보하여 실시간 처리에도 적합한 경량 구조를 구현하였다. 실제 차량에서 획득한 RGB-D(Depth) 데이터셋을 기반으로 학습을 진행하였으며, 조명 조건과 탑승자 배치가 변화하는 다양한 차량 내 환경에서도 안정적인 성능을 보였다.

본 논문의 구성은 다음과 같다. 제2장에서는 관련 연구를 검토하고 차별성을 논의한다. 제3장에서는 제안된 모델의 구조와 학습 방법을 상세히 기술하며, 제4장에서

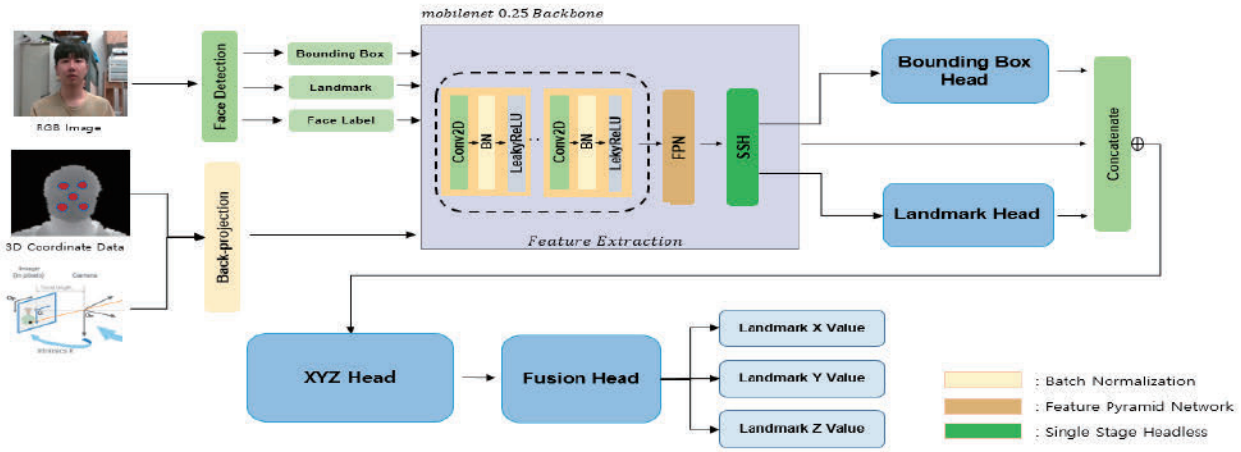

Fig. 1 Architecture of the proposed multi-head network for 3D face landmark coordinate estimation

는 실제 차량 환경에서 수집한 데이터를 기반으로 수행한 실험 결과를 제시한다. 마지막으로 제5장에서는 연구의 의의와 한계, 그리고 향후 발전 방향을 제안한다.

2. 관련 연구

2.1 얼굴 및 랜드마크 검출 관련 연구

2.1.1 YOLO5Face

YOLO5Face[3]는 YOLOv5 아키텍처를 기반으로 얼굴 영역 검출에 특화된 One-Stage 모델이다.

얼굴 검출에 최적화된 앵커 설정과 랜드마크 회귀 헤드를 추가하여, 단일 네트워크에서 얼굴 경계 상자와 눈·코·입꼬리 위치를 동시에 산출할 수 있도록 설계되었다.또한 CSPNet 기반 백본과 PAN-FPN 구조를 활용하여 다중 스케일 특징을 효과적으로 통합하며, 다양한 크기와 비율의 얼굴을 안정적으로 검출할 수 있다. 실제로 WIDER FACE 등 공개 데이터셋에서 높은 정확도와 빠른 추론 속도를 동시에 달성하였으며, 모바일 및 임베디드 환경에서도 활용 가능한 효율성을 보인다.

그러나 YOLO 계열의 근본적인 구조는 범용 객체 검출을 목표로 한 그리드 기반 설계와 손실 함수에 기초하고 있어, 픽셀 단위의 고정밀 위치 추정이 요구되는 얼굴 랜드마크 회귀에는 한계가 있다. 특히 조명 변화나 가림이 빈번한 환경에서는 프레임 간 위치 안정성이 떨어져 누적 오차가 발생하기 쉽다. 또한 기존 검출 헤드에 3차원 좌표 회귀 모듈을 직접 결합할 경우 손실 함수의 균형 조정과 학습 안정성 확보, 앵커 설계 변경 등 구조적 수정이 요구되는 제약이 존재한다.

2.1.2 BlazeFace

BlazeFace[4]는 경량화와 온디바이스 추론을 목적으로 설계된 얼굴 검출 모델로, Google의 MediaPipe 프레임워크에서 널리 사용되고 있다. SSD(Single Shot Detector) 계열의 변형된 앵커 전략과 경량 합성곱 연산을 적용하여 연산량을 크게 줄였으며, 128 × 128 해상도의 입력에서도 수백 FPS(Frame Per Second) 수준의 초고속 추론이 가능하다. 또한 YOLO5Face와 유사하게 얼굴 경계 상자와 5점 랜드마크를 동시에 예측할 수 있어, 모바일 및 엣지 디바이스에서의 실시간 추론에 적합하다.

하지만 실시간성을 확보한 대신 네트워크의 표현 용량이 제한되어 있어, 다중 스케일 특징 학습 능력과 정밀한 랜드마크 회귀 성능이 낮다는 단점이 있다. 특히 차량 내부와 같이 다양한 자세 변화, 좌석 거리 차이, 부분 가림, 강한 조명 대비 변화가 존재하는 복잡한 환경에서는 검출 안정성이 저하된다. 초경량 구조를 전제로 설계된 모델 특성상, 3D 좌표 회귀 헤드를 추가할 경우 학습 안정성이 낮아지고 태스크 간 손실 균형이 쉽게 무너지는 한계가 있다.

2.2 3차원 얼굴 추정 관련 연구

2.2.1 PRNet

PRNet[5]은 얼굴의 3차원 형상을 복원하기 위해 제안된 대표적 모델로, 단일 2D 이미지로부터 고밀도 3차원 얼굴 복원을 End-to-End 방식으로 수행한다. 이 모델은 3차원 얼굴 표면을 2D 위치 맵 형태로 표현하고 이를 회귀하는 구조를 채택하여, 랜드마크 기반 접근보다 더 풍부한 기하학 정보를 복원할 수 있다. 단일 프레임에서도 얼굴의 전체 깊이와 형상을 예측할 수 있어 포즈 변화나 표정 변형에 강인한 성능을 보였으며, 얼굴 정렬, 표정 인식, 3D 애니메이션 생성 등 다양한 분야에 활용되었다.

그러나 PRNet은 고해상도 입력에서만 우수한 성능을 보이고 연산량이 많아 모바일 및 임베디드 환경에서의 실시간 추론에는 부적합하므로, 차량 내부처럼 비정형 조명 조건이나 극단적 가림이 존재하는 환경에서는 복원된 3차원 형상의 정밀도가 저하되는 한계가 있다.

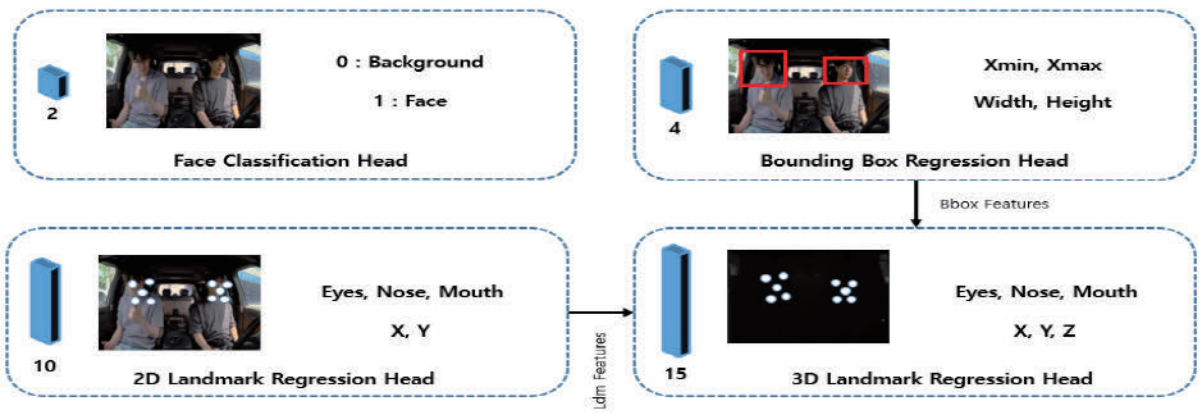

Fig. 2 Multi-head structure and functional roles of each prediction head in the proposed network

2.2.2 3DDFA_v2

3DDFA_v2[6]는 기존 3D 얼굴 복원 연구의 한계를 개선하기 위해 제안된 모델로 경량 네트워크 설계와 정규화 전략을 통해 빠르고 안정적인 3차원 얼굴 정렬을 수행한다.

얼굴의 포즈 정규화와 3D 랜드마크 추정을 동시에 수행할 수 있어 다양한 조명 · 자세 · 가림 환경에서도 안정적인 결과를 산출한다. 특히 대규모 데이터셋 기반의 사전 학습과 효율적인 피처 융합 방식을 통해 기존 방식보다 추론 속도가 빠르며, 실시간성 측면에서 실용적인 장점을 갖는다. 입력 영상의 정규화를 통해 데이터 다양성을 효과적으로 보정함으로써, 얼굴 인식 및 3D 분석의 전처리 단계로도 활용 가능하다.

다만, 얼굴의 정규화 과정을 거쳐 포즈를 보정하는 구조적 특성상 차량 내부와 같이 극단적인 회전이나 가림이 빈번한 환경에서는 정규화 과정이 불안정해져 복원 정확도가 저하될 수 있다. 또한 랜드마크 추정과 3D 형상 복원을 동시에 회귀하는 다중 모듈 구조는 손실 함수 간 균형 조정이 어려워, 조명 · 자세 변화가 심한 상황에서는 특정 태스크의 성능 저하를 초래할 가능성이 있다.

3. 연구 방법

3.1 전체 구조 개요

본 연구에서는 단일 RGB 영상을 입력으로 차량 내 다중 탑승자의 얼굴을 검출하고 주요 랜드마크의 3차원 좌표를 동시에 산출하기 위한 통합 네트워크를 설계했다. 제안된 네트워크는 RetinaFace[7]를 기반으로 하되, 기존 2차원 검출 중심 구조를 확장하여 각 랜드마크의 3차원 좌표를 직접 회귀할 수 있도록 다중 예측 구조로 구성하였다.

RetinaFace는 얼굴 검출 및 랜드마크 추정을 동시에 수행할 수 있는 단일 단계 기반 구조를 갖는다. 제안 당시 SOTA(State-of-the-Art)를 달성했으며 고성능 백본을 활용한 신뢰도 높은 GT 생성이 가능하고, 경량 백본을 적용한 실시간 추론 환경으로의 확장이 용이하다는 장점이 있다. 본 연구에서는 실시간 처리를 중점적으로 고려하여 학습 및 추론 단계 모두에서 MobileNet[8] 기반 백본을 적용하였다. 이러한 특성은 랜드마크 기반 GT 생성 과정의 신뢰성과 실시간성을 확보하는 데에도 유리하게 작용한다.

또한, RetinaFace는 C++ 기반 추론 환경 및 다양한 임베디드 플랫폼으로의 모델 포팅을 지원하여, 실제 차량 시스템과의 연동 측면에서도 높은 활용성을 제공한다. 다양한 크기와 자세의 얼굴에 대해 강인한 특징 표현이 가능한 모델로 차량 내부와 같이 시야 및 조명 변화가 큰 환경에서도 안정적인 검출 성능을 기대할 수 있다. 이러한 실시간성, 확장성 및 임베디드 환경 적용 가능성을 종합적으로 고려하여 본 연구의 기반 네트워크로 RetinaFace를 채택하였다.

전체 네트워크는 Backbone, FPN[9], SSH[10] 그리고 네 개의 예측 헤드로 구성된다. 입력 영상 $I \in \mathbb{R}^{H \times W \times 3}$는 MobileNetV1 기반의 경량 합성곱 네트워크를 통해 다단계 피처맵 $\{F_1, F_2, F_3\}$으로 변환되며, 이후 피처 융합 모듈은 서로 다른 해상도의 정보를 상호 보완적으로 통합하여 얼굴 크기나 시야각 변화와 관계없이 안정적인 공간 표현을 형성한다.

$$
\begin{aligned}
P_3 &= Conv_{1\times1}(F_3), \\
P_2 &= Conv_{1\times1}(F_2) + Up(P_3), \\
P_1 &= Conv_{1\times1}(F_1) + Up(P_2), \\
F_{fpn} &= \{P_1,\ P_2,\ P_3\}
\end{aligned}
\tag{1}
$$

FPN 융합 과정은 식 (1)과 같이 정의되며, 각 단계에서 상위 레벨 특징을 업샘플링하여 동일 해상도의 하위 레벨 특징과 결합하는 방식으로 다중 해상도 정보를 통합한다. 이를 통해 고수준 의미 정보와 저수준 공간 정보가 동시에 반영된 피처 표현이 형성되며, 생성된 다단계 피처맵은 이후 SSH 모듈을 거쳐 검출 및 회귀를 담당하는 다중 헤드의 입력으로 사용된다.

제안된 구조의 핵심은 기존 얼굴 검출 네트워크에 3차원 좌표 회귀 헤드(XYZ Head)와 보조 특성 융합 헤드(Fusion Head) 를 추가한 점이다. 이 확장 구조를 통해 네트워크는 별도의 깊이 맵을 생성하지 않고, 각 랜드마크의 3차원 위치를 직접 회귀하도록 설계되었다.

3차원 좌표 회귀 헤드는 각 단계의 FPN 피처맵을 입력으로 받아 랜드마크별 공간 좌표를 예측하고 기존의 2차원 랜드마크 헤드와 병렬로 동작한다. 이를 통해 얼굴의 전후 깊이 정보를 명시적으로 학습할 수 있으며, RGB 영상 내의 명암 대비 · 형태 · 투영 왜곡 등의 시각적 단서를 활용하여 각 랜드마크의 3차원 좌표를 추정한다.

이 헤드는 피처맵의 공간 해상도를 유지한 상태에서 각 위치의 앵커 박스에 대해 5개 주요 랜드마크의 좌표를 동시에 회귀하도록 설계되었다. 즉, 하나의 위치가 얼굴 후보를 나타낼 경우, 해당 위치에서 15차원(5개 랜드마크 × 3차원 좌표) 벡터를 직접 예측한다. 이는 기존 2차원 회귀 결과를 후처리로 3차원 보정하는 방식보다 간결하면서도 효율적인 구조적 이점을 갖는다.

추가적으로, 보조 특성 융합 헤드는 기하학적 보조 정보를 결합하여 좌표 추정의 정밀도를 향상시킨다. 이때

활용되는 보조 특성은 바운딩 박스의 상대 크기, 얼굴의 종횡비, 그리고 랜드마크 간 거리의 통계적 분산으로 구성되어 자세 변화·조명 조건·카메라 위치 차이에 따른 추정 불안정을 완화한다.

헤드의 구체적인 설계와 수식적 정의는 3.2.3~3.2.4절에서 상세히 기술한다.

3.2 다중 헤드 설계

제안된 네트워크의 출력부는 서로 다른 목적의 예측을 수행하는 다중 헤드 구조로 구성된다.

각 헤드는 공통 피처맵을 입력받아 병렬적으로 연산되며, 얼굴 존재 확률, 바운딩 박스 회귀, 2차원 랜드마크, 3차원 랜드마크 좌표를 동시에 산출한다. 이러한 다중 예측 구조는 공통 표현을 공유함으로써 학습 효율성과 일반화를 동시에 확보할 수 있다. 네트워크의 다중 헤드는 다음 네 가지로 구성된다:

3.2.1 분류 및 바운딩 박스 헤드

얼굴 분류와 바운딩 박스 헤드는 얼굴의 존재 확률과 위치 오프셋을 동시에 예측하는 모듈로, 다단계 피처맵 $F_i (i=1,2,3)$를 입력으로 받아 각각 독립적인 1×1 합성곱 계층을 통해 결과를 산출한다(식 (2)참조).

$$\hat{p_i} = Conv_{1\times1}^{(cls)}(F_i), \quad \hat{b_i} = Conv_{1\times1}^{(bbox)}(F_i) \tag{2}$$

각 앵커는 하나의 얼굴 후보를 나타내며, 두 헤드는 얼굴 검출 정확도를 결정하는 주요 학습 신호로 작용한다.

3.2.2 랜드마크 회귀 헤드

랜드마크 회귀 헤드는 2차원 공간에서 얼굴의 주요 특징점(눈, 코, 입꼬리 등)의 위치를 예측하는 모듈로, 각 피처맵 위치에 대해 10차원(5개 포인트 × 2차원 좌표) 벡터를 회귀한다. (식 (3) 참조)

$$\hat{L_i} = Conv_{1\times1}^{(lm)}(F_i), \quad \hat{L_i} \in \mathbb{R}^{N_i \times 10} \tag{3}$$

얼굴의 기하학적 구조를 세밀히 인식하도록 피처를 정렬시키는 역할을 하며, 3차원 좌표 회귀 헤드가 안정적으로 수렴할 수 있도록 보조적인 감독 신호를 제공한다.

3.2.3 3차원 좌표 회귀 헤드

3차원 좌표 회귀 헤드는 각 얼굴 랜드마크의 3차원 위치를 직접 회귀하는 핵심 모듈이다. 기존 2차원 랜드마크 결과에 의존하지 않고, RGB 피처맵 상의 지역 특징을 기반으로 각 앵커 위치에서 3차원 좌표 벡터를 예측한다. 이는 장면 전체의 깊이 맵을 생성하는 방식보다 연산 효율이 높고, 개별 얼굴 단위의 깊이 해상도를 향상시킬 수 있다.

네트워크의 각 단계에서 출력되는 FPN 피처맵을 $F_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ 라고 할 때, 좌표 추정 헤드는 각 위치에 대해 5개의 랜드마크 좌표 (x, y, z)를 직접 회귀하여 15차원 출력 벡터를 생성한다. 연산 과정은 식 (4)와 같이 정의된다.

$$\begin{aligned} X_i^{3D} &= W_{xyz}^{(i)} * F_i + b_{xyz}^{(i)}, \quad W_{xyz}^{(i)} \in \mathbb{R}^{1\times1\times C_i \times 15} \\ X_i^{3D}(u,v) &= [(\hat{x}_1, \hat{y}_1, \hat{z}_1), (\hat{x}_2, \hat{y}_2, \hat{z}_2), \cdots, (\hat{x}_5, \hat{y}_5, \hat{z}_5)] \end{aligned} \tag{4}$$

결과적으로 각 피처맵 위치 (u, v) 에서 예측된 벡터는 얼굴 크기 대비 깊이 변화를 정규화하여 서로 다른 거리의 피사체를 안정적으로 비교할 수 있도록 한다.

단계별 피처맵에서 얻은 결과는 채널 방향으로 결합하여 최종 예측으로 통합된다. (식 (5) 참조)

$$X^{3D} = Concat\left(X_1^{3D}, X_2^{3D}, X_3^{3D}\right) \tag{5}$$

이후 Prior box 및 Variance 정보를 이용한 디코딩 과정을 통해 실제 이미지 좌표계로 복원된다.

해당 구조는 좌표 추정 과정에서 불필요한 전역 시각 정보를 제거하고, 얼굴 후보 주변의 지역 피처 표현에 학습을 집중하도록 설계되었다. 이를 통해 연산 효율을 극대화하면서도 실시간 추론에 적합한 경량 구조를 유지한다. 또한 2차원 랜드마크 회귀 헤드와 병렬적으로 동작함으로써, 두 태스크가 동일한 표현 공간을 공유하며 깊이 방향의 상관관계를 공동 학습할 수 있다.

결과적으로 제안된 네트워크는 RGB 영상 내 명암 대비, 윤곽, 투영 왜곡 등 시각적 단서를 활용하여 3차원 공간 정보를 직접 회귀하며, 깊이 센서 없이도 얼굴의 전후 위치 변화를 정밀하게 추정할 수 있다.

헤드의 학습은 각 랜드마크에 대해 Charbonnier[11] 기반 회귀 손실로 정의되고, 각 랜드마크 k에 대한 잔차 $r_k = \|\hat{x}_k^{3D} - x_k^{3D}\|_2$를 이용해 식 (6)과 같이 계산된다. 여기서 ϵ는 수치적 안정성을 위한 상수이며, K는 사용된 랜드마크의 총 개수를 의미한다.

$$L_{xyz} = \frac{1}{K}\sum_{k=1}^{K}\sqrt{r_k^2 + \epsilon^2}, \quad \epsilon = 10^{-3}, \quad K = 5 \tag{6}$$

3.2.4 보조 특성 융합 헤드

보조 특성 융합 헤드는 3차원 좌표 회귀 헤드에서 예측된 결과를 추가적인 기하학적 특성과 결합하여 보정

하는 역할을 수행한다. 단일 RGB 영상에서는 조명 · 자세 · 시점 변화에 따른 깊이 단서의 불안정성이 크기 때문에, 얼굴의 형태적 · 통계적 특성을 보조 입력으로 활용해 예측의 안정성과 정밀도를 향상시킨다.

보조 입력 벡터는 바운딩 박스의 크기, 종횡비, 랜드마크 간 거리의 분산으로 구성되며, 이는 식 (7)과 같이 정의된다.

$$A_{aux} = [s, a, \sigma], \quad a = \frac{w}{h}$$
$$s = \frac{\sqrt{wh}}{\sqrt{w^2 + h^2}}, \quad \sigma = \sqrt{\frac{1}{N}\sum_{(i,j)} (\| p_i - p_j \|_2 - \bar{d})^2} \tag{7}$$

MLP를 통해 임베딩 공간으로 변환되는 동시에 피처 맵은 3 × 3 합성곱을 통해 동일 차원의 표현 F'로 투영된다. 이후 두 표현은 채널 방향으로 결합되어 1 × 1 합성곱을 통해 보정량 $\triangle X^{3D}$을 회귀한다. (식 (8) 참조)

$$Z_{aux} = \phi(A_{aux}) = ReLU(W_2 ReLU(W_1 A_{aux} + b_1) + b_2)$$
$$\Delta X^{3D} = Conv_{1\times1}([F', Z_{aux}]) \tag{8}$$

보정 과정은 얼굴 크기 · 자세 변화 및 내부 구조 변형으로 인한 깊이 예측의 불안정성을 완화하며, RGB 피처와 기하학적 벡터를 조건부로 결합함으로써 깊이 추정의 일관성과 공간적 정합성을 강화한다.

그 결과 동일 인물의 시점 변화나 조명 차이가 존재하더라도 $\hat{z}$ 예측의 분산이 크게 감소하고, 네트워크의 수렴 안정성과 일반화 성능이 향상되는 효과를 얻었다.

4. 실험 및 결과

본 연구는 SDV 환경에서의 실제 적용을 목표로 하는 얼굴 인식 시스템을 구축하기 위해 수행되었다. 기존 공개 데이터셋에서는 차량 내부에서 촬영된 RGB-D 얼굴 데이터를 제공하지 않기 때문에, 실제 차량 환경을 기반으로 한 데이터 수집 시스템을 직접 구성하였다.

기아의 EV9 차량 내부에 Intel Realsense D455 Depth Camera와 AR0233 RGB 카메라를 동기화하여 설치하였으며, 캘리브레이션된 두 센서에서 동시에 640 × 360 해상도의 RGB 영상과 깊이 맵을 취득하였다. 총 66명의 탑승자를 대상으로 얼굴 영상을 촬영하고, 프레임 단위로 분할 · 정제하여 약 53,600장의 학습용 이미지를 확보하였다. 각 이미지에는 바운딩 박스와 5개 주요 랜드마크에 대한 주석을 포함하였으며, 모든 픽셀에는 대응되는 GT 깊이 값을 함께 저장하였다. 이때 RGB 영상에 RetinaFace 기반 얼굴 검출 모델을 적용하여 랜드마크의 2차원 픽셀 좌표를 추출하고, 이를 깊이 맵 상의 대응 위치로 투영하여 각 랜드마크에 대한 깊이 값을 획득하였다. 이후 카메라 내부 파라미터를 이용하여 깊이 값을 3차원 공간 좌표 (x, y, z)로 변환하였으며, 변환된 mm 단위좌표를 학습 및 평가를 위한 GT로 활용하였다.

모델 학습은 NVIDIA RTX 4090 GPU 환경에서 수행되었으며, 최대 200 epoch까지 진행되었다. 학습 과정에서 학습 및 검증 손실 곡선을 분석한 결과, 약 100 epoch 이후부터 손실 감소 폭이 미미해지고 변동 폭이 안정화되는 경향을 보여 해당 시점을 기준으로 수렴 구간으로 판단하였다.

이에 따라 수렴 이후에도 성능의 안정성과 과적합 여부를 확인하기 위해 충분한 학습 구간을 확보하였다. 입력 이미지는 RGB 평균값 (104, 117, 123)으로 정규화되었고, 배치 크기는 32로 설정하였다. 학습률은 두 개의 감쇠 구간에서 감쇠 계수(γ)를 0.1로 적용하였으며, 초기 5 epoch 동안은 warm-up 구간을 두어 학습률을 10^{-6}에서 선형적으로 증가시켰다.

3차원 좌표 추정의 안정성과 수렴 속도를 높이기 위해 학습 과정에서 손실 가중치 점진 조정 전략을 적용하였으며, 구체적인 가중치 변화는 Table 1에 제시하였다.

4.1 데이터 전처리 및 증강 기법

모델의 일반화 성능을 높이고 다양한 조명 · 자세 · 시

Table 1 Weight schedule for 3D coordinate regression loss

Epoch range	Weight
0 - 19	0.01
20 - 39	0.05
40 - 79	0.10
80 - 119	0.15
≥ 120	0.20 (fixed)

Table 2 Data augmentation strategies applied in training

Method	Description
Distortion	Random changes in brightness (-16 to +16), Saturation (0.8-1.2×), and Hue (±8)
Rotation	Random rotation (-10° to +10°) with corresponding 3D-coordinate rotation
Crop	Random facial-region crop (scale ratio: 0.3-1.0) followed by normalization
Mirror	Horizontal flipping (p = 0.5) with left/right landmark and 3D x-coordinate swapping
Resize	Input resizing to 640 × 640 followed by mean subtraction

점 변화에 대응하기 위해, 학습 데이터에 여러 형태의 이미지 증강(Image augmentation) 기법을 적용하였다.

본 과정에서는 밝기・채도・색상 변화와 같은 광학적 왜곡뿐 아니라, 회전・크롭・좌우 반전 등 기하학적 변환을 함께 수행하여 실제 주행 환경에서 발생 가능한 시각적 다양성을 반영하였다. 적용된 주요 증강 기법은 Table 2에 요약되어 있다.

4.2 모델 성능 분석

4.2.1 평가 프로토콜 및 설정

모델의 성능은 테스트용 데이터셋 540장을 기준으로 평가하였다. 각 이미지는 단일 또는 두 명의 인물이 포함된 차량 실내 장면으로 구성되어 있으며, 탐지 결과의 신뢰성을 확보하기 위해 얼굴 분류 확률 0.9 이상의 예측 결과만 평가에 포함하였다.

평가 과정에서는 모델이 출력한 다중 얼굴 후보 중 NMS[12]를 통해 중복 영역을 제거한 최종 검출 결과를 대상으로 하였다. 이때 각 예측 얼굴 바운딩 박스는 IoU(Intersection over Union) ≥0.5 기준으로 GT와 매칭하였으며, 매칭된 쌍에 대해 바운딩 박스 위치, 2D 랜드마크, 그리고 3차원 좌표의 오차를 각각 계산하였다. 정량적 평가는 다음 세 가지 지표를 중심으로 수행하였다.

1) 2D 랜드마크 오차 (Position error) – 예측된 5개 랜드마크와 GT 간의 평균 유클리드 거리 (px 단위)

2) 바운딩 박스 오차 (BBox error) – 예측된 얼굴 영역과 GT 영역의 폭・높이 차이의 평균 (px 단위)

3) 3차원 좌표 오차 (XYZ error) – 각 랜드마크의 3차원 좌표 차이의 평균 (mm 단위)

모든 오차는 GT 주석이 포함된 CSV 데이터를 기준으로 산출하였으며, 각 예측 결과는 IoU가 가장 높은 GT 항목과 1:1로 대응시켰다. 이후 축 방향별 편차와 전체 3차원 위치 오차의 분포를 시각화하여, 모델의 공간적 안정성과 일관된 좌표 회귀 성능을 분석하였다.

4.2.2 얼굴 검출 및 2D 회귀 성능 평가

검출 결과, 모델은 모든 테스트 프레임에서 얼굴을 안정적으로 탐지하여 모든 얼굴을 검출했다. 랜드마크 위치에 대한 평균 2D 오차는 3.7 px, 바운딩 박스 크기 오차는 9.9 px로 측정되어, 탐지 네트워크와 회귀 모듈 간의 공간적 정합성이 매우 높음을 확인하였다.

또한 3D 좌표의 평균 오차는 약 105.8 mm 수준으로, 단일 RGB 입력만으로도 전반적으로 양호한 회귀 성능을 보였다. 특히 조명 변화나 시점 이동이 존재하는 상황에서도 탐지 및 랜드마크 회귀가 안정적으로 유지되어, 학습 단계에서 적용된 데이터 증강 전략과 보조 특성 융합 구조가 모델의 일반화 성능 향상에 기여한 것으로 판단된다. 관련 결과는 Fig. 3을 통해 확인할 수 있다.

4.2.3 3차원 좌표 회귀 성능 평가

Fig. 4는 테스트 세트에서 추정된 3차원 랜드마크의 축별 절대 오차 분포를 시각화한 결과이다. 평균 오차는 X축 13.4 mm, Y축 40.3 mm, Z축 95.1 mm로 나타났다. 좌우(X) 및 상하(Y) 방향의 경우 비교적 작은 오차 범위를 보였으나, 깊이(Z) 방향에서는 약 100 mm 내외의 분산이 관찰되었다. 이는 단일 RGB 영상으로부터 깊이 정보를 직접 복원해야 하는 구조적 한계에 기인하며, 조명 반사

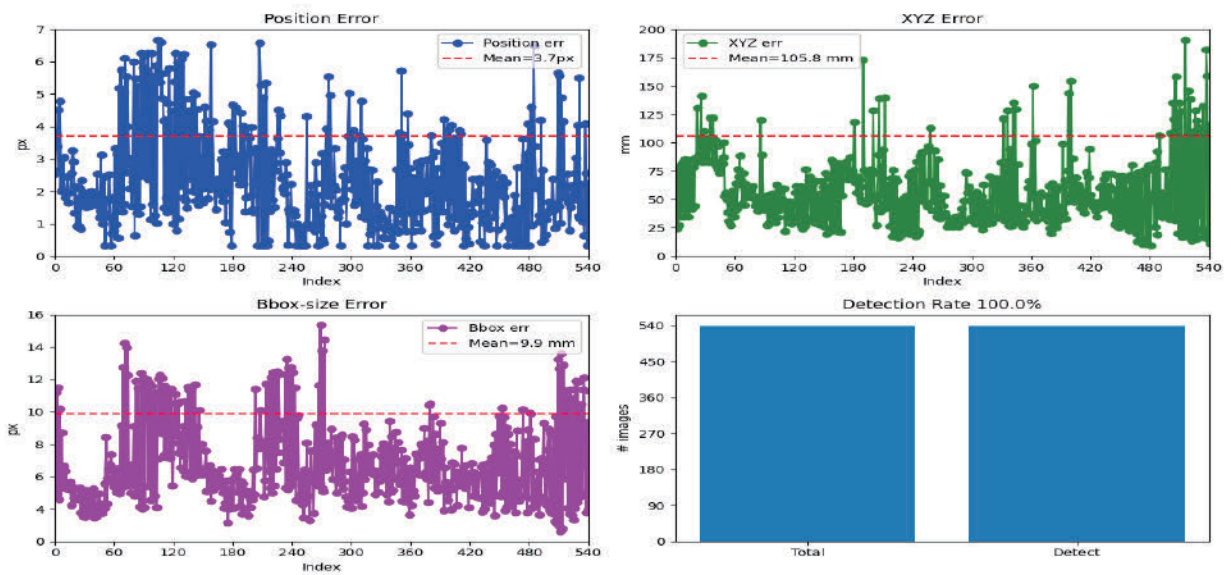

Fig. 3 Performance evaluation results of the proposed network on the test dataset

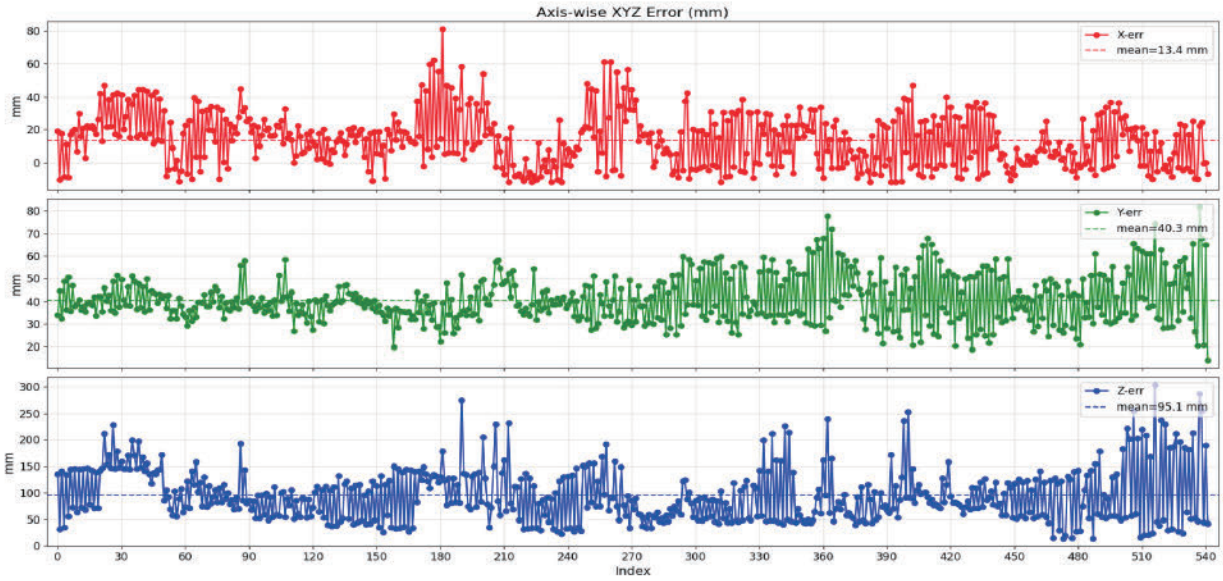

Fig. 4 Axis-wise 3D coordinate error analysis of the proposed network on the test dataset

나 얼굴 자세 변화가 깊이 단서 추출에 영향을 미친 것으로 분석된다. 그러나 보조 특성 융합 헤드의 적용을 통해 Z축 오차의 분산이 완화되고 평균 오차가 안정적으로 수렴하였다.

결과적으로, 제안된 네트워크는 깊이 센서 없이도 RGB 입력만으로 얼굴의 3차원 구조를 비교적 정확히 추정할 수 있음을 입증하였다. 이는 실시간 운전자 모니터링과 같은 차량 내 응용 시스템에서 센서 비용을 최소화하면서도 충분한 좌표 추정 성능을 확보할 수 있음을 시사한다.

4.2.4 연산 복잡도 및 실시간 처리 성능 분석

상기 성능 평가 결과를 바탕으로 제안된 네트워크의 실제 적용 가능성을 검증하기 위해 연산 복잡도와 실시간 처리 성능을 분석하였다. 본 모델은 RetinaFace 구조를 기반으로 3차원 좌표 회귀 헤드와 보조 특성 융합 모듈을 확장한 형태로 구조 확장에 따른 연산 증가와 실시간성 변화를 정량적으로 확인하기 위해 기존 RetinaFace 모델과 비교하였다.

차량 내 임베디드 환경을 고려하여 CPU 기반 추론 성능을 중심으로 평가하였으며, 모든 실험은 Intel® Core™ i7-14700F 프로세서에서 640 × 360 해상도를 기준으로 수행하였다.

추론 속도 평가는 약 1분간 연속 프레임을 처리하여 평균 FPS와 지연 시간을 산출하는 방식으로 수행하였다. 이는 순간적인 처리 속도가 아닌 일정 시간 동안 지속적으로 동작하는 환경에서의 평균적인 성능 특성을 확인하기 위함이다. 측정 결과 기존 모델은 약 36 FPS 수준을 보였고, 제안된 네트워크는 약 24 FPS 수준의 추론 속도를 나타냈다.

연산량 측면에서 제안된 네트워크는 추가 모듈의 도입에 따라 기존 RetinaFace 대비 파라미터 수와 FLOPs가 증가하였다. 그러나 평균 24 FPS 이상의 속도를 유지해 실시간 적용이 가능한 수준임을 확인하였다. 파라미터 수, FLOPs, FPS 및 지연 시간에 대한 상세 비교 결과는 Table 3에 제시하였다.

해당 결과는 RGB 입력만으로 3차원 좌표를 직접 회귀하는 구조적 확장을 고려할 때 연산 증가 대비 효율적인 설계가 이루어졌음을 의미한다. 향후 경량화 및 모델 압축 기법을 적용한다면 제한된 임베디드 환경에서도 안정적인 실시간 처리가 가능할 것으로 판단된다.

Table 3 Computational complexity and inference performance comparison between RetinaFace and the proposed network

Model	Parameters	FLOPs	FPS (CPU)
RetinaFace (Baseline)	426.6 K	585.3 M	36.83
Proposed network	1.12 M	1.68 G	24.32

5. 결 론

본 논문은 SDV 환경을 대상으로 단일 RGB 카메라만을 이용하여 차량 내 탑승자의 얼굴을 검출하고 주요 랜드마크의 3차원 좌표를 직접 회귀하는 경량 통합 네트워크를 제안하였다. RetinaFace 구조를 기반으로 3차원 좌표 회귀 헤드와 보조 특성 융합 모듈을 확장함으로써, 별

도의 깊이 센서나 깊이 맵 생성 과정 없이 RGB 영상만으로 공간 좌표를 추정할 수 있도록 설계하였다.

EV9 차량에서 수집한 RGB-D 데이터셋 기반 실험 결과, 제안된 네트워크는 다양한 조명・자세・시점 변화 조건에서도 탐지율 100 %, 평균 2D 랜드마크 오차 3.7 px, 3차원 좌표 오차 105.8 mm (X=13.4 mm, Y=40.3 mm, Z=95.1 mm)의 성능을 보였다. 특히 보조 특성 융합 모듈은 Z축 방향 오차 분산을 완화하여 깊이 추정의 안정성을 향상시키는 효과를 확인하였다.

또한 기존 RetinaFace 모델과의 비교를 통해 구조 확장에 따른 연산 증가를 정량적으로 분석하였으며, CPU 기반 환경에서 평균 24 FPS 수준의 추론 속도를 유지함을 확인하였다. 이는 3차원 좌표 직접 회귀 구조가 추가되었음에도 불구하고 실시간 운용을 고려할 수 있는 범위 내의 처리 성능을 확보하였음을 보여준다.

해당 연구는 고가의 깊이 센서를 사용하지 않고 RGB 영상만으로 3차원 얼굴 좌표를 직접 추정하는 경량 통합 구조를 제안하였다는 점에서 의의를 갖는다. 제안된 방법은 센서 비용과 시스템 복잡도를 줄이면서도 안정적인 좌표 추정 성능을 유지할 수 있음을 실험적으로 확인하였다. 향후에는 다양한 차종과 주행 환경을 포함한 확장 데이터셋을 기반으로 추가 검증을 수행하고, 시퀀스 기반 학습 기법 및 경량화 기법을 도입하여 보다 일반화된 3차원 추정 모델로 발전시킬 계획이다.

후 기

이 논문은 2025년도 한국공학대학교 학술연구진흥사업에 의하여 연구되었음(“This work was supported by the Academic Promotion System Tech University of Korea”).

References

1) K. S. Nam, “Future Mobility Innovation: Safety Assurance in Software-Defined Vehicles (SDV) and Autonomous Driving Technology,” Auto Journal, Vol.46, No.6, pp.41-44, 2024.
2) J. W. Kim, M. Jang, D. K. Kang, H. Jang and T. Oh, “Study on Requirements for Driver Monitoring in Autonomous Vehicles,” KSAE Annual Conference Proceedings, pp.2063-2065, 2023.
3) D. Qi, W. Tan, Q. Yao and J. Liu, “YOLO5Face: Why Reinventing a Face Detector,” arXiv Preprint arXiv:2105.12931, 2021.
4) V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran and M. Grundmann, “BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs,” arXiv Preprint arXiv:1907.05047, 2019.
5) Y. Feng, F. Wu, X. Shao, Y. Wang and X. Zhou, “Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network,” arXiv Preprint arXiv:1803.07835, 2018.
6) J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei and S. Z. Li, “Towards Fast, Accurate and Stable 3D Dense Face Alignment,” arXiv Preprint arXiv:2009.09960, 2020.
7) J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia and S. Zafeiriou, “RetinaFace: Single-stage Dense Face Localisation in the Wild,” arXiv Preprint arXiv:1905.00641, 2019.
8) A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto and H. Adam, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” arXiv Preprint arXiv:1704.04861, 2017.
9) T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, “Feature Pyramid Networks for Object Detection,” arXiv Preprint arXiv:1612.03144, 2017.
10) M. Najibi, P. Samangouei, R. Chellappa and L. Davis, “SSH: Single Stage Headless Face Detector,” arXiv Preprint arXiv:1708.03979, 2017.
11) J. T. Barron, “A General and Adaptive Robust Loss Function,” arXiv Preprint arXiv:1701.03077, 2017.
12) R. Rothe, M. Guillaumin and L. Van Gool, “Non-Maximum Suppression for Object Detection by Passing Messages between Windows,” Lecture Notes in Computer Science, Vol.9003, pp.290-306, 2015. DOI: 10.1007/978-3-319-16865-4_19.