특집논문 (Special Paper)
방송공학회논문지 제31권 제2호, 2026년 3월 (JBE Vol.31, No.2, March 2026)
https://doi.org/10.5909/JBE.2026.31.2.227
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

단일 RGB 영상으로부터 Normal-to-3DMM 회귀를 통한 3차원 얼굴 복원

이 호 영[a)], 장 주 용[a)‡]

3D Face Reconstruction from a Single RGB Image via Normal-to-3DMM Regression

Hoyoung Lee[a)] and Ju Yong Chang[a)‡]

요 약

본 논문에서는 단일 RGB 영상으로부터 효율적이고 정확한 3차원 얼굴 복원을 수행하기 위한 normal-to-3DMM 회귀 기반 방법을 제안한다. 최근 3차원 얼굴 복원 분야에서는 픽셀 단위의 기하 정보를 활용하여 높은 정확도를 달성하는 하이브리드 방법이 제안되었으나, 반복적인 최적화 과정으로 인해 계산 비용이 크고 추론 속도가 느리다는 한계를 가진다. 이를 해결하기 위해, 본 연구에서는 기존 하이브리드 방법인 Pixel3DMM에서 추정된 노말 맵을 입력으로 하여 3차원 가변형 얼굴 모델(FLAME)의 파라미터를 직접 회귀하는 normal-to-FLAME network(N2FNet)를 제안한다. 제안 방법은 RGB 영상에서 고차원 3DMM 파라미터를 직접 추정하는 대신, RGB-to-normal 회귀를 통해 기하학적 모호성을 완화한 후 normal-to-3DMM 회귀를 수행함으로써 학습 안정성과 계산 효율성을 동시에 향상시킨다. 또한 FLAME 파라미터 공간을 샘플링하여 노말 맵과 참값 FLAME 파라미터로 구성된 대규모 합성 데이터셋을 생성하고, 실제 데이터셋을 활용한 미세 조정을 통해 도메인 차이를 효과적으로 완화한다. Multiface 데이터셋을 포함한 다양한 실험 결과, 제안 방법은 기존 회귀 기반 방법 대비 복원 정확도를 크게 향상시키며, Pixel3DMM과 비교하여 유사한 수준의 기하학적 정확도를 유지하면서도 추론 속도 측면에서 현저한 개선을 달성함을 확인하였다.

Abstract

In this paper, we propose a normal-to-3DMM regression-based approach for efficient and accurate 3D face reconstruction from a single RGB image. Recent advances in 3D face reconstruction have introduced hybrid methods that leverage pixel-level geometric information to achieve high reconstruction accuracy. However, such methods typically rely on iterative optimization, resulting in high computational cost and slow inference speed. To address this limitation, we propose a normal-to-FLAME network (N2FNet), which directly regresses the parameters of a 3D morphable face model (FLAME) from the surface normal map estimated by an existing hybrid method, Pixel3DMM. Instead of directly predicting high-dimensional 3DMM parameters from RGB images, the proposed approach first alleviates geometric ambiguity through RGB-to-normal regression and then performs normal-to-3DMM regression, thereby improving both training stability and computational efficiency. In addition, we generate a large-scale synthetic dataset composed of normal maps and corresponding ground-truth FLAME parameters by sampling the FLAME parameter space, and further mitigate domain gaps through fine-tuning on real-world datasets. Extensive experimental results on the Multiface dataset and other benchmarks demonstrate that the proposed method significantly outperforms existing regression-based approaches in reconstruction accuracy, while achieving comparable geometric accuracy to Pixel3DMM with substantially improved inference speed.

Keyword : 3D face reconstruction, 3D morphable model, Normal-to-3DMM regression

Ⅰ. 서 론

최근 증강현실(augmented reality) 및 가상현실(virtual reality)과 같은 실감형 콘텐츠 기술의 발전과 함께 영상 기반 3차원 얼굴 복원(3D face reconstruction)에 대한 연구의 중요성이 크게 증가하고 있다. 특히 신경 방사장(neural radiance field; NeRF)이나 3차원 가우시안 스플래팅(3D Gaussian splatting)을 활용한 고품질 헤드 아바타 복원 기술이 활발히 연구되고 있으며, 이러한 기법의 대부분은 안정적인 학습과 기하학적 제어를 위하여 3차원 가변형 모델(3D morphable model; 3DMM) 기반의 얼굴 복원 결과를 사전 정보(prior) 혹은 초기값으로 활용한다. 헤드 아바타를 포함한 실감형 콘텐츠 기술은 만족스러운 사용자 경험을 제공하기 위하여 빠른 실행 속도와 높은 정확도를 동시에 요구한다.

기존의 3차원 얼굴 복원 방법은 최적화(optimization) 기반 방법, 회귀(regression) 기반 방법, 그리고 하이브리드(hybrid) 방법으로 구분할 수 있다. 첫째, 최적화 기반 방법[1,2]은 포토메트릭(photometric) 비용 함수, 랜드마크(landmark) 비용 함수 등 다양한 비용 함수의 최소화를 통해 3DMM의 파라미터를 추정한다. 이러한 방법은 높은 해석 가능성(interpretability), 기하학적 제어 가능성(controllability), 그리고 상대적으로 낮은 데이터 의존성 등의 장점을 가진다. 반면, 적절한 초기값이 필요하며 계산 비용이 커 수행 속도가 느리다는 단점을 가진다. 둘째, 회귀 기반 방법[3,4,5,6]은 입력 영상으로부터 3DMM을 직접 출력하도록 모델을 학습하는 방식이다. 이 방법은 매우 빠른 복원 속도를 제공하는 장점이 있으나, RGB 영상으로부터 고차원의 3DMM 파라미터를 직접 예측해야 하므로 학습 난이도가 높고 일반화를 위해 대규모 학습 데이터가 필요하다는 단점을 가진다. 셋째, 하이브리드 방법[7]은 회귀 기반 추정과 최적화 과정을 결합한 방식으로, 일반적으로 딥러닝 기반 회귀 모델을 통해 얼굴 기하의 초기 추정을 수행한 후 명시적인 제약 조건을 포함한 비용 함수의 최적화를 통해 추정 결과를 정제(refine)한다.

최근 제안된 Pixel3DMM[7]은 단일 영상으로부터 추정된 픽셀 단위 기하 정보를 활용하여 3DMM을 최적화하는 하이브리드 얼굴 복원 방법으로, 우수한 3차원 얼굴 복원 성능을 보이며 주목을 받고 있다. Pixel3DMM에서는 먼저 DINO[8] 기반 회귀 모델을 이용하여 입력 RGB 영상으로부터 표면 노말(normal) 맵과 UV 맵을 예측한다. 이와 같이 입력 영상으로부터 2.5차원 기하 정보를 추정하도록 설계한 것은 기존의 3DMM 파라미터를 직접 회귀하는 방식에 비해 여러 장점을 가진다. 첫째, RGB 영상으로부터 고차원의 3DMM 파라미터 공간으로의 직접적인 매핑은 비선형성이 커 학습이 어려운 반면, RGB 영상으로부터 2.5차원 기하 맵으로의 매핑은 상대적으로 학습이 용이하다. 둘째, 기하 맵에 대한 픽셀 단위의 감독(supervision)은 기존의 3DMM 파라미터나 랜드마크 기반 감독 방식에 비해 보다 강력한 기하학적 제약을 제공한다. 그러나 Pixel3DMM은 예측된 2.5차원 기하 맵을 기반으로 3DMM을 반복적으로 최적화하는 과정을 포함하므로, 계산 비용이 커 추론 속도가 상대적으로 느리다는 한계를 가진다.

이러한 반복 최적화 기반 접근은 오프라인 분석 환경에서는 효과적일 수 있으나, 방송 및 미디어 제작과 같이 실시간 또는 준실시간 처리가 요구되는 응용 환경에는 적합하지 않다. 예를 들어 실시간 아바타 구동, 방송용 얼굴 캡처, 라이브 가상 스튜디오 합성과 같은 응용에서는 얼굴 복원의 정확도뿐만 아니라 처리 지연 또한 중요한 성능 요소로 작용한다. 특히 처리 지연은 정확도 저하와 마찬가지로 사용자 경험을 크게 저해할 수 있다. 따라서 이러한 응용 환경에서는 반복 최적화 과정 없이도 정확한 기하 복원을 수행할 수 있는 효율적인 구조가 요구된다.

a) 광운대학교 전자통신공학과(Kwangwoon University)
‡ Corresponding Author : 장주용(Ju Yong Chang)
E-mail: jychang@kw.ac.kr
Tel: +82-2-940-5136
ORCID: https://orcid.org/0000-0003-3710-7314
※ This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2023-00219700, Development of FACS-compatible Facial Expression Style Transfer Technology for Digital Human, 90%) and the Excellent researcher support project of Kwangwoon University in 2025 (10%).
※ 이 논문의 연구 결과 중 일부는 한국방송·미디어공학회 2025년 추계학술대회에서 발표한 바 있음.
· Manuscript January 6, 2026; Revised February 13, 2026; Accepted February 13, 2026.

본 논문에서는 Pixel3DMM의 계산 비용이 큰 최적화 과정을 딥러닝 기반 회귀 모델로 대체함으로써 해당 방법의 한계를 극복하고자 한다. 구체적으로, 제안 방법에서는 Pixel3DMM이 출력한 노말 맵을 입력으로 하여 3DMM 파라미터를 추정하는 회귀 모델인 normal-to-FLAME network(N2FNet)를 제안한다. 여기서 FLAME[9]은 최근 널리 사용되고 있는 헤드 영역의 3DMM을 의미한다. 이러한 normal-to-3DMM 회귀는 RGB 영상으로부터 직접 3DMM 파라미터를 추정하는 기존 방식에 비해, RGB-to-normal 회귀 단계에서 상당 부분의 모호성(ambiguity)이 해소되므로 학습이 상대적으로 용이하다는 장점을 가진다. 또한 계산 효율성을 고려하여 UV 맵은 사용하지 않고 노말 맵만을 입력으로 활용하였으며, 실험 결과 노말 맵 기반 접근이 UV 맵을 사용하는 경우보다 더 우수한 복원 성능을 보임을 확인하였다. 한편, N2FNet의 학습을 위해서는 노말 맵과 이에 대응하는 FLAME 파라미터 쌍으로 구성된 학습 데이터셋이 필요하다. 이를 위해 본 논문에서는 FLAME 파라미터 공간을 샘플링하여, 노말 맵과 해당 참값(ground-truth) FLAME 파라미터를 포함하는 대규모 데이터셋을 생성하는 방법을 제안한다. 제안 방식은 노말 맵을 입력으로 사용하므로 텍스처 및 조명 렌더링이 필요 없으며, 이에 따라 기존 RGB 기반 합성 데이터셋 대비 구축 비용과 파이프라인 복잡도 측면에서 효율적인 장점을 제공한다.

정리하면, 본 논문에서는 Pixel3DMM에서 반복적 최적화를 통해 수행되던 노말 맵 기반 FLAME 파라미터 추정 과정을 normal-to-FLAME 회귀 네트워크(N2FNet)로 대체하는 새로운 프레임워크를 제안한다. 제안 방법은 Pixel3DMM의 노말 맵 기반 기하 분리 전략을 유지하면서, RGB 영상으로부터 직접 FLAME 파라미터를 회귀하는 기존 접근보다 더 정확한 FLAME 복원을 달성한다. 또한 반복적인 최적화 과정 없이 단일 피드포워드 추론만으로 동작하므로, 실시간 또는 준실시간 응용에 적합한 실용적인 속도-정확도 절충점을 제공한다.

Multiface[10] 데이터셋을 사용한 평가 결과, 제안 방법은 기존 회귀 기반 방법 대비 복원 정확도에서 향상을 보였으며, 하이브리드 방법인 Pixel3DMM과 비교하여 동등한 수준의 정확도를 유지하면서도 추론 속도 측면에서는 유의미한 개선을 달성하였다.

II. 관련 연구

1. 3차원 얼굴 복원을 위한 회귀 기반 방법

회귀 기반 3차원 얼굴 복원 방법은 입력 영상으로부터 3DMM 파라미터를 직접 예측하도록 신경망을 학습하는 방식이다. 이러한 접근은 반복적인 최적화 과정을 필요로 하지 않으므로 매우 빠른 추론 속도를 제공하며, 실시간 응용이 요구되는 증강현실 및 가상현실 환경에서 널리 활용되고 있다. 대표적인 회귀 기반 방법으로는 PRNet[3], 3DDFAv2[4], DECA[5], TokenFace[6] 등이 있으며, 단일 RGB 영상만을 이용하여 얼굴 기하를 효과적으로 복원할 수 있음을 보였다. 그러나 회귀 기반 방법은 입력 RGB 영상과 고차원의 3DMM 파라미터 공간 간의 복잡한 비선형 매핑을 학습해야 하므로 대규모 학습 데이터에 대한 의존성이 높고, 픽셀 단위의 명시적인 기하학적 제약이 부족하여 정밀한 표면 기하 복원에는 취약한 경향을 보인다.

2. 3차원 얼굴 복원을 위한 최적화 기반 방법

최적화 기반 3차원 얼굴 복원 방법은 포토메트릭 오차, 랜드마크 오차, 기하학적 정합 오차 등 명시적으로 정의된 비용함수들을 최소화함으로써 3DMM 파라미터를 추정하는 방식이다. 이러한 접근은 분석-합성(analysis-by-synthesis) 구조에 기반하여 높은 해석 가능성과 기하학적 제어 가능성을 제공한다. FlowFace[1]는 광학 흐름(optical flow)을 활용하여 프레임 간 기하학적 일관성을 에너지 항으로 모델링하고, 이를 기반으로 비디오 시퀀스 전반에 대해 3DMM 파라미터를 반복적으로 최적화하는 방법을 제안하였으며, VHAP[2] 역시 다중 프레임 공동 최적화를 통해 높은 기하학적 정확도와 시간적 일관성을 달성하였다. 그러나 이러한 최적화 기반 방법들은 반복적인 에너지 최소화 과정으로 인해 계산 비용이 크고, 입력 프레임 수가 증가할수록 추론 속도가 저하되어 실시간 응용에 적용하기에는 한계를 가진다.

3. 3차원 얼굴 복원을 위한 하이브리드 방법

하이브리드 방법은 회귀 기반 예측과 최적화 기반 정

제를 결합하여 두 접근 방식의 장점을 동시에 활용하고자 하는 방법이다. 일반적으로 딥러닝 기반 회귀 모델을 통해 얼굴 기하에 대한 초기 추정을 수행한 후, 명시적인 기하학적 제약을 포함한 비용 함수의 최적화를 통해 복원 결과를 정제한다. Pixel3DMM[7]은 이러한 하이브리드 접근의 대표적인 사례로, 단일 RGB 영상으로부터 픽셀 단위의 표면 노말 맵과 UV 맵을 예측하고 이를 기반으로 3DMM 파라미터를 최적화함으로써 높은 복원 정확도를 달성하였다. 그러나 반복적인 최적화 과정을 필수적으로 포함하므로 계산 비용이 크며, 정확도와 계산 효율성 간의 상충 관계로 인해 추론 속도 측면에서 여전히 개선의 여지가 남아 있다.

Ⅲ. 제안하는 방법

1. 제안 방법의 개요

제안하는 방법은 입력 단일 RGB 영상에 기존의 off-the-shelf 모델을 적용하여 2.5차원의 중간 기하 정보를 획득한 후 이를 기반으로 FLAME 파라미터를 회귀하는 두 단계 구조를 가진다.

첫 번째 단계에서는 한 사람의 얼굴을 포함하는 입력 RGB 영상 $I \in \mathbb{R}^{256\times256\times3}$에 Pixel3DMM[7]과 FaRL[11]을 적용하여 노말 맵 $N \in \mathbb{R}^{256\times256\times3}$과 분할 마스크(segmentation mask) $S \in \{0,1\}^{256\times256\times K}$를 획득한다. 여기서 K는 분할 영역의 클래스 수이다.

두 번째 단계에서는 제안하는 N2FNet이 앞서 획득된 노말 맵 N과 분할 마스크 S로부터 FLAME 파라미터 $\Omega = \{\beta, \theta, \psi, \gamma\}$와 카메라 파라미터 $C = \{R, t\}$를 추정한다. FLAME 파라미터 $\beta \in \mathbb{R}^{300}$, $\theta \in \mathbb{R}^{6}$, $\psi \in \mathbb{R}^{100}$, $\gamma \in \mathbb{R}^{2}$는 각각 얼굴의 형상(shape), 턱 관절의 회전(jaw rotation), 표정(expression), 그리고 눈꺼풀의 개폐 상태를 나타낸다. 이러한 파라미터는 미분 가능한 FLAME 디코더에 입력되어 5,023개의 정점들로 구성된 메쉬 $M \in \mathbb{R}^{5023\times3}$가 출력된다. 카메라의 외부 파라미터 $R \in SO(3)$과 $t \in \mathbb{R}^{3}$는 각각 얼굴의 3차원 회전과 평행이동(translation)을 나타낸다. 내부 파라미터인 초점거리(focal length) $f \in \mathbb{R}$와 주점(principal point) $c \in \mathbb{R}^{2}$는 사전에 주어졌다고 가정한다. 본 연구에서는 카메라 파라미터에 대한 정확한 추정을 목표로 하지 않는다. 카메라 파라미터는 복원된 메쉬가 입력 얼굴 영상에 잘 정렬되게끔 하는 역할만을 수행한다. 제안하는 방법의 개요는 그림 1에 제시되어 있다.

2. N2FNet

N2FNet은 먼저 입력 분할 마스크에서 얼굴 영역만을 선택하여 이진 얼굴 마스크 $S_{\text{face}} \in \{0,1\}^{256\times256}$를 구성한 후 이에 기반하여 얼굴 영역 노말 맵 $N_{\text{face}} \in \mathbb{R}^{256\times256\times3}$을 계산한다:

$$N_{\text{face}} = N \odot S_{\text{face}}, \tag{1}$$

여기서 $\odot$는 원소별 곱(element-wise multiplication)을 나

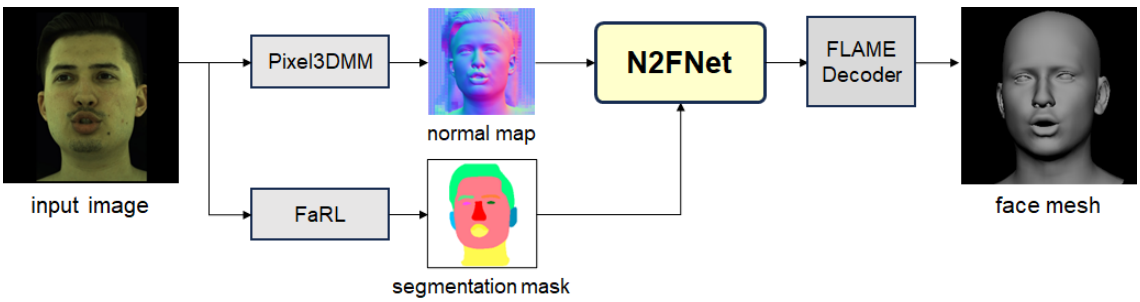

그림 1. 제안하는 3차원 얼굴 복원 방법의 개요
Fig. 1. Overview of the proposed 3D face reconstruction method

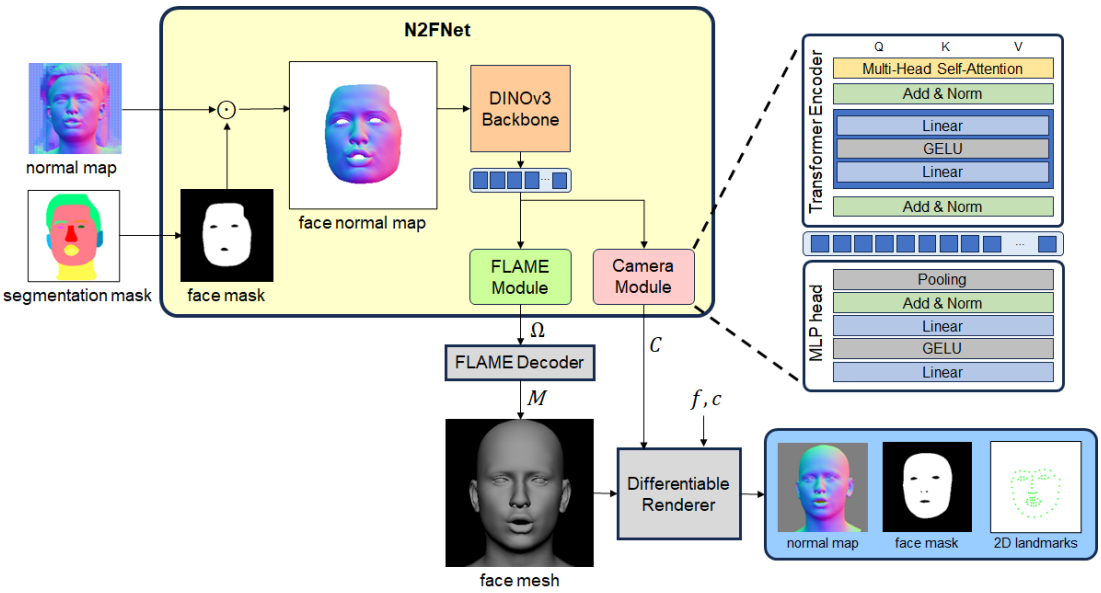

그림 2. N2FNet의 세부 구조
Fig. 2. Details of the N2FNet architecture

타낸다. 얼굴 영역 노말 맵 N_{face}은 DINOv3[12] 백본을 통해 패치 토큰 시퀀스 $T \in \mathbb{R}^{N\times d}$로 변환된다. 여기서 N과 d는 각각 토큰 개수와 차원을 나타낸다. 패치 토큰 시퀀스는 FLAME 모듈과 카메라 모듈에 입력되어 FLAME 파라미터와 카메라 파라미터가 출력된다.

FLAME 모듈에서는 T가 트랜스포머 $Block_{\text{F}}(\cdot)$를 통과한 후 풀링되어 전역 특징 $f_{\text{F}} \in \mathbb{R}^{d}$으로 변환된다. 여기서 $Block_{\text{F}}(\cdot)$는 d의 입출력 차원을 가지는 단일 레이어로 구성된 트랜스포머 인코더[13]로서 다중 헤드 셀프 어텐션(multi-head self-attention)과, GELU[16] 활성화 함수를 사용하는 순전파 신경망(feed-forward network; FFN)으로 구성된다. 이어서 다층 퍼셉트론(multi-layer perceptron; MLP) $g_{\text{F}}(\cdot)$에 의해 FLAME 파라미터 Ω가 추정된다:

$$f_{\text{F}} = Pool\big(Block_{\text{F}}(T)\big),\ \Omega = g_{\text{F}}(f_{\text{F}}). \tag{2}$$

카메라 모듈은 FLAME 모듈과 동일한 구조를 가진다. 패치 토큰 시퀀스 T는 트랜스포머 $Block_{\text{C}}(\cdot)$를 통과한 후 풀링되어 전역 특징 $f_{\text{C}} \in \mathbb{R}^{d}$로 변환된다. 이어서 MLP $g_{\text{C}}(\cdot)$에 의해 카메라 파라미터 C가 추정된다:

$$f_{\text{C}} = Pool\big(Block_{\text{C}}(T)\big),\ C = g_{\text{C}}(f_{\text{C}}). \tag{3}$$

FLAME 모듈에 의해 획득된 FLAME 파라미터 Ω는 FLAME 디코더에 입력되어 얼굴 메쉬 $M \in \mathbb{R}^{5023\times 3}$이 출력된다. 이는 카메라 모듈에 의해 획득된 카메라 파라미터 C, 초점거리 f, 주점 c에 기반해 렌더링 되어 노말 맵 N_{face}과 얼굴 분할 맵 S_{face}가 얻어진다. 또한 메쉬를 구성하는 정점들 중 선택된 68개의 3차원 랜드마크 $lmk_{3D} \in \mathbb{R}^{68\times 3}$ 역시 스크린 좌표계 위로 투영되어 2차원 랜드마크 $lmk_{2D} \in \mathbb{R}^{68\times 2}$가 얻어진다. 노말 맵, 얼굴 분할 맵, 그리고 2차원 랜드마크는 N2FNet의 학습을 위한 손실 함수를 위해 사용된다. N2FNet의 자세한 구조는 그림 2에 제시되어 있다.

3. 학습 방법

제안하는 N2FNet의 학습을 위한 손실 함수는 다음과 같다. 먼저 FLAME 모듈에서 추정된 FLAME 파라미터 Ω와 그 참값 $\Omega^{*} = \{\beta^{*}, \theta^{*}, \psi^{*}, \gamma^{*}\}$에 대한 L2 차이로 정의되는 파라미터 손실이다:

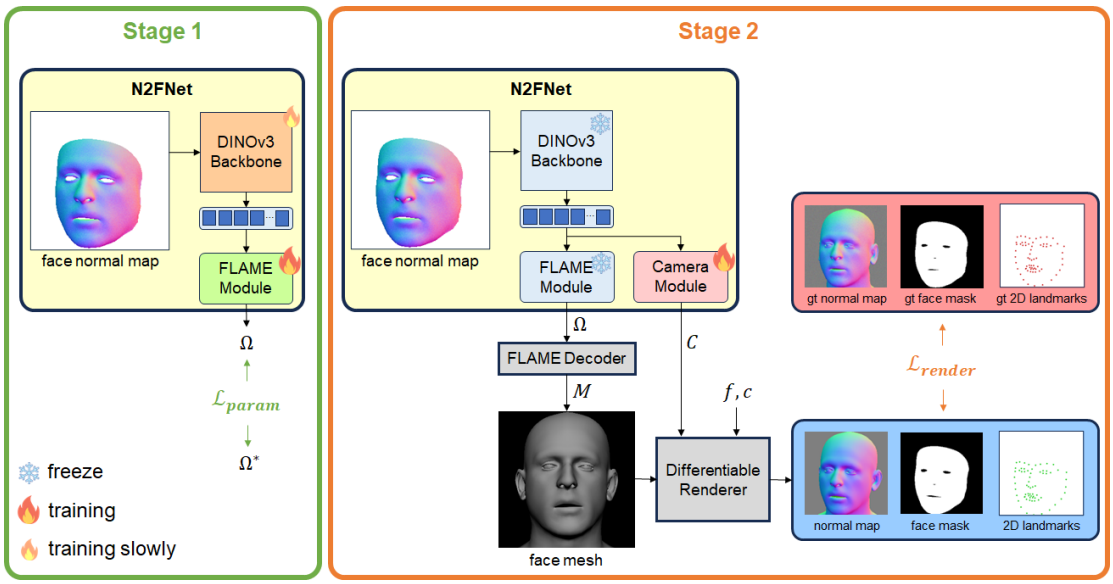

그림 3. N2FNet의 두 단계 학습 방식
Fig. 3. Two-stage training strategy for N2FNet

$$L_{\text{param}} = \lambda_\beta \|\beta - \beta^*\|_2^2 + \lambda_\theta \|\theta - \theta^*\|_2^2 + \lambda_\psi \|\psi - \psi^*\|_2^2 + \lambda_\gamma \|\gamma - \gamma^*\|_2^2. \quad (4)$$

다음은 노말 맵 N_{face}, 얼굴 분할 맵 S_{face}, 그리고 2차원 랜드마크 lmk_{2D}와 그에 대응하는 참값 N^*_{face}, S^*_{face}, lmk^*_{2D} 사이의 코사인 거리(cosine distance) 및 L2 차이로 정의되는 렌더링 손실이다:

$$L_{\text{render}} = \lambda_{N_{\text{face}}} d_{cos}(N_{\text{face}}, N^*_{\text{face}}) + \lambda_{S_{\text{face}}} \|S_{\text{face}} - S^*_{\text{face}}\|_2^2 + \lambda_{lmk_{2D}} \|lmk_{2D} - lmk^*_{2D}\|_2^2, \quad (5)$$

여기서 $d_{cos}(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}$는 코사인 거리를 의미한다.

본 연구에서는 안정적인 학습을 위해 두 단계 학습 방식을 사용한다. 첫 번째 단계에서는 백본과 FLAME 모듈만을 학습하여 입력 노말 맵으로부터 형상, 표정, 자세 정보를 안정적으로 회귀하도록 한다. 두 번째 단계에서는 백본과 FLAME 모듈을 고정하고, 카메라 모듈을 학습한다. 첫 번째 단계와 두 번째 단계에서는 각각 파라미터 손실과 렌더링 손실이 사용된다. 이러한 두 단계 학습 방식은 그림 3에 제시되어 있다.

4. 합성 데이터셋 생성 방법

본 연구에서는 N2FNet의 학습을 위해 FLAME 파라미터에 대한 샘플링과 렌더링 과정을 통해 합성 데이터셋을 구축하였다. 먼저 FLAME 파라미터 $\Omega^* = \{\beta^*, \theta^*, \psi^*, \gamma^*\}$, 카메라 파라미터 $C^*_{\text{orig}} = \{R^*_{\text{orig}}, t^*_{\text{orig}}\}$, 초점거리 f^*_{orig}, 그리고 주점 c^*_{orig}를 가우시안 분포에 기반하여 샘플링한다. 샘플링된 Ω^*를 FLAME 디코더에 입력하여 메쉬 M^*를 획득한 후 3차원 랜드마크 lmk^{orig}_{2D}를 얻는다. 샘플링된 C^*_{orig}, f^*_{orig}, 그리고 c^*_{orig}를 사용해 lmk^{orig}_{3D}를 스크린 좌표계로 투영하여 2차원 랜드마크 lmk^{orig}_{2D}를 얻는다. DECA[5]의 전처리 방법을 따라 투영된 랜드마크 lmk^{orig}_{2D}의 최소/최대 좌표를 이용해 얼굴 바운딩 박스(bounding box) $b = (x_0, y_0, w, h)$를 정의한다. 정사각형 얼굴 영역을 크롭(crop)하기 위해 패딩(padding) 계수 α를 도입하여, 바운딩 박스의 한 변 길이를 $side = \max(w, h) \cdot \alpha$로 정의한다. 얼굴 정렬(face alignment)을 통해 크롭된 정사각형 얼굴 영상에 대응하는 카메라 파라미터 $C^* = \{R^*, t^*\}$, 초점거리 f^*, 그리고 주점 c^*는 다음과 같이 획득된다:

$$R^* = R^*_{\text{orig}}, \tag{6}$$

$$t^* = t^*_{\text{orig}}, \tag{7}$$

$$f^* = f^*_{\text{orig}} \cdot \frac{S_{\text{orig}}}{side}, \tag{8}$$

$$c^* = \left(c^*_{\text{orig}} - (x_0, y_0)\right) \cdot \frac{S_{\text{out}}}{side}, \tag{9}$$

여기서 S_{orig}는 투영 및 바운딩 박스 계산이 이루어지는 원본 영상의 해상도이며, S_{out}은 N2FNet의 입력 영상 해상도(=256)를 의미한다. 샘플링된 FLAME 파라미터 Ω^*, 카메라 파라미터 C^*, 초점거리 f^*, 그리고 주점 c^*에 FLAME 디코더와 렌더러를 적용해 노말 맵 N^*, 얼굴 분할 맵 S^*, 2차원 랜드마크 $lmk^*_{2\text{D}}$를 얻을 수 있다.

Ⅳ. 실험 결과

1. 구현 세부사항

N2FNet의 백본으로 ImageNet-1K 데이터셋[14]과 대규모 웹 데이터 LVD-1689M[12]에 의해 사전 학습된 DINOv3-ViT-L/16[12]을 사용하였다. 이 백본은 $d = 1{,}024$ 차원의 패치 임베딩을 출력한다. FLAME 및 카메라 모듈의 트랜스포머 인코더에서 다중 헤드의 개수와 FFN의 은닉층(hidden layer) 차원은 각각 16과 $4 \times d = 4{,}096$으로 설정된다.

합성 데이터셋 생성 과정에서 FLAME 파라미터 Ω^*에 대한 분포는 $\beta^* \sim \mathcal{N}(0, 0.8^2)$, $\theta^*_x \sim \mathcal{N}(0, (15°)^2)$, $\theta^*_{y,z} \sim \mathcal{N}(0, (5°)^2)$, $\psi^* \sim \mathcal{N}(0, 0.5^2)$, $\gamma^* \sim \mathcal{N}(0, 0.3^2)$으로 설정된다. 여기서 얼굴의 형상($\beta^*$), 표정($\psi^*$), 눈꺼풀의 개폐 상태($\gamma^*$)는 실제 사람의 얼굴과 유사한 형태를 유지하면서도 적절한 다양성을 확보할 수 있도록 실험을 통해 결정되었다. 턱관절의 회전(θ^*)은 해부학적 범위를 준수하되, 학습의 강건성을 고려하여 설정하였다. 카메라 파라미터 C^*_{orig}의 회전 R^*_{orig}을 결정하는 오일러 각 ϕ^*는 $\phi^*_x \sim \mathcal{N}(0, (15°)^2)$, $\phi^*_y \sim \mathcal{N}(0, (20°)^2)$, $\phi^*_z \sim \mathcal{N}(0, (10°)^2)$로 설정되며, 평행이동 t^*_{orig}은 $t^*_{x,y} \sim \mathcal{N}(0, 0.02^2)$, $t^*_z \sim \mathcal{N}(1.0, 0.2^2)$의 분포로 설정된다. 여기서 t^*_{orig}의 단위는 m이다.

파라미터 손실 함수 L_{param}에 대한 λ는 $\lambda_\beta = 10$, $\lambda_\theta = 10$, $\lambda_\psi = 100$, $\lambda_\gamma = 10$으로 설정되었다. 렌더링 손실 함수 L_{render}에 대한 λ는 $\lambda_{N_{\text{face}}} = 1$, $\lambda_{S_{\text{face}}} = 10$으로 설정된다. $\lambda_{lmk_{2D}}$는 합성 데이터셋 학습에서는 100, 실제 데이터셋 학습에서는 10000으로 설정되었다. 이는 실험을 통해 경험적으로 결정된 값이다.

손실 함수의 최소화를 위해 Adam 옵티마이저[15]를 사용하였으며, 배치 크기는 16으로 설정하였다. 학습의 첫 번째 단계에서는 백본과 FLAME 모듈의 학습률(learning rate)을 각각 1e-6과 1e-3으로 설정하였다. 다만 입력 노말 맵에 대한 백본의 적응도를 높이기 위해 백본에서 가장 앞 단에 위치한 2차원 합성곱 계층(2D convolutional layer)의 학습률은 백본 내 다른 부분보다 상대적으로 큰 1e-3으로 설정하였다. 두 번째 단계에서 카메라 모듈의 학습률은 1e-4로 설정하였다. 학습률 스케줄링을 위해 선형 웜업(linear warmup)이 포함된 코사인 어닐링(cosine annealing) 방식[16]을 채택하였다. 학습은 NVIDIA RTX 4090 GPU 1개에서 총 200K 스텝 동안 진행되었다. 학습 과정에서 일정 주기마다 검증 데이터셋에 대한 평가를 수행하여 최고의 성능을 보이는 모델 가중치를 선택하여 최종 평가에 사용하였다. 추론 속도에 대한 평가는 단일 NVIDIA RTX 3090 GPU에서 수행되었다.

2. 데이터셋

본 연구에서는 N2FNet의 학습을 위해 III.4절에서 제안된 합성 데이터셋을 사용하였다. 학습 반복(iteration) 마다 샘플링을 통해 새로운 데이터셋을 생성하여 모델의 일반화 성능을 향상시키고자 하였다. 합성 데이터셋 생성을 위한 FLAME 디코더와 렌더러는 Pixel3DMM[7]의 구현을 따른다.

하지만 합성 데이터셋에 포함된 노말 맵과 실제 영상에서 추출된 노말 맵 사이에는 텍스처 디테일이나 머리카락과 같은 도메인 차이가 존재한다. 이러한 도메인 차이를 해소하기 위해 실제(real) 데이터셋을 활용한 추가적인 미세 조정(fine-tuning)을 수행한다. 이러한 미세 조정 과정에서

도 III.3절에서의 두 단계 학습 방식을 사용한다.

첫 번째 단계에서는 백본 및 FLAME 모듈의 미세 조정을 위해 고해상도 얼굴 데이터셋인 CelebA-HQ[17]를 활용하였다. 해당 데이터셋은 참값 FLAME 파라미터를 포함하지 않으므로 최적화 기반의 state-of-the-art(SOTA) 방법인 Pixel3DMM을 통해 추정된 FLAME 파라미터를 파라미터 손실을 위한 의사 참값(pseudo-GT)으로 사용하였다.

두 번째 단계에서는 카메라 모듈의 미세 조정을 위해 다양한 자세와 표정을 가진 실환경(in-the-wild) 영상들로 구성된 300W[18]를 활용하였다. Pixel3DMM과 FaRL을 통해 얻어진 노말 맵과 얼굴 마스크를 렌더링 손실을 위한 의사 참값으로 사용하였다. 또한 해당 데이터셋에 포함된 2차원 랜드마크는 렌더링 손실을 위한 참값으로 사용하였다.

제안하는 방법의 평가에는 Multiface[10] 데이터셋이 사용되었다. 이는 여러 카메라 시점에서 다양한 표정을 짓는 13명의 휴먼 객체가 특정 문장을 말하는 동작이 촬영된 다시점 얼굴 시퀀스로 구성되어 있다. 본 논문의 평가 벤치마크 구성을 위해 FlowFace[1]를 참고하였는데, 공식 코드의 부재로 인해 해당 논문에 명시된 절차를 참고하여 10명의 휴먼 객체에 대한 평가 벤치마크를 직접 재구축하여 사용하였다.

3. 평가 지표

평가 지표로는 CD(chamfer distance), MNE(mean normal error), CR(completeness rate), NME(normalized mean error), FPS(frame per second)를 사용하였다. CD의 계산은 Now 벤치마크[19]에서 제안된 방식을 따른다. 즉, 복원된 얼굴 메쉬와 참값 메쉬 사이의 강체 정렬을 수행한 후, 정렬된 두 메쉬 간의 평균 점-대-점(vertex-to-vertex) 거리가 계산된다. 이는 카메라 파라미터의 영향이 배제된 FLAME 파라미터 추정의 정확도를 나타낸다. MNE와 CR의 계산은 [20]을 따른다. MNE는 참값 메쉬의 각 정점과 이에 대응되는 복원된 메쉬 표면 상의 가장 가까운 점 사이의 법선 벡터 차이의 평균으로 정의된다. 이는 단순한 좌표 거리 오차(CD)로는 포착하기 어려운 피부의 곡률 등 표면 복원의 품질을 나타낸다. CR은 참값 메쉬의 전체 정점 중 복원된 메쉬와의 거리가 특정 임계값(2mm) 이내인 점들의 비율(%)로 정의되며, 형상이 누락 없이 얼마나 완전하게 복원되었는지를 나타낸다. NME는 카메라 파라미터 추정 및 2차원 랜드마크 정렬의 정확성을 평가하기 위해 사용된다. 이는 복원된 3차원 얼굴 랜드마크를 영상 평면에 투영한 좌표와 2차원 참값 랜드마크 좌표 사이의 유클리드 거리를 계산한 뒤, [18]의 기준에 따라 양 눈 사이의 거리로 나누어 정규화한 값이다. 마지막으로 FPS는 초당 복원된 얼굴 영상의 개수로 정의되며 모델의 추론 속도를 나타낸다.

4. 절제 실험

표 1은 각각 Ⅲ.3과 Ⅳ.2에서 제안된 두 단계 학습 방법과 실제 데이터 기반 미세 조정의 효과를 보여준다. 얼굴 영역(face)에 대한 전체적인 복원 성능(CD) 관점에서, 제안하는 두 단계 학습 방법(+two-stage training)은 백본, FLAME 모듈, 카메라 모듈을 동시에 학습하는 방법(baseline)보다 더 나은 결과(1.19mm)를 보여준다. 이는 카메라 파라미터를 함께 추론하도록 하는 것이 FLAME 파라미터 추정 성능 향상에 기여하지 못함을 의미한다. 또한 실제 데이터셋 기반의 미세 조정(+fine-tuning on real data) 역시 얼굴 복원 성능의 추가적인 향상(1.18mm)을 가져온다. 이러한 결과는 2차원 랜드마크 추정 성능(NME)에서도 관찰된다. 특별히 실제 데이터 기반 미세 조정이 6.64에서 5.50으로 성능을 약 17.2% 개선하였다. FLAME 복원 성능(CD)의 개선폭이 상대적으로 제한적인 이유는, 실제 데이터 기반 미세 조정에서 사용되는 FLAME 파라미터 감독 신호가 기존 추정 결과에 기반한 간접적인 형태이기 때문으로 해석될 수 있다. 반면 투영 기반 지표인 NME는 정렬 정확도의 향상에 직접적인 영향을 받기 때문에 상대적으로 더 큰 성능 개선이 관찰된다. 이는 앞선 FLAME 복원과 달리 정확한 카메라 파라미터 복원을 위해서는 실제 데이터를 학습에 활용하는 것이 필요함을 보여준다.

표 1. Multiface 및 300W 데이터셋에 대한 절제 실험 결과
Table 1. Ablation results on the Multiface and 300W datasets

Method	CD ↓					NME ↓
	face	eyes	nose	mouth	ears	
baseline	1.26	1.05	1.30	1.22	2.52	6.68
+ two-stage training	1.19	1.01	1.16	1.14	2.71	6.64
+ fine-tuning on real data	1.18	1.00	1.15	1.13	2.69	5.50

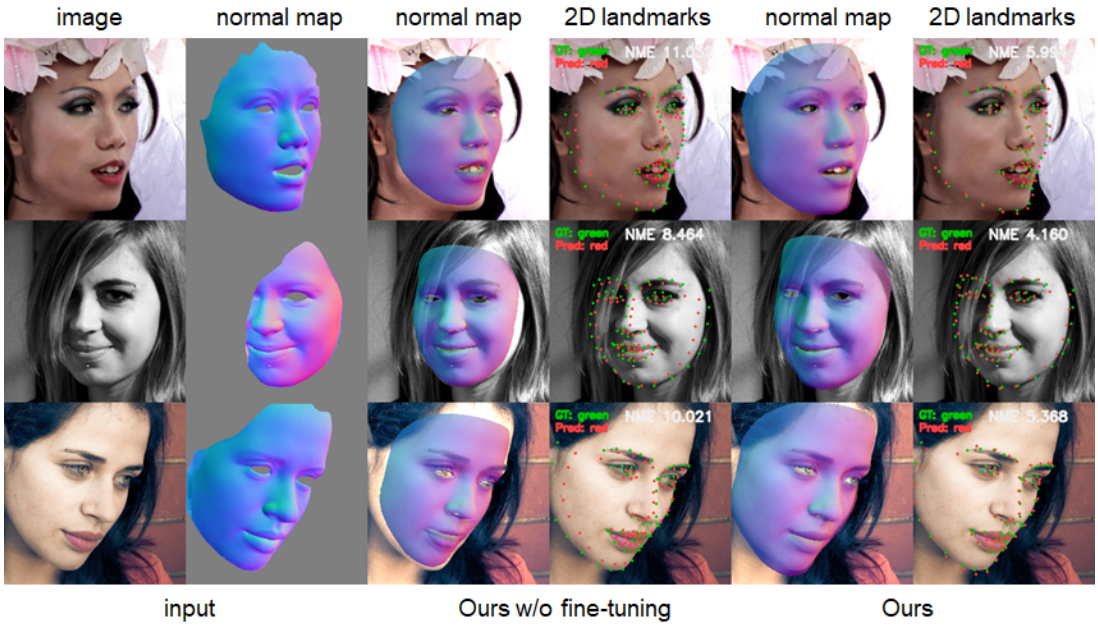

그림 4. 실제 데이터에 대한 미세 조정 절제 실험의 정성 결과
Fig. 4. Qualitative results of fine-tuning ablation on real data

그림 4는 합성 데이터셋만으로 학습된 모델(w/o fine-tuning)과, 실제 데이터셋으로 미세 조정을 수행한 모델(Ours)의 정성 결과를 함께 보여준다. 좌측부터 입력 영상, 입력 노말 맵, 합성 데이터셋 만으로 학습된 모델이 복원한 노말 맵과 2차원 랜드마크, 실제 데이터셋으로 미세 조정을 수행한 모델이 복원한 노말 맵과 2차원 랜드마크이다. 2차원 랜드마크의 경우 초록색과 빨간색은 각각 참값과 예측값을 의미한다. 시각화 결과, 머리카락에 의해 얼굴의 일부가 가려진 경우(2행)와 가려짐이 적은 일반적인 경우(1, 3행) 모두 정렬 오차가 감소하였다. 결론적으로, 실제 데이터셋에 대한 미세 조정을 통해 합성 데이터와 실제 데이터 간의 도메인 차이를 줄여 전반적인 정렬 성능이 향상되었고, 머리카락과 같은 국소적 가려짐에 대해서도 강건성이 확보되었음을 알 수 있다.

표 2는 제안하는 N2FNet의 입력 구성에 따른 복원 성능(CD)을 비교한 절제 실험 결과이다. UV 맵과 노말 맵을 둘 다 사용하는 경우(UV + normal)는 그 둘을 먼저 동일한 범위(-1,1)로 정규화 한 후 채널 단위로 단순 결합(concatenation)하였다. 실험 결과, 제안하는 방법에서 채택한 노말 맵 단독 입력(normal (ours))의 경우 1.18mm의 가장 낮은 오차를 보였다. 이러한 결과는 UV 맵에서 FLAME 파라미터로의 매핑이 제안하는 N2FNet에서의 normal-to-FLAME 매핑에 비해 상대적으로 학습하기 어렵기 때문으로 생각된다.

표 2. Multiface 데이터셋에 대한 UV 맵, 노말 맵 입력에 따른 절제 실험 결과
Table 2. Ablation results with UV map and normal map inputs on the Multiface dataset

Input	CD↓				
	face	eyes	nose	mouth	ears
UV	1.21	1.01	1.19	1.18	2.90
UV + normal	1.26	1.08	1.20	1.24	2.68
normal (ours)	1.18	1.00	1.15	1.13	2.69

5. 기존 방법과의 정량적 비교

우리는 제안 방법을 얼굴 영상으로부터 직접 3DMM 파라미터를 회귀하는 방식인 DECA[5], EMOCA[21], SMIRK[22]와, 영상으로부터 2.5차원 기하 정보를 추출한 후 이를 이용해 3DMM 파라미터를 최적화하는 Pixel3DMM[7]과 정량적으로 비교하였다. Multiface 데이터셋에 대한 평가 결과는 표 3에 제시된다.

표 3. Multiface 데이터셋에 대한 기존 방법과의 정량적 비교 실험 결과
Table 3. Quantitative comparison with existing methods on the Multiface dataset

Method	CD ↓	MNE ↓	CR ↑	FPS ↑
DECA[5]	1.36	0.266	0.77	48.38
EMOCA[21]	1.48	0.289	0.73	36.38
SMIRK[22]	1.34	0.264	0.77	100.36
Pixel3DMM[7]	1.13	0.227	0.84	0.27
Ours w/o seg	1.21	0.266	0.82	9.48
Ours	1.18	0.235	0.82	6.42

표 3의 결과에 따르면, 최적화 기반 방법인 Pixel3DMM은 CD, MNE, CR 측면에서 가장 우수한 성능을 보이며 높은 기하학적 정확도를 달성하였다. 그러나 FPS가 0.27에 불과하여 추론 속도는 가장 느린 것으로 나타났다. 반면 회귀 기반 방법인 DECA, EMOCA, SMIRK는 빠른 추론 속도를 보이지만, 기하학적 복원 정확도는 상대적으로 낮은 성능을 보였다. 제안 방법은 Pixel3DMM과 비교할 때 CD(4.42%), MNE(3.52%), CR(2.38%) 측면에서 성능이 소폭 감소하였으나, FPS는 2,378% 향상되어 추론 속도 측면에서 큰 개선을 달성하였다. 또한 회귀 기반 방법 중 가장 우수한 성능을 보인 SMIRK와 비교하면, CD(11.94%), MNE(10.98%), CR(6.60%) 측면에서 뚜렷한 정확도 향상을 보였으나 FPS는 93.60% 감소하여 추론 속도는 상대적으로 낮았다. 추가적으로, 분할 맵을 사용하지 않는 설정(Ours w/o seg)과 비교할 경우 제안 방법은 CD(2.48%)와 MNE(11.65%) 측면에서 복원 정확도가 향상되었으나, FPS는 32.28% 감소하여 추론 속도는 다소 저하되는 경향을 보였다.

6. 정성 결과

그림 5는 Multiface 데이터셋에서의 정성적 비교 결과를 보여준다. 왼쪽부터 입력 영상, 노말 맵, DECA, EMOCA, SMIRK, Pixel3DMM, 제안 방법의 복원된 FLAME 메쉬 및 컬러맵, 그리고 참값 메쉬를 순서대로 나타낸다. 컬러맵은 CD 오차를 시각화한 것으로, 빨간색에 가까울수록 오차가 크고 파란색에 가까울수록 오차가 작음을 의미한다. 정성적 비교 결과, DECA, EMOCA, SMIRK와 같은 회귀 기반 방법들은 전반적으로 입술, 광대 등 얼굴 형상의 세부적인 디테일이 일부 손실되는 경향을 보였다. 특히 DECA는 극단적인 표정(1행: 볼에 바람을 불어 넣는 표정, 2행: 입을 양 옆으로 크게 당겨 웃는 표정, 4행: 입술을 내미는 표정)을 충분히 복원하지 못하는 모습을 보인다. EMOCA는 DECA 대비 표정 표현 능력이 향상되었으나, 1행과 4행에서 볼 팽창과 입술 돌출을 완전히 재현하지 못한다. SMIRK는 극단적인 표정에 대해 DECA 및 EMOCA보다 전반적으로 정확한 복원 성능을 보이지만, 3행에서 부정확한 복원이 관찰되며 1행의 볼 팽창 역시 충분히 복원하지 못한다. Pixel3DMM은 최적화 기반 접근을 통해 전반적으로 안정적인 복원 성능을 보인다. 제안하는 방법 또한 안정적인 복원 성능을 보이며, 입술 윤곽, 얼굴 골격, 표정과 같은 세밀한 구조를 충실히 복원하는 모습을 확인할 수 있다. 제안하는 방법은 회귀 방법임에도 불구하고, 최적화 기반

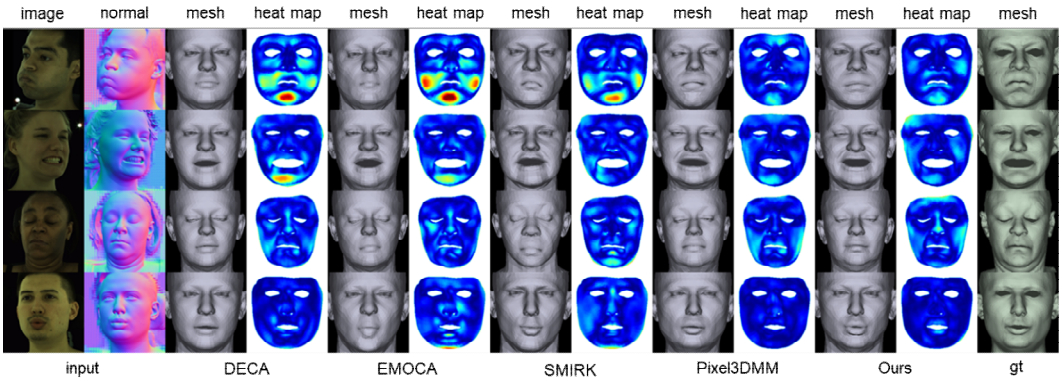

그림 5. Multiface 데이터셋에 대한 정성 결과
Fig. 5. Qualitative results on the Multiface dataset

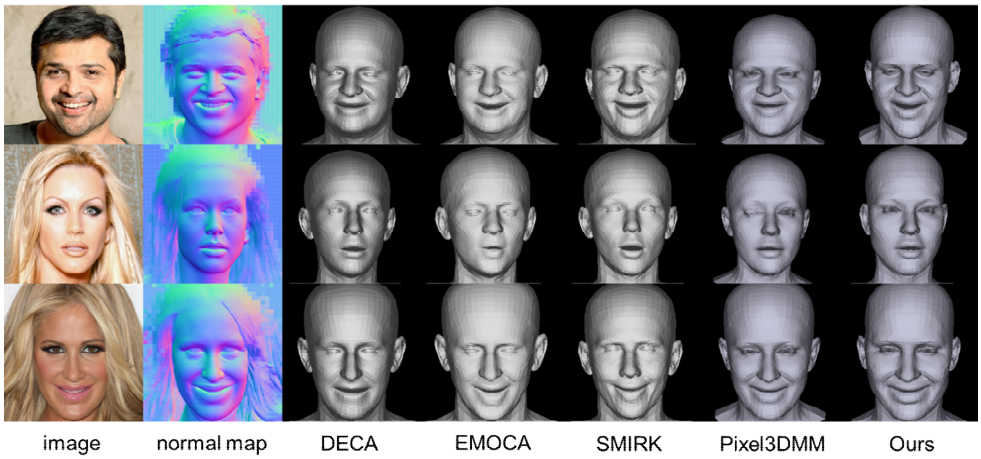

그림 6. CelebA-HQ 데이터셋에 대한 정성 결과
Fig. 6. Qualitative results on the CelebA-HQ dataset

인 Pixel3DMM과 비교했을 때 정성적으로 유사한 수준의 복원 성능을 보인다.

그림 6은 CelebA-HQ 데이터셋에서의 정성적 비교 결과를 보여준다. 왼쪽부터 입력 영상, 입력 노말 맵, DECA, EMOCA, SMIRK, Pixel3DMM, 그리고 제안 방법의 복원된 FLAME 메쉬를 순서대로 나타낸다. 정성적 결과를 통해 제안 방법이 기존 방법들과 마찬가지로 실제 환경(in the wild) 데이터셋에서도 안정적으로 동작함을 확인할 수 있다. 세부적으로, DECA는 1행과 3행에서 나타나는 활짝 웃는 표정을 충분히 복원하지 못하는 경향을 보인다. EMOCA는 DECA 대비 1행과 3행의 웃는 표정을 비교적 잘 재현하지만, 2행의 자연스러운 입 벌림 표현은 정확히 복원하지 못한다. SMIRK는 DECA와 EMOCA보다 전반적으로 우수한 표정 복원 성능을 보이나, 3행의 웃는 표정에서 볼의 부풀어 오름과 얼굴 골격이 부정확하게 복원되는 문제가 관찰된다. Pixe3DMM은 전반적으로 충실히 복원하지만, 2행에서 왼쪽 눈꺼풀에 대해 복원이 실패한 모습을 보인다. 반면 제안하는 방법은 광대와 같은 얼굴 골격, 표정과 같은 세밀한 구조를 충실히 복원하는 것을 볼 수 있다.

V. 결 론

본 논문에서는 반복적 최적화에 의존하는 기존 하이브리드 3차원 얼굴 복원 방법의 한계를 분석하고, 이를 회귀 기반 방식으로 효과적으로 대체할 수 있음을 보였다. 제안한 normal-to-3DMM 회귀 접근은 픽셀 단위 기하 정보를 유지하면서도 추론 속도를 크게 개선하여, 정확도와 효율성 간의 상충 관계를 완화한다. 실험 결과는 제안 방법이 실시간 처리가 요구되는 3차원 얼굴 복원 응용에서 실용적인 대안이 될 수 있음을 보여준다. 한편, 제안 방법은 기존 노말 맵 및 분할 맵 추출기에 의존하기 때문에 RGB 기반 회귀 모델에 비해 추가적인 연산 비용이 발생한다는 한계를 가진다. 향후 연구로는 분할 맵에 대한 의존성을 줄이거나 제거함으로써, 추론 속도를 더욱 향상시키는 방향으로 제안 방법을 확장할 계획이다.

참 고 문 헌 (References)

[1] F. Taubner, P. Raina, M. Tuli, E. W. Teh, C. Lee, and J. Huang, “3D face tracking from 2D video through iterative dense UV to image flow,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.00123

[2] S. Qian, T. Kirschstein, L. Schoneveld, D. Davoli, S. Giebenhain, and M. Nießner, “GaussianAvatars: Photorealistic head avatars with rigged 3D Gaussians,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20299 - 20309, 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.01919

[3] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, “Joint 3D face

reconstruction and dense alignment with position map regression network," European Conference on Computer Vision (ECCV), pp. 534 – 551, 2018.
doi: https://doi.org/10.1007/978-3-030-01264-9_33

[4] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3D dense face alignment," European Conference on Computer Vision (ECCV), pp. 152 – 168, 2020.
doi: https://doi.org/10.1007/978-3-030-58529-7_10

[5] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3D face model from in-the-wild images," SIGGRAPH, 2021.
doi: https://doi.org/10.1145/3450626.3459936

[6] T. Zhang, X. Chu, Y. Liu, L. Lin, Z. Yang, and Z. Xu, "Accurate 3D face reconstruction with facial component tokens," IEEE/CVF International Conference on Computer Vision (ICCV), 2023.
doi: https://doi.org/10.1109/ICCV51070.2023.00829

[7] S. Giebenhain, T. Kirschstein, M. Rünz, L. Agapito, and M. Nießner, "Pixel3DMM: Versatile screen-space priors for single-image 3D face reconstruction," arXiv preprint arXiv:2505.00615, 2025.
doi: https://doi.org/10.48550/arXiv.2505.00615

[8] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby et al., "Dinov2: Learning robust visual features without supervision," arXiv preprint arXiv:2304.07193, 2023.
doi: https://doi.org/10.48550/arXiv.2304.07193

[9] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," SIGGRAPH Asia, 2017.
doi: https://doi.org/10.1145/3130800.3130813

[10] C. H. Wuu, N. Zheng, S. Ardisson, R. Bali, D. Belko, E. Brockmeyer, L. Evans, T. Godisart, H. Ha, X. Huang, A. Hypes, T. Koska, S. Krenn, S. Lombardi, X. Luo, K. McPhail, L. Millerschoen, M. Perdoch, M. Pitts, A. Richard, J. Saragih, J. Saragih, T. Shiratori, T. Simon, M. Stewart, A. Trimble, X. Weng, D. Whitewolf, C. Wu, S. I. Yu, and Y. Sheikh, "Multiface: A dataset for neural face rendering," IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2023.
doi: https://doi.org/10.48550/arXiv.2207.11243

[11] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, and F. Wen, "General facial representation learning in a visual-linguistic manner," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.01814

[12] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski, "DINOv3," arXiv preprint arXiv:2508.10104, 2025.
doi: https://doi.org/10.48550/arXiv.2508.10104

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in Neural Information Processing Systems (NIPS), 2017.
doi: https://doi.org/10.48550/arXiv.1706.03762

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211-252, 2015.
doi: https://doi.org/10.1007/s11263-015-0816-y

[15] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization," International Conference on Learning Representations (ICLR), 2015.
doi: https://doi.org/10.48550/arXiv.1412.6980

[16] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," arXiv preprint arXiv:1606.08415, 2016.
doi: https://doi.org/10.48550/arXiv.1606.08415

[17] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," International Conference on Learning Representations (ICLR), 2018.
doi: https://doi.org/10.48550/arXiv.1710.10196

[18] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," IEEE International Conference on Computer Vision Workshops (ICCVW), 2013.
doi: https://doi.org/10.1109/ICCVW.2013.59

[19] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black, "Learning to regress 3D face shape and expression from an image without 3D supervision," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
doi: https://doi.org/10.1109/CVPR.2019.00795

[20] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "FaceScape: A large-scale high quality 3D face dataset and detailed riggable 3D face prediction," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
doi: https://doi.org/10.1109/CVPR42600.2020.00068

[21] R. Daněček, M. J. Black, and T. Bolkart, "Emoca: Emotion driven monocular face capture and animation," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), p. 20311-20322, 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.01967

[22] G. Retsinas, P. P. Filntisis, R. Daněček, V. F. Abrevaya, A. Roussos, T. Bolkart, and P. Maragos, "SMIRK: 3D facial expressions through analysis-by-neural-synthesis," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.00241

저 자 소 개

이 호 영

- 2024년 2월 : 광운대학교 전자통신공학과 학사
- 2024년 3월 ~ 현재 : 광운대학교 전자통신공학과 석사과정
- ORCID : https://orcid.org/0009-0009-2273-2652
- 주관심분야 : 컴퓨터비전 및 머신러닝

장 주 용

- 2001년 2월 : 서울대학교 전기공학부 학사
- 2008년 2월 : 서울대학교 전기컴퓨터공학부 박사
- 2008년 2월 ~ 2009년 1월 : Postdoctoral Researcher, Mitsubishi Electric Research Laboratories (MERL), US
- 2009년 4월 ~ 2011년 1월 : 삼성전자 DMC 연구소 책임연구원
- 2011년 4월 ~ 2012년 2월 : 서울대학교 BK 조교수
- 2012년 3월 ~ 2017년 2월 : 한국전자통신연구원 선임연구원
- 2024년 3월 ~ 2025년 2월 : Visiting Scholar, University of Birmingham, UK
- 2017년 3월 ~ 현재 : 광운대학교 전자통신공학과 교수
- ORCID : https://orcid.org/0000-0003-3710-7314
- 주관심분야 : 컴퓨터비전 및 머신러닝