Journal of Real-Time Image Processing (2023) 20:18
https://doi.org/10.1007/s11554-023-01257-z

ORIGINAL RESEARCH PAPER

Towards reduced dependency and faster unsupervised 3D face reconstruction

Hitika Tiwari[1,2] · Venkatesh K. Subramanian[2] · Yong-Sheng Chen[1]

Received: 8 August 2022 / Accepted: 27 December 2022 / Published online: 13 February 2023

Abstract

Recent monocular 3D face reconstruction methods demonstrate performance improvement regarding 3D face geometry retrieval. However, these methods pose numerous challenges, particularly during testing. One of the significant challenges is the requirement of processed (cropped and aligned) input, which leads to the dependency on the facial landmark coordinates detector. Moreover, input processing time degrades the network's testing speed, thus increasing the test time. Therefore, we propose a ***RE**duced **D**ependency **F**ast* **U***nsuperviSEd 3D Face Reconstruction* (***RED-FUSE***) framework, which exploits unprocessed (uncropped and unaligned) face images to estimate reliable 3D face shape and texture, waiving off the requirement for prior facial landmarks information, and improving the network's estimation speed. More specifically, we utilize a (1) *Multi-pipeline training architecture* to reconstruct accurate 3D faces from challenging (transformed) unprocessed test inputs without posing additional requirements and (2) *Pose transfer module* that ensures reliable training for unprocessed challenging images by attaining the inter-pipeline face pose consistency without requiring the respective facial landmark information. We performed qualitative and quantitative analysis of our model on the unprocessed CelebA-test dataset, LFW-test set, NoW selfie challenge set and various open-source images. Our RED-FUSE outperforms a current method on the unprocessed CelebA-test dataset, e.g., for 3D shape-based, color-based, and 2D perceptual errors, the proposed method shows an improvement of **46**.**2**%, **15**.**1**%, and **27**.**4**%, respectively. Moreover, our approach demonstrates a significant improvement of **29**.**6**% on NoW selfie challenge. Furthermore, RED-FUSE requires lesser test time (a reduction from **7**.**30** m.sec. to **1**.**85** m.sec. per face) and poses minimal test time dependencies, demonstrating the effectiveness of the proposed method.

Keywords Unsupervised training · 3D Morphable Model (3DMM) · 3D face reconstruction · Reduced testing requirements · Improved testing speed

1 Introduction

3D face reconstruction from a monocular face image is a longstanding problem in the field of computer graphics and computer vision, which has numerous applications, such as face recognition [1], interaction in augmented/virtual environments [2], media manipulation, and animation [3]. For recovering 3D face shape and texture from monocular images, a statistical 3D Morphable Model (3DMM) [4] is most popularly utilized, built from hundreds of 3D face scans. 3DMM facilitates a search space spanning the range of 3D human faces. The points from the 3DMM search space contain information on 3D face geometry and texture. Along with these points, face illumination and pose coefficients are required to generate desired 3D faces. The reconstructed 3D faces imitate the face shape and color of the corresponding face images; thus the processing, i.e., *cropping and alignment*, of input face images is needed. Facial image processing poses dependencies such as pre-trained landmark detectors. Moreover, processing requires significant time, which is a major issue, particularly during testing.

Numerous deep learning-based monocular 3D face reconstruction methods [5–7] have been proposed, but

✉ Hitika Tiwari
hitika@iitk.ac.in

Venkatesh K. Subramanian
venkats@iitk.ac.in

Yong-Sheng Chen
yschen@nycu.edu.tw

[1] Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

[2] Department of Electrical Engineering, Indian Institute of Technology Kanpur, Kanpur, India

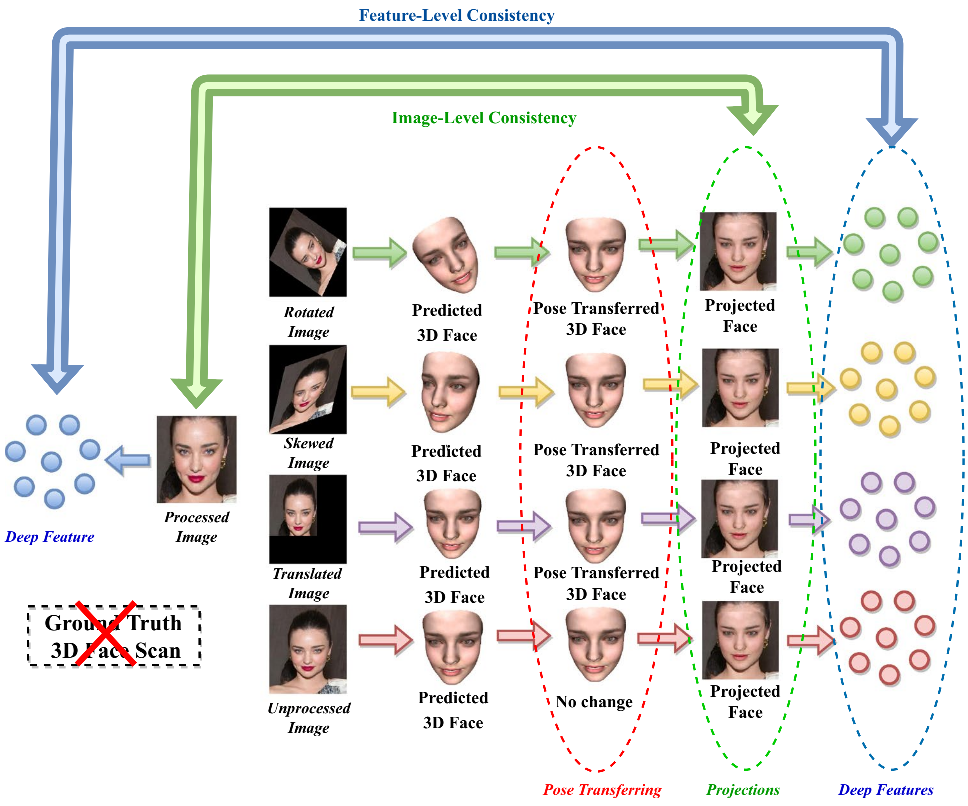


Fig. 1 An overview of our ***RE****duced* ***D****ependency* ***F****ast* **U***nsuperviS****E****d 3D Face Reconstruction* (***RED-FUSE***) framework. The proposed method addresses the problem of unprocessed monocular 3D face reconstruction in an unsupervised manner by exploiting novel pose transferring module and speeds up the testing process, without posing the requirement of 3D ground-truth face scans

dependency reduction and test speed improvement are beyond the scope of these approaches, which are crucial for real-time application. Tewari et al. [5, 8] produce 3D faces in consistency with the processed inputs. Deng et al. [6] exploit dlib [9] and (or) MTCNN [10] to process the input face images for reconstructing accurate 3D faces. Tiwari et al. [7, 11] require processed face images at the input for generating occlusion robust 3D faces. Feng et al. [12] reconstruct detailed 3D faces from monocular processed face images. Although these methods improve the 3D face reconstruction accuracy, they require processed input at the test time, leading to the need for facial landmark information. Moreover, processing reduces the testing speed, thus increasing the time required for testing these methods. Therefore, a novel training pipeline is required to overcome these issues and obtain accurate 3D faces from unprocessed (uncropped and unaligned) monocular images.

In this work, our aim is to estimate the 3D faces from unprocessed monocular face images to reduce the test time dependencies and improve the testing speed. Furthermore, an unsupervised training scheme is needed to overcome the requirement of *difficult to procure* ground truth 3D face scans. To achieve these objectives, we propose a ***RE****duced* ***D****ependency* ***F****ast* **U***nsuperviS****E****d 3D Face Reconstruction* (***RED-FUSE***) framework, which estimates statistical 3D face coefficients for unprocessed face images in an unsupervised manner, as in Fig. 1. More specifically, RED-FUSE exploits a (1) *Multi-pipeline architecture* to ensure a reliable reconstruction of 3D faces from challenging unprocessed inputs and (2) *Pose transfer module* that facilitates the elimination of landmark requirements for training the network on various variants of unprocessed inputs. Due to the inclusion of challenging image variants (affine transformed) as the inputs to the training pipeline and landmark free network's training for unprocessed variants, RED-FUSE produces accurate 3D face from real-world in-the-wild face images.[1] The proposed RED-FUSE is qualitatively and quantitatively evaluated on the numerous open source unprocessed images, CelebA-test dataset [13], LFW-test set [14], and NoW selfie-based validation dataset [15]. Our method demonstrates superior performance over several methods. For example, we obtain an improvement of **46.2**%, **15.1**%, **29.6**% and **27.4**% for 3D shape-based error, color-based error, NoW selfie challenge, and visual similarity (perceptual) error, respectively, compared to a recent approach. Moreover, our test time improves from **7.30** m.sec. to **1.85** m.sec. per face compared to various 3D face reconstruction methods.

A summary of our multi-fold contributions[2] is as follows.

[1] Note that we refer to those images in which face occupies significant area such as selfies.

[2] This paper is an extended version of [16].

1. We propose ***RE**duced **D**ependency **F**ast* **U***nsuperviSEd 3D Face Reconstruction* (***RED-FUSE***) to perform 3D face reconstruction from unprocessed face images without posing additional requirements and dependencies.
2. We propose a pose transfer module, which integrates with our training framework to facilitate landmark-free training of unprocessed variants, thus aiding in eliminating the landmark requirement at the test time.
3. We leverage a multi-pipeline training scheme to learn the statistical representation of 3D faces for unprocessed variants of face images in an unsupervised manner, overcoming the need for difficult-to-procure ground-truth 3D faces.
4. Our method demonstrates improvements for several 2D and 3D evaluation metrics. For example, the proposed approach improves 3D shape accuracy by **46**%+ and 2D visual error by **27**%+, demonstrating the effectiveness of the proposed approach.
5. Our method does not require input processing during testing, thus eliminating the test time landmark dependency and producing reliable 3D faces.
6. The proposed approach provides **75**% faster inference than recent state-of-the-art monocular 3D face reconstruction methods and shows real-time performance.

2 Related work

The literature behind 3D face reconstruction method [17–21] is vast. Therefore, we focus on morphable model-based [4, 22–24] monocular 3D face reconstruction approaches and unsupervised training strategies.

3D Face Reconstruction Methods: 3D face shape retrieval from an unconstrained monocular face image is a mathematically ill-posed problem. A geometric prior is required to address the issue. 3DMM [4] has gained immense popularity in recent years, which serves as a strong prior for reconstructing 3D faces. Tewari et al. [5, 25] exploit 3DMM to reconstruct 3D faces from face images by exploiting a cycle-consistent approach. Sela et al. [26] provide high-quality reconstructions by utilizing depth images and facial correspondence maps. Feng et al. [22] disentangle shape features such that the tasks of reconstructing 3D face shapes and learning discriminative shape features for face recognition are accomplished simultaneously. Tran et al. [27] produce accurate 3D faces from non-frontal, obstructed face images. Genova et al. [28] use synthetic images with corresponding ground-truth data, where label-free instances of real images are exploited to reconstruct 3D faces. Deng et al. [6] attain deep-feature consistency to improve the reconstructed 3D face shape accuracy. Gecer et al. [29] produce high-fidelity 3D face texture and shape by estimating facial texture in UV space. Tu et al. [30] use sparse 2D facial landmark heatmaps to produce high-quality 3D faces. Feng et al. [12] generate a UV displacement map containing person-specific details to reconstruct detailed 3D faces from monocular images. Zeng et al. [31] integrate a fitting-based approach with the shape-from-shading method [32] to reconstruct detailed 3D face geometry. Tiwari et al. [11] distill knowledge for tackling occlusions to reconstruct accurate 3D faces. Tiwari et al. [7] deploy a self-supervision strategy to generate occlusion robust 3D faces. However, these approaches require the processing of face images, which poses a dependency on prior landmark information and degrades the testing speed of the model. Besides, our method reconstructs reliable 3D faces from unprocessed face data without posing additional dependencies, thus demonstrating reduced dependency and faster testing speed.

Unsupervised Learning: Recently, there has been a surge of interest in the unsupervised training scheme for monocular 3D face reconstruction using processed inputs, as it can learn statistical 3D face coefficients without human intervention. The key is to design a 3D face reconstruction task that relates the projected 3D faces with the corresponding processed face images such that the 3D face coefficients can be self-annotated. Most recent developments for 3D face reconstruction tasks [8, 28, 29, 33] utilize the unsupervised approach mentioned above. Tewari et al. [8] establish consistency between processed input and the rendered face to overcome the requirement of external supervision. Genova et al. [28] exploit labeled synthetic data, whereas label-free instances of processed real inputs are used to perform unsupervised 3D face learning. Gecer et al. [29] estimate the relationship between the facial identity features and the parameters of a 3DMM for shape and texture for processed data in an unsupervised manner. Besides, our proposed task exploits unprocessed images as input to learn the accurate 3D face representation without external supervision.

3 Technical details

In this section, we present the preliminaries of 3D face reconstruction (Sect. 3.1). Moreover, we provide the details of proposed **RE**duced **D**ependency **F**ast **U**nsupervi***SE***d 3D Face Reconstruction (**RED-FUSE**), which reconstructs 3D faces from unprocessed face images without requiring external supervision (Sect. 3.2).

3.1 Preliminaries

We present the preliminaries for reconstructing 3D faces from monocular face images. More specifically, we provide the details on the 3D Morphable Model (Sect. 3.1.1), which serves as a prior for facilitating fitting-based monocular 3D face reconstruction. Moreover, we present the illumination assumption (Sect. 3.1.2), and face projection (Sect. 3.1.3).

3.1.1 3D Morphable Model (3DMM)

A 3DMM reconstructs the desired 3D face by exploiting the linear combination of shape ($\boldsymbol{\alpha} \in \mathbb{R}^{80}$), expression ($\boldsymbol{\beta} \in \mathbb{R}^{64}$), and texture ($\boldsymbol{\gamma} \in \mathbb{R}^{80}$) coefficients with their respective basis vectors $\boldsymbol{\mathcal{B}}_\alpha \in \mathbb{R}^{80\times3N}$, $\boldsymbol{\mathcal{B}}_\beta \in \mathbb{R}^{64\times3N}$, and $\boldsymbol{\mathcal{B}}_\gamma \in \mathbb{R}^{80\times3N}$, respectively, as follows.

$$\boldsymbol{\mathcal{S}} = \overline{\boldsymbol{\mathcal{S}}} + \boldsymbol{\alpha}\boldsymbol{\mathcal{B}}_\alpha + \boldsymbol{\beta}\boldsymbol{\mathcal{B}}_\beta, \quad \boldsymbol{\mathcal{T}} = \overline{\boldsymbol{\mathcal{T}}} + \boldsymbol{\gamma}\boldsymbol{\mathcal{B}}_\gamma, \tag{1}$$

where, $\overline{\boldsymbol{\mathcal{S}}} \in \mathbb{R}^{3N}$ and $\overline{\boldsymbol{\mathcal{T}}} \in \mathbb{R}^{3N}$ are the mean 3D face shape and texture, respectively. It is worth noting that $\overline{\boldsymbol{\mathcal{S}}}$, $\overline{\boldsymbol{\mathcal{T}}}$, $\boldsymbol{\mathcal{B}}_\alpha$, and $\boldsymbol{\mathcal{B}}_\gamma$ are obtained from the Basel Face Model (BFM) [34]. BFM produces 3D faces with neutral expressions; thus, the expression basis $\boldsymbol{\mathcal{B}}_\beta$ is extracted from the Facewarehouse model [23]. Besides, our network estimates the face coefficients $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$. Moreover, we exclude the reconstruction of ear and neck regions of 3D faces following [6], leading to $N = 36\text{K}$ face vertices.

3.1.2 Illumination assumption

We illuminate the reconstructed 3D faces (from Eq. (1)) using Spherical Harmonics (SH) under the assumption of a *Lambertian* surface reflectance, following [6]. In particular, we exploit SH basis vector $\boldsymbol{\phi} : \mathbb{R}^3 \rightarrow \mathbb{R}$, i-th vertex normal $\boldsymbol{n}_i \in \mathbb{R}^3$, illumination coefficient $\boldsymbol{\gamma}_x \in \mathbb{R}^3$, and texture $\mathcal{T}_i \in \mathbb{R}^3$ corresponding to i-th vertex $\boldsymbol{v}_i \in \mathbb{R}^3$ to illuminate 3D faces, as follows.

$$\boldsymbol{\Gamma}(\boldsymbol{v}_i, \boldsymbol{n}_i \mid \boldsymbol{\gamma}) = \mathcal{T}_i \cdot \sum_{x=1}^{9} \boldsymbol{\gamma}_x \boldsymbol{\phi}(\boldsymbol{n}_i). \tag{2}$$

In Eq. (2), Γ represents the illumination function for reconstructed 3D faces.

3.1.3 3D face projection

To project the 3D faces onto the screen space, we map each 3D face vertex (containing shape $\boldsymbol{\mathcal{S}}_i$, texture $\boldsymbol{\mathcal{T}}_i$, illumination $\boldsymbol{\Gamma}$ and pose $\boldsymbol{p}$ information such that $i \in \{1, 2, \dots, 3N\}$) to the image plane by assuming a pinhole camera under full perspective projection, as follows.

$$\boldsymbol{I'} = \Upsilon(\boldsymbol{\mathcal{S}}_i, \boldsymbol{\mathcal{T}}_i, \boldsymbol{\Gamma}, \boldsymbol{p}), \tag{3}$$

where $\boldsymbol{p}$ contains $\boldsymbol{R} \in SO(3)$ and $\boldsymbol{t} \in \mathbb{R}^3$. It is worth noting that $\boldsymbol{\mathcal{S}} = [\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_{3N}]$ and $\boldsymbol{\mathcal{T}} = [\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_{3N}]$, where $N = 36\text{K}$. Moreover, Υ is projection function, whereas $\boldsymbol{I'}$ denotes the projected 3D face.

3.2 Reduced dependency fast unsupervised 3D face reconstruction

Despite the recent advancements in monocular 3D face reconstruction methods, there is still a large scope for improvement concerning the test time dependencies. Moreover, the issue of test speed improvement is still under-addressed, which is crucial for real-time applications. One possible way to address the problem is to reconstruct 3D faces from unprocessed data, which eliminates the facial landmark requirement at the test time and improves the estimation speed; thus, we aim to reconstruct accurate 3D faces from unprocessed single-view face images without posing additional dependencies. To achieve our objective, we propose ***RE****duced* ***D****ependency* ***F****ast* **U***nsupervi****SE****d 3D Face Reconstruction* (***RED-FUSE***) framework, which exploits unprocessed face images and their variants to estimate the corresponding 3D face coefficients. More specifically, the proposed network exploits unprocessed (Original) image $\boldsymbol{I_O}$ and it's three variants i.e., *Rotated* $\boldsymbol{I_R}$, *Skewed* $\boldsymbol{I_S}$, *and Translated* $\boldsymbol{I_T}$ as the inputs to multi-pipeline framework, estimates corresponding 3D face coefficients $\boldsymbol{C_R}$, $\boldsymbol{C_S}$, $\boldsymbol{C_T}$, $\boldsymbol{C_O}$, generates corresponding 3D face meshes $\boldsymbol{M_R}$, $\boldsymbol{M_S}$, $\boldsymbol{M_T}$, $\boldsymbol{M_O}$, transfers the pose of $\boldsymbol{M_O}$ to the remaining 3D face meshes, and projects them on the processed face image $\boldsymbol{I_P}$ (obtained after processing $\boldsymbol{I_O}$) to get the 2D images $\boldsymbol{I_{R'}}, \boldsymbol{I_{S'}}, \boldsymbol{I_{T'}}, \boldsymbol{I_{O'}}$, all similar to processed image $\boldsymbol{I_P}$ as shown in Fig. 2. Furthermore, $\boldsymbol{C_R}$, $\boldsymbol{C_S}$, $\boldsymbol{C_T}$, $\boldsymbol{C_O}$ are learned by ensuring the consistency between processed image $\boldsymbol{I_P}$ and projected 3D faces $\boldsymbol{I_{R'}}$, $\boldsymbol{I_{S'}}$, $\boldsymbol{I_{T'}}$, $\boldsymbol{I_{O'}}$. It should be noted that 3DMM coefficient $\boldsymbol{C}_i$ contains shape $\boldsymbol{\alpha}_i \in \mathbb{R}^{80}$, expressions $\boldsymbol{\beta}_i \in \mathbb{R}^{64}$, texture $\boldsymbol{\gamma}_i \in \mathbb{R}^{80}$, illumination $\boldsymbol{\delta}_i \in \mathbb{R}^{27}$, rotation and translation vectors (together known as pose coefficients) $\boldsymbol{R}_i \in \mathbb{R}^3$ and $\boldsymbol{t}_i \in \mathbb{R}^3$ such that $i \in \{\boldsymbol{R}, \boldsymbol{S}, \boldsymbol{T}, \boldsymbol{O}\}$, for generating 3D faces. All the components required to train the proposed framework are given below.

Pose Transfer Module: The conventional approaches ensure cycle consistency of the estimated 3D faces with their processed counterparts. The processing of face images requires facial landmark information. However, deriving facial landmarks becomes tedious and infeasible for tough unprocessed variants. Therefore, these methods fail to solve the problem of unprocessed monocular 3D face reconstruction. To overcome these issues, we exploit a novel pose transfer scheme. For this purpose, we transfer the pose coefficients of the 3D face ($\boldsymbol{M_O}$) obtained from the unprocessed image to the 3D faces ($\boldsymbol{M_R}$, $\boldsymbol{M_S}$, $\boldsymbol{M_T}$) generated from the variants of unprocessed input (as shown in Fig. 3), thus assisting the RED-FUSE to attain consistency of all projected 3D faces ($\boldsymbol{I_{R'}}$, $\boldsymbol{I_{S'}}$, $\boldsymbol{I_{T'}}$, $\boldsymbol{I_{O'}}$) with a single processed facial image ($\boldsymbol{I_P}$) (using Eq. (3)), without posing requirement of landmark information for unprocessed variants, as follows.

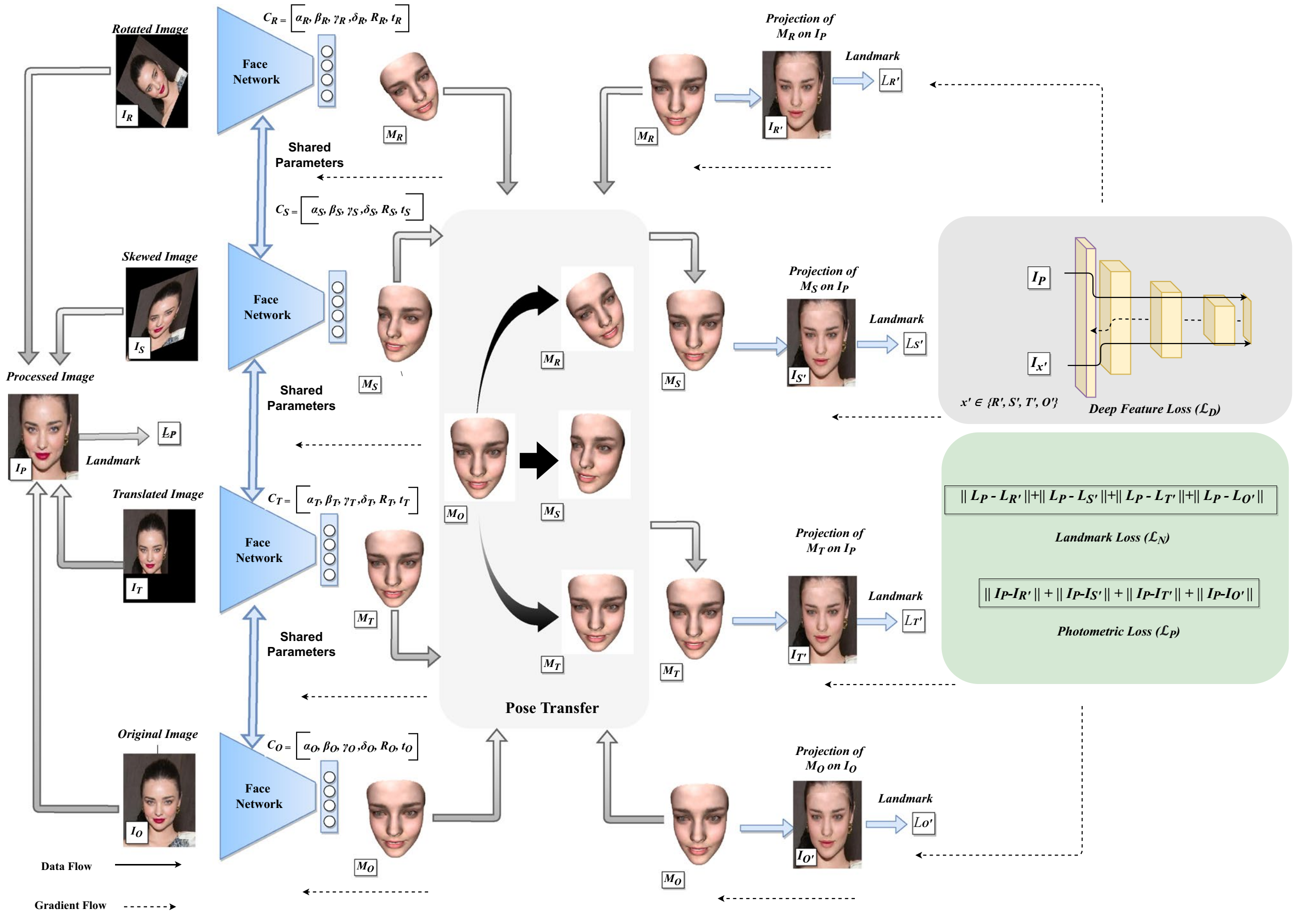

Rotated Image
I_R
Face Network
$C_R = [\alpha_R, \beta_R, \gamma_R, \delta_R, R_R, t_R]$
M_R
Shared Parameters
$C_S = [\alpha_S, \beta_S, \gamma_S, \delta_S, R_S, t_S]$
Skewed Image
I_S
Face Network
M_S
Processed Image
I_P
Landmark
L_P
Shared Parameters
Translated Image
I_T
$C_T = [\alpha_T, \beta_T, \gamma_T, \delta_T, R_T, t_T]$
Face Network
M_T
Shared Parameters
Original Image
I_O
$C_O = [\alpha_O, \beta_O, \gamma_O, \delta_O, R_O, t_O]$
Face Network
M_O
Data Flow
Gradient Flow
Pose Transfer
M_R
M_O
M_S
M_T
Projection of M_R on I_P
Landmark
$L_{R'}$
$I_{R'}$
M_R
Projection of M_S on I_P
Landmark
$L_{S'}$
$I_{S'}$
M_S
I_P
$I_{x'}$
$x' \in \{R', S', T', O'\}$
Deep Feature Loss ($\mathcal{L}_D$)
$|| L_P - L_{R'} || + || L_P - L_{S'} || + || L_P - L_{T'} || + || L_P - L_{O'} ||$
Landmark Loss ($\mathcal{L}_N$)
$|| I_P - I_{R'} || + || I_P - I_{S'} || + || I_P - I_{T'} || + || I_P - I_{O'} ||$
Photometric Loss ($\mathcal{L}_P$)
Projection of M_T on I_P
Landmark
$L_{T'}$
$I_{T'}$
M_T
Projection of M_O on I_O
Landmark
$L_{O'}$
$I_{O'}$
M_O

Fig. 2 An overview of our ***RE**duced **D**ependency **F**ast **U**nsuperviSEd 3D Face Reconstruction* (***RED-FUSE***) framework. The proposed method addresses the problem of unprocessed monocular 3D face reconstruction by exploiting novel pose transferring module in an unsupervised manner and speeds up the testing process, without the requirement of 3D ground-truth face scans

$$\boldsymbol{I_{j'}} = \Upsilon(\boldsymbol{\mathcal{S}}_{ij}, \boldsymbol{\mathcal{T}}_{ij}, \boldsymbol{\Gamma}_j, \boldsymbol{p}_{j=O}), \tag{4}$$

where $\boldsymbol{\mathcal{S}}_{ij}$ and $\boldsymbol{\mathcal{T}}_{ij}$ are the i-th element of shape vector $\boldsymbol{\mathcal{S}}$ and texture vector $\boldsymbol{\mathcal{T}}$ (from Eq. (1)), respectively, such that $j \in \{R, S, T, O\}$. $\boldsymbol{\Gamma}_j$ (from Eq. (2)) is the illumination vector, whereas $\boldsymbol{p}_{j=O}$ represents the pose coefficients of estimated 3D face $\boldsymbol{M_O}$. Also, Υ is the 3D face projection function, which aids in producing projected 3D face $\boldsymbol{I_{j'}}$ on to the processed image. It is worth noting that the module facilitates the network to waive-off facial landmark coordinate requirements during testing, thus reducing test time dependencies and improving estimation speed.

Obtaining 3D Face Alignment: To obtain the accurate pose of estimated 3D faces, we align the projected 3D faces $\boldsymbol{I_{R'}}, \boldsymbol{I_{S'}}, \boldsymbol{I_{T'}}, \boldsymbol{I_{O'}}$ with the corresponding processed face image $\boldsymbol{I_P}$. Therefore, as follows, we obtain the consistency between 68 facial landmark coordinates using *Landmark Loss* $\mathcal{L}_N$.

$$\begin{aligned}\mathcal{L}_N =& || \boldsymbol{L_P} - \boldsymbol{L_{R'}} || + || \boldsymbol{L_P} - \boldsymbol{L_{S'}} || \\ &+ || \boldsymbol{L_P} - \boldsymbol{L_{T'}} || + || \boldsymbol{L_P} - \boldsymbol{L_{O'}} || .\end{aligned} \tag{5}$$

In Eq. (5), $\boldsymbol{L_P}$ is a set of landmark coordinates obtained for $\boldsymbol{I_P}$, whereas $\boldsymbol{L_{R'}}, \boldsymbol{L_{S'}}, \boldsymbol{L_{T'}}, \boldsymbol{L_{O'}}$ are the facial landmark coordinates of $\boldsymbol{I_{R'}}, \boldsymbol{I_{S'}}, \boldsymbol{I_{T'}}, \boldsymbol{I_{O'}}$, respectively. Also, $|| \cdot ||$ is the L2 loss.

Obtaining Photometric Consistency: To learn the 3D face color, we regress the pixels of projected 3D faces $\boldsymbol{I_{R'}}, \boldsymbol{I_{S'}}, \boldsymbol{I_{T'}}, \boldsymbol{I_{O'}}$ on to the corresponding processed face image $\boldsymbol{I_P}$, thus attaining the pixel-consistency using *Photometric Loss* $\mathcal{L}_P$, as follows.

$$\begin{aligned}\mathcal{L}_P =& \frac{\boldsymbol{\mathcal{A}} \cdot || \boldsymbol{I_P} - \boldsymbol{I_{R'}} ||}{|| \boldsymbol{\mathcal{A}} ||} + \frac{\boldsymbol{\mathcal{A}} \cdot || \boldsymbol{I_P} - \boldsymbol{I_{S'}} ||}{|| \boldsymbol{\mathcal{A}} ||} \\ &+ \frac{\boldsymbol{\mathcal{A}} \cdot || \boldsymbol{I_P} - \boldsymbol{I_{T'}} ||}{|| \boldsymbol{\mathcal{A}} ||} + \frac{\boldsymbol{\mathcal{A}} \cdot || \boldsymbol{I_P} - \boldsymbol{I_{O'}} ||}{|| \boldsymbol{\mathcal{A}} ||},\end{aligned} \tag{6}$$

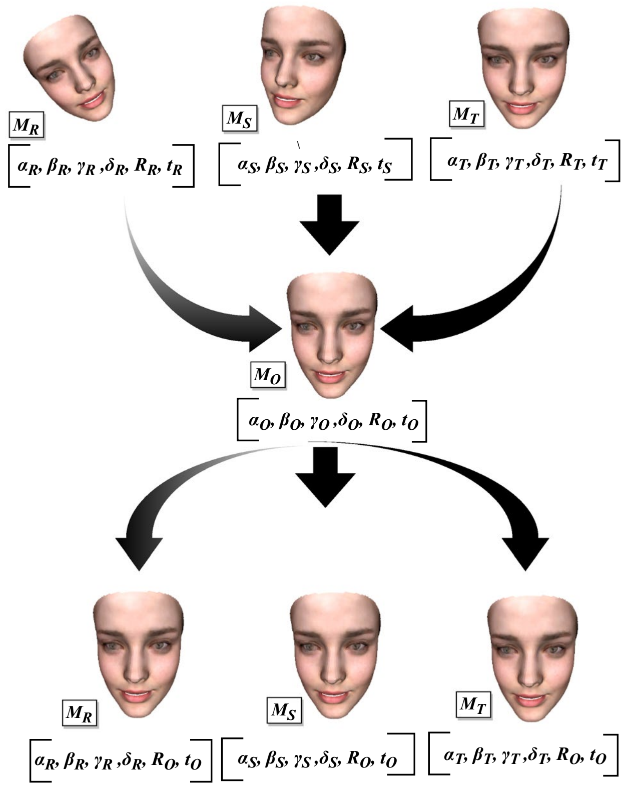

M_R
M_S
M_T
$[\alpha_R, \beta_R, \gamma_R, \delta_R, R_R, t_R]$
$[\alpha_S, \beta_S, \gamma_S, \delta_S, R_S, t_S]$
$[\alpha_T, \beta_T, \gamma_T, \delta_T, R_T, t_T]$
M_O
$[\alpha_O, \beta_O, \gamma_O, \delta_O, R_O, t_O]$
M_R
M_S
M_T
$[\alpha_R, \beta_R, \gamma_R, \delta_R, R_O, t_O]$
$[\alpha_S, \beta_S, \gamma_S, \delta_S, R_O, t_O]$
$[\alpha_T, \beta_T, \gamma_T, \delta_T, R_O, t_O]$

Fig. 3 A demonstration of the proposed pose transfer module. It is worth noting that apart from rotation and translation coefficients, we do not transfer other 3D face coefficients

where $\mathcal{A}$ represents the skin attention mask [6] obtained for $\boldsymbol{I_P}$. $\cdot$ denotes the element-wise multiplication.

Obtaining Deep Feature Similarity: To ensure the visual similarity between the processed image $\boldsymbol{I_P}$ and the projected 3D faces $\boldsymbol{I_{R'}}$, $\boldsymbol{I_{S'}}$, $\boldsymbol{I_{T'}}$, $\boldsymbol{I_{O'}}$, we use *Deep Feature Loss* $\mathcal{L}_D$, as follows.

$$\mathcal{L}_D = 4 - \left(\frac{< \zeta_P, \zeta_{R'} >}{|| \zeta_P |||| \zeta_{R'} ||} + \frac{< \zeta_P, \zeta_{S'} >}{|| \zeta_P |||| \zeta_{S'} ||} + \frac{< \zeta_P, \zeta_{T'} >}{|| \zeta_P |||| \zeta_{T'} ||} + \frac{< \zeta_P, \zeta_{O'} >}{|| \zeta_P |||| \zeta_{O'} ||} \right), \tag{7}$$

where $\boldsymbol{\zeta_P}$ is the deep feature for $\boldsymbol{I_P}$, whereas $\zeta_{R'}$, $\zeta_{S'}$, $\zeta_{T'}$, $\zeta_{O'}$ represent the deep-feature vectors of $\boldsymbol{I_{R'}}$, $\boldsymbol{I_{S'}}$, $\boldsymbol{I_{T'}}$, $\boldsymbol{I_{O'}}$, respectively. It should be noted that the deep features are obtained using pre-trained face recognition model FaceNet [35].

Regularization: For ensuring the plausibility of reconstructed 3D face shape, expressions and texture, we enforce the estimated shape ($\boldsymbol{\alpha}_R$, $\boldsymbol{\alpha}_S$, $\boldsymbol{\alpha}_T$, $\boldsymbol{\alpha}_O$), expression ($\boldsymbol{\beta}_R$, $\boldsymbol{\beta}_S$, $\boldsymbol{\beta}_T$, $\boldsymbol{\beta}_O$) and texture ($\boldsymbol{\gamma}_R, \boldsymbol{\gamma}_S, \boldsymbol{\gamma}_T, \boldsymbol{\gamma}_O$) coefficients to follow the BFM distribution (normal), using *Regularization* term, as follows.

$$\begin{aligned}\mathcal{L}_R = {} & w_{\alpha_R} \parallel \boldsymbol{\alpha}_R \parallel + w_{\beta_R} \parallel \boldsymbol{\beta}_R \parallel + w_{\gamma_R} \parallel \boldsymbol{\gamma}_R \parallel \\ & + w_{\alpha_S} \parallel \boldsymbol{\alpha}_S \parallel + w_{\beta_S} \parallel \boldsymbol{\beta}_S \parallel + w_{\gamma_S} \parallel \boldsymbol{\gamma}_S \parallel \\ & + w_{\alpha_T} \parallel \boldsymbol{\alpha}_T \parallel + w_{\beta_T} \parallel \boldsymbol{\beta}_T \parallel + w_{\gamma_T} \parallel \boldsymbol{\gamma}_T \parallel \\ & + w_{\alpha_O} \parallel \boldsymbol{\alpha}_O \parallel + w_{\beta_O} \parallel \boldsymbol{\beta}_O \parallel + w_{\gamma_O} \parallel \boldsymbol{\gamma}_O \parallel,\end{aligned} \tag{8}$$

where w_{α_i}, w_{β_i} and w_{γ_i} are the weights associated with $\boldsymbol{\alpha}_i$, $\boldsymbol{\beta}_i$ and $\boldsymbol{\gamma}_i$, respectively such that $i \in \{R, S, T, O\}$.

Obtaining Overall Supervision: The overall supervisory signal for training the *Reduced dependency fast unsupervised 3D face reconstruction* framework is obtained using the pose transferring module (Eq. (4)), landmark loss $\mathcal{L}_N$ (Eq. (5)), photometric loss $\mathcal{L}_P$ (Eq. (6)), deep feature loss $\mathcal{L}_D$ (Eq. (7)), and regularization $\mathcal{L}_R$ (Eq. (8)). The mathematical formulation of overall loss function is as follows.

$$\mathcal{L} = w_N \mathcal{L}_N + w_P \mathcal{L}_P + w_D \mathcal{L}_D + w_R \mathcal{L}_R, \tag{9}$$

where w_N, w_P, w_D, and w_R are the weights associated with $\mathcal{L}_N$, $\mathcal{L}_P$, $\mathcal{L}_D$, and $\mathcal{L}_R$, respectively.

4 Experiments

In this section, we present the details of the training and testing datasets (Sect. 4.1). Also, the evaluation metrics and algorithms are detailed to evaluate the performance of the proposed method (Sect. 4.2). Moreover, we provide the implementation details of our approach (Sect. 4.3).

4.1 Datasets

We gathered various standard face datasets, such as 300W-LP [36], LFW [14], etc., to form a training dataset. To validate the reconstruction accuracy, we use the test data-set of CelebA [37], NoW selfie-based validation dataset [15] and LFW-test set [14].

4.2 Evaluation metrics

For evaluating the reconstruction accuracy of proposed RED-FUSE, we exploit various 3D and 2D evaluation metrics. Furthermore, we demonstrate the test speed improvement of our using the time analysis. The details of the metrics are as follow.

Algorithm 1 3D Color and Shape-based Error

```
Require: Ground-truth 3D Face Dataset: $\boldsymbol{\Upsilon_G} \in \mathbb{R}^n$, Unprocessed CelebA-test
  Faces: $\boldsymbol{\Psi_O} \in \mathbb{R}^n$, RED-FUSE: $\boldsymbol{\tau}$, 3D Face Function: $\boldsymbol{\mu}$, Number of 3D face
  vertices: $m$, Number of Color Components per Vertex: $c$, Number of Spatial
  Components per Vertex: $s$
Ensure: 3D Color-based Error: $\mathcal{L}_C$, 3D Shape-based Error: $\mathcal{L}_S$
  Initialize $\mathbf{E}_{3DC_{i_j}} = \mathbf{E}_{3DS_{i_j}} = 0$;
  while $i \leq n$ do
      $\boldsymbol{F_{3D_G}} \leftarrow \boldsymbol{\Upsilon_G}[i]$;
      $\boldsymbol{I_O} \leftarrow \boldsymbol{\Psi_O}[i]$;
      $\boldsymbol{\alpha_O}, \boldsymbol{\beta_O}, \boldsymbol{\gamma_O}, \boldsymbol{\delta_O}, \boldsymbol{R_O}, \boldsymbol{t_O} \leftarrow \boldsymbol{\tau}(\boldsymbol{I_O})$;
      $\boldsymbol{F_{3D_O}} \leftarrow \boldsymbol{\mu}(\boldsymbol{\alpha_O}, \boldsymbol{\beta_O}, \boldsymbol{\gamma_O}, \boldsymbol{\delta_O}, \boldsymbol{R_O}, \boldsymbol{t_O})$;
      while $j \leq m$ do
          while $k \leq c$ do
              $\mathbf{E}_{3DC_{i_j}} \leftarrow \mathbf{E}_{3DC_{i_j}} + (\boldsymbol{F_{3D_G}}[j][k] - \boldsymbol{F_{3D_O}}[j][k])^2$;
              $k \leftarrow k + 1$;
          end while
          while $t \leq s$ do
              $\mathbf{E}_{3DS_{i_j}} \leftarrow \mathbf{E}_{3DS_{i_j}} + (\boldsymbol{F_{3D_G}}[j][t] - \boldsymbol{F_{3D_O}}[j][t])^2$;
              $t \leftarrow t + 1$;
          end while
          $j \leftarrow j + 1$;
      end while
      $i \leftarrow i + 1$;
  end while
  $\mathcal{M}_{3DS} \leftarrow \frac{1}{n \times m \times s} \sum_{i,j,k} \mathbf{E}_{3DS_{i_{j_k}}}$;
  $\mathcal{S}_{3DS} \leftarrow \frac{1}{n \times m \times s} \sum_{i,j} (\mathbf{E}_{3DS_{i_j}} - \mathcal{M}_S)^2$;
  $\mathcal{L}_S \leftarrow \mathcal{M}_{3DS} \pm \mathcal{S}_{3DS}$
  $\mathcal{M}_C \leftarrow \frac{1}{n \times m \times c} \sum_{i,j,k} \mathbf{E}_{3DC_{i_{j_k}}}$;
  $\mathcal{S}_{3DC} \leftarrow \frac{1}{n \times m \times s} \sum_{i,j} (\mathbf{E}_{3DC_{i_j}} - \mathcal{M}_C)^2$;
  $\mathcal{L}_C \leftarrow \mathcal{M}_{3DC} \pm \mathcal{S}_{3DC}$
```

3D Shape and Color-based Error: The 3D shape and color-based error metrics evaluate the spatial and color differences between the estimated 3D faces and the corresponding ground truth. Specifically, each 3D face contains $N = 36$K vertices; each vertex has an associated spatial location (x, y, z) and color values (r, g, b). The estimated vertex locations and texture values are compared with the ground-truth data using root mean square and standard deviation error metric. The mathematical formulation of the 3D shape-based error $(M_{3DS} \pm S_{3DS})$ is given below.

$$\begin{aligned} M_{3DS} &= \frac{1}{3N} \sum_i \mathbf{E}_{3DS_i}, \\ S_{3DS} &= \frac{1}{3N} \sum_i (\mathbf{E}_{3DS_i} - M_{3DS})^2 \quad \text{where,} \\ \mathbf{E}_{3DS} &= \sqrt{(x_{i^G} - x_{i^P})^2 + (y_{i^G} - y_{i^P})^2 + (z_{i^G} - z_{i^P})^2}, \end{aligned} \tag{10}$$

where M_{3DS} and S_{3DS} are the mean and standard deviation of shape error, respectively. Moreover, k_{i^G} and k_{i^P} are the ground-truth and predicted spatial locations of i-th vertex such that $k \in \{x, y, z\}$. Also, the mathematical formulation of the 3D color-based error $(M_{3DC} \pm S_{3DC})$ is given below.

$$
\begin{aligned}
M_{3DC} &= \frac{1}{3N}\sum_i \mathbf{E}_{3DC_i}, \\
S_{3DC} &= \frac{1}{3N}\sum_i (\mathbf{E}_{3DC_i} - M_{3DC})^2 \quad \text{where,} \\
\mathbf{E}_{3DC} &= \sqrt{(r_{i^G} - r_{i^P})^2 + (g_{i^G} - g_{i^P})^2 + (b_{i^G} - b_{i^P})^2}.
\end{aligned}
\tag{11}
$$

$$
\begin{aligned}
M_{2DP} &= \sum_i || (\mathbf{v}_{i^G} - \mathbf{v}_{i^P}) ||, \\
S_{2DP} &= \frac{1}{M}\sum_i (|| (\mathbf{v}_{i^G} - \mathbf{v}_{i^P}) || - M_{2DP})^2,
\end{aligned}
\tag{12}
$$

Algorithm 2 Perceptual Error

```
Require: Unprocessed CelebA-test Dataset: $\boldsymbol{\Psi_O} \in \mathbb{R}^n$, Processed CelebA-test
  Dataset: $\boldsymbol{\Psi_P} \in \mathbb{R}^n$, Projection Function: $\boldsymbol{\zeta}$, Face Recognition Model: $\Xi$, RED-
  FUSE: $\boldsymbol{\tau}$
Ensure: Perceptual Error: $\mathcal{L}_{PE}$
  while $i \leq n$ do
    $\boldsymbol{I_P} \leftarrow \boldsymbol{\Psi_P}[i]$;
    $\boldsymbol{I_O} \leftarrow \boldsymbol{\Psi_O}[i]$;
    $\boldsymbol{\alpha_O}, \boldsymbol{\beta_O}, \boldsymbol{\gamma_O}, \boldsymbol{\delta_O}, \boldsymbol{R_O}, \boldsymbol{t_O} \leftarrow \boldsymbol{\tau}(\boldsymbol{I_O})$;
    $\boldsymbol{I_{O'}} \leftarrow \boldsymbol{\zeta}(\boldsymbol{\alpha_O}, \boldsymbol{\beta_O}, \boldsymbol{\gamma_O}, \boldsymbol{\delta_O}, \boldsymbol{R_O}, \boldsymbol{t_O})$;
    $\boldsymbol{v_G} \leftarrow \Xi(\boldsymbol{I_{O'}}), \boldsymbol{v_P} \leftarrow \Xi(\boldsymbol{I_P})$;
    $\mathcal{L}_{P_i} \leftarrow \|\boldsymbol{v_P} - \boldsymbol{v_G}\|$;
    $i \leftarrow i + 1$;
  end while
  $M_{2DP} \leftarrow \frac{1}{n}\sum_i \mathcal{L}_{P_i}$;
  $S_{2DP} \leftarrow \frac{1}{n}\sum_i (\mathcal{L}_{P_i} - M_{2DP})^2$;
  $\mathcal{L}_{PE} \leftarrow M_{2DP} \pm S_{2DP}$;
```

M_{3DC} and S_{3DC} are the mean and standard deviation of the color error, respectively. Furthermore, k_{i^G} and k_{i^P} are the ground-truth and predicted color values associated with i-th face vertex such that $k \in \{r, g, b\}$ where r denotes red, g represents green and b is blue color values. We exploit a total of 80 subjects for the comparison. An algorithm for 3D color and shape-based error evaluation is given in Algo. 1.

NoW Challenge: NoW selfie challenge [15] computes the scan-to-mesh distance between the ground truth scan and the estimated 3D faces on the selfie images. Our method produces 3D faces from unprocessed images such as selfies and near-face pictures; thus, the evaluation is crucial for demonstrating the 3D shape accuracy of the proposed approach.

Perceptual Error: In addition to 3D evaluation, we also evaluate the performance of our model on the 2D perceptual metric using 3K, and 1.5K images of the CelebA-test, and LFW-test datasets, respectively. The metric emphasizes the visual similarity between the 2D face image and the rendered counterpart. Therefore, the metric is crucial for evaluating the visual consistency between the input data and the estimated faces. To perform the evaluation, we leverage, seven high performing face recognition models VGG-Face [38], FaceNet [35], FaceNet-512 [35], OpenFace [39], DeepFace [40], ArcFace [41] and SFace [42], as follows.

where $\mathbf{v}_{i^G} \in \mathbb{R}^M$ and $\mathbf{v}_{i^P} \in \mathbb{R}^M$ are the ground truth and predicted vectors for i-th face image, respectively. $|| \cdot ||$ denotes L2 norm. Moreover, M_{2DP} and S_{2DP} are the mean and standard deviation of perceptual error vectors, respectively. Please refer to Algo. 2 for details.

Test-time Analysis: Finally, we evaluate the improvement in the testing time by deriving its average percentage decrease compared to SOTA methods. For the comparison, we tested the models on 3K images from the CelebA-test dataset and derived the average time taken by each network.

Note that the training and testing datasets are distinct, and the testing data is not accessible during training.

4.3 Implementation details

Our ***RE****duced* ***D****ependency* ***F****ast* **U***nsuperviSEd 3D Face Reconstruction* (***RED-FUSE***) framework contains 3D face prediction networks, which estimate 3D face vector $\boldsymbol{C}_i \in \mathbb{R}^{257}$, containing shape $\boldsymbol{\alpha}_i \in \mathbb{R}^{80}$, expression $\boldsymbol{\beta}_i \in \mathbb{R}^{64}$, texture $\boldsymbol{\gamma}_i \in \mathbb{R}^{80}$, illumination $\boldsymbol{\delta}_i \in \mathbb{R}^{27}$, rotation $\boldsymbol{R}_i \in \mathbb{R}^3$, and translation $\boldsymbol{t}_i \in \mathbb{R}^3$ coefficients such that $i \in \{R, S, T, O\}$. Therefore, the last fully-connected (FC) layer of our

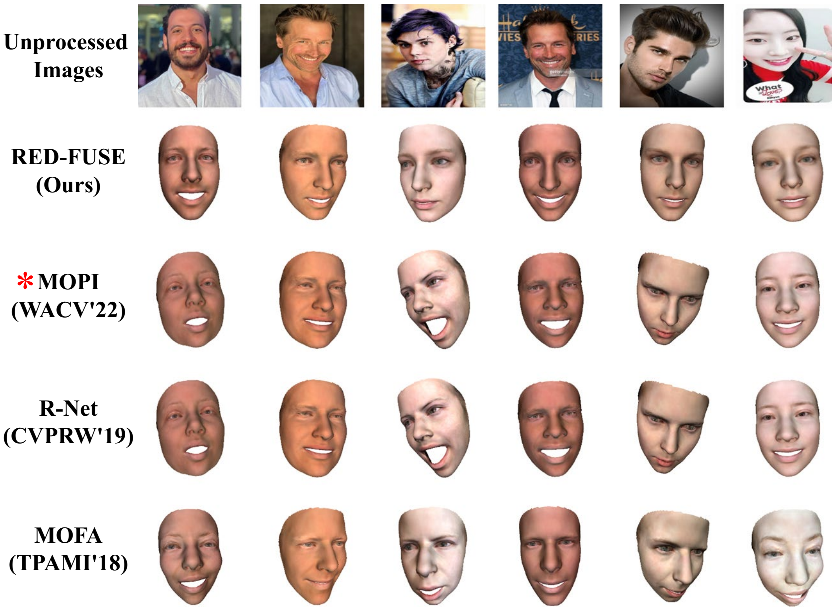


Fig. 4 A comparison of qualitative performance of the proposed **RED-FUSE** model with R-Net and MoFA methods on *open source images*. Results show the superior 3D face reconstruction using the proposed approach. *MOPI distills the knowledge from R-Net for occlusion robustness, resulting in the same performance

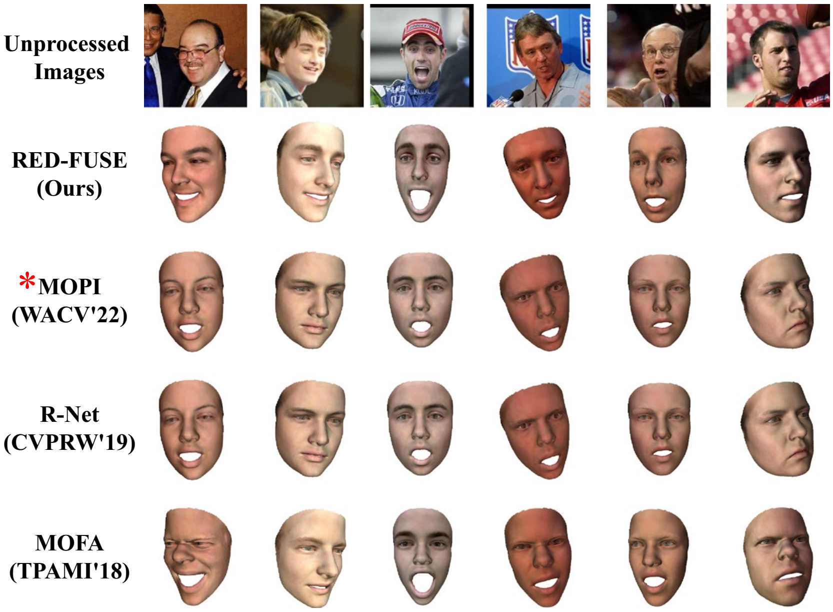


Fig. 5 A comparison of qualitative performance of the proposed **RED-FUSE** model with R-Net and MoFA methods on *LFW datasets*. *MOPI distills the knowledge from R-Net for occlusion robustness, resulting in the same performance

backbone architecture contains 257 nodes. Following [6], we exploit ResNet-50 as our backbone architecture such that the classification layer is replaced by 257 nodal FC layer. Moreover, the in-the-wild (unprocessed) face images and their variants (*rotated, skewed, and translated*) of size 224 × 224 serve as the inputs to our framework. Furthermore, the unprocessed face images (not their variants) are cropped, aligned (using the method in [43]), and reshaped to size 224 × 224, which facilitate the unsupervised training. Besides, we opt for a batch of 5 for each case: rotated, skewed, and translated original unprocessed face images. Thus, the proposed framework is trained with a net batch size of 20. Our framework is initialized with ImageNet weights [44]. In addition, an Adam optimizer [45] is utilized with an initial learning rate of 10^{-4}, and 500K training iterations. The proposed framework contains the weights associated with the losses as $w_N = 1.6 \times 10^{-3}$, $w_P = 1.92, w_D = 0.2$, $w_R = 3 \times 10^{-4}$, following R-Net [6].

4.4 Results

In this section, we compare the qualitative (Sect. 4.4.1) and quantitative results (Sect. 4.4.2) of our method with various methods, MoFA [5, 8], R-Net [6], and MOPI [11] on the several open source images, test dataset of CelebA [37], LFW-test set [14] and NoW selfie dataset [15]. MoFA is

Table 1 A quantitative comparison of the perceptual error with other approaches on *CelebA-test dataset*, where the error numbers are the lower the better

	Methods			
Backbones	MoFA (**TPAMI'18**)	R-Net (**CVPRW'19**)	MOPI (**WACV'22**)	**RED-FUSE (Ours)**
VGG-Face [38]	1.00 ± 0.130	0.936 ± 0.130	0.934 ± 0.130	**0.731 ± 0.191**
FaceNet [35]	1.296 ± 0.134	1.201 ± 0.141	1.197 ± 0.139	**0.801 ± 0.239**
FaceNet-512 [35]	1.329 ± 0.123	1.210 ± 0.135	1.205 ± 0.134	**0.789 ± 0.211**
OpenFace [39]	0.953 ± 0.181	0.885 ± 0.194	0.881 ± 0.193	**0.659 ± 0.243**
DeepFace [40]	0.785 ± 0.164	0.781 ± 0.149	0.781 ± 0.149	**0.646 ± 0.228**
ArcFace [41]	1.315 ± 0.171	1.259 ± 0.153	1.254 ± 0.150	**0.987 ± 0.223**
SFace [42]	1.260 ± 0.110	1.230 ± 0.118	1.227 ± 0.116	**1.036 ± 0.203**

Table 2 A quantitative comparison of the perceptual error with other approaches on *LFW-test set*, where the error numbers are the lower the better

	Methods			
Backbones	MoFA (**TPAMI'18**)	R-Net (**CVPRW'19**)	MOPI (**WACV'22**)	**RED-FUSE (Ours)**
VGG-Face [38]	0.909 ± 0.140	0.915 ± 0.120	0.908 ± 0.119	**0.664 ± 0.180**
FaceNet [35]	1.284 ± 0.210	1.254 ± 0.153	1.206 ± 0.128	**0.883 ± 0.250**
FaceNet-512 [35]	1.294 ± 0.184	1.234 ± 0.156	1.201 ± 0.121	**0.835 ± 0.211**
OpenFace [39]	0.932 ± 0.219	0.963 ± 0.215	0.960 ± 0.213	**0.827 ± 0.240**
DeepFace [40]	0.740 ± 0.159	0.788 ± 0.147	0.782 ± 0.147	**0.668 ± 0.192**
ArcFace [41]	1.314 ± 0.219	1.308 ± 0.136	1.298 ± 0.132	**1.040 ± 0.198**
SFace [42]	1.231 ± 0.161	1.267 ± 0.111	1.250 ± 0.107	**1.067 ± 0.169**

a preliminary CNN-based 3D face reconstruction method, whereas R-Net and MOPI generate the accurate 3D face from single-view face images using the CNN framework, and thus we choose these methods for the comparisons.

4.4.1 Qualitative results

With a single monocular unprocessed face image, RED-FUSE reconstructs 3D face shape and texture without posing additional dependencies. The second rows of Figs. 4 and 5 show that the proposed approach attains high visual similarity between 3D faces and the corresponding unprocessed face images.

Figure 4 qualitatively compares RED-FUSE results with the recent methods, namely MoFA [5], R-Net [6] and MOPI [11] on open source unprocessed images (such as YouTube, Google, etc.). Compared to these methods, RED-FUSE reconstructs superior overall 3D face shape (row 2, 3, 4 and 5) and estimates reliable 3D face pose (column 3). In addition, RED-FUSE predicts better 3D face expressions than all the other approaches. More specifically, MoFA either drags the search outside the 3DMM space (column 3 and 6, row 5) or maps to a coordinate distant from the true coordinate in the search space, resulting in unreliable reconstruction results (column 1, 2, 4 and 5, row 5). Moreover, R-Net fails to capture accurate expressions from unprocessed face images, resulting in poor 3D face shape accuracy (column 1 and 3, row 4). Similar to R-Net, MOPI produces inaccurate face shapes and poses from unprocessed inputs (row 3). It is worth noting that all these methods are producing 3D faces with $N = 36$K face vertices facilitating a fair comparison.

Table 3 A quantitative evaluation on the NoW validation selfie dataset. Our results show superior performance compared to recent methods

Methods	Median (↓)	Mean (↓)	Std. (↓)
MoFA (**TPAMI'18**)	1.99	2.54	2.32
R-Net (**CVPRW'19**)	1.81	2.41	2.41
MOPI (**WACV'22**)	1.81	2.41	2.41
RED-FUSE (Ours)	**1.40**	**2.02**	**2.40**

Figure 5 demonstrates the performance of our method on LFW [14] unprocessed images. The second row shows variations in the expressions and poses of 3D faces emphasizing the ability of RED-FUSE to re-produce difficult-to-produce facial expressions on 3D faces (column 3 and 5, row 2). Also, our model holds the ability to capture a range of accurate 3D face shapes from unprocessed images. It is worth

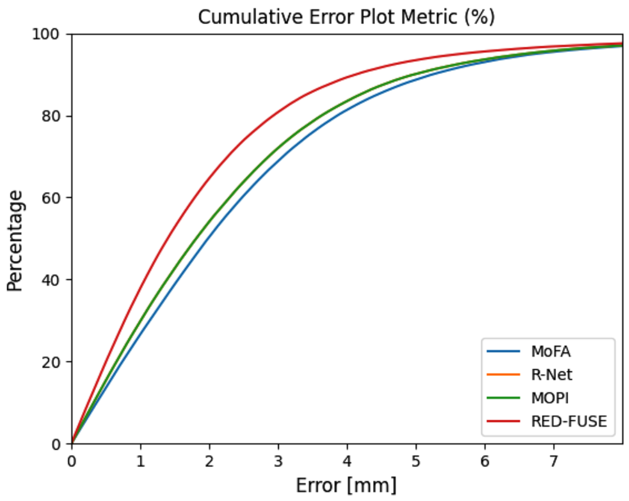


Fig. 6 A cumulative error plot obtained for the NoW validation selfie dataset. In the plot, the x-axis shows the scan-to-mesh distance error (in mm), whereas the y-axis displays the cumulative percentage such that the higher the curve, the better the shape-based accuracy. It is worth noting that the error curves for R-Net (orange) and MOPI (green) are overlapping

Table 4 A comparison of our method with methods MoFA, R-Net and MOPI, the principle of the lesser the better principle

Methods	Shape-based Error (↓)	Color-based Error (↓)
MoFA (**TPAMI'18**)	8.78 ± 0.23	4.23 ± 0.17
R-Net (**CVPRW'19**)	5.84 ± 0.16	3.50 ± 0.14
MOPI (**WACV'22**)	5.82 ± 0.16	3.50 ± 0.14
RED-FUSE (Ours)	**3.14 ± 0.11**	**2.97 ± 0.09**

noting that RED-FUSE reliably predicts eyebrow patterns, gaze details, etc., resulting in the high perceptual similarity between the unprocessed input and the resultant 3D face. Finally, our approach effectively tackles minor occlusions such as caps and spectacles (column 1 and 3, row 2). MoFA aims to attain cycle consistency with the processed input images (row 5), thus resulting in poor visual appearance and sometimes may lead to not-a-human looking face (column 1, row 5). R-Net exploits deep-feature loss to improve the accuracy of 3D faces using cropped and aligned face images in the input, thus producing better results than MoFA but still estimates unreliable 3D face shape and expression for unprocessed face images (row 4). Furthermore, MOPI distills the knowledge from R-Net, showing similar performance as R-Net (row 3). Besides, RED-FUSE exploits the unprocessed images and their variants to estimate 3D faces using a novel pose transfer module and regress the projection of predicted 3D faces over the corresponding processed variant of unprocessed face images for obtaining accurate 3D faces. Therefore, our approach shows significant improvement in performance as compared to other recent methods.

In summary, RED-FUSE generates better reconstruction results, outperforming recent 3D face reconstruction approaches in terms of shape robustness, while producing reliable 3D face expression and pose. Moreover, the proposed method effectively tackles minor occlusions and generates occlusion robust 3D faces.

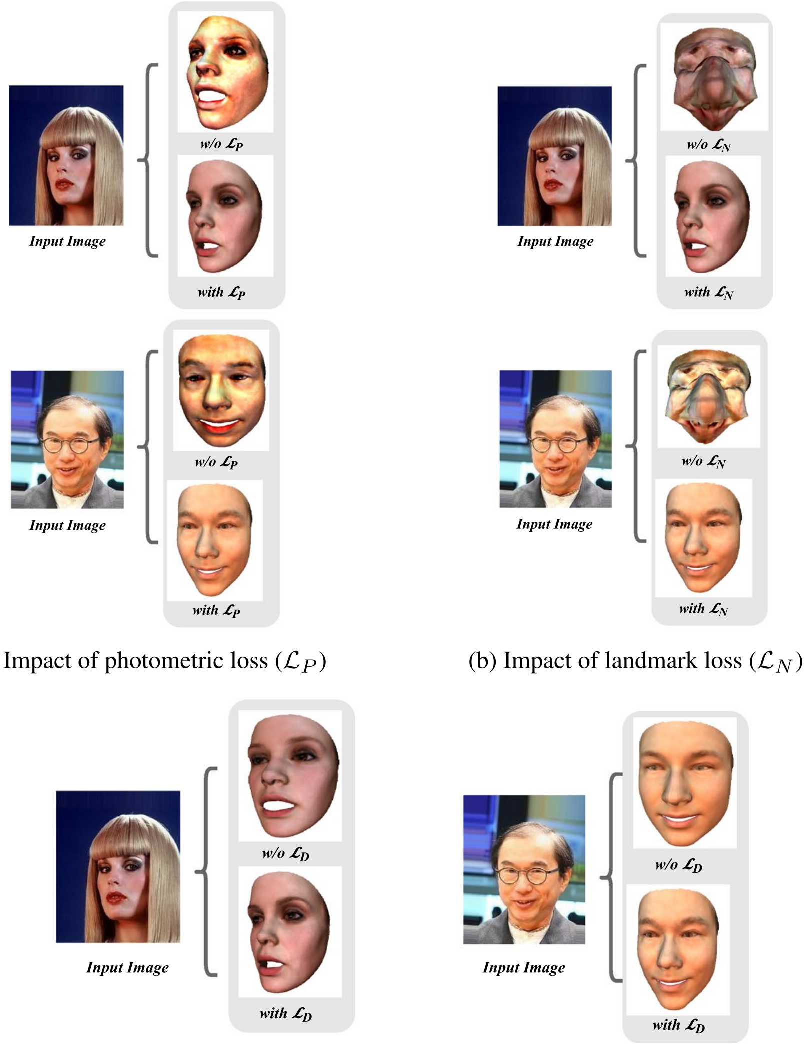
4.4.2 Quantitative results

We compare the quantitative performance of our RED-FUSE framework with methods MoFA [8], R-Net [6], and MOPI [11] on four criteria: (1) Perceptual Error, (2) NoW Selfie Challenge, (3) 3D Shape-based and Color-based Errors, and (4) Required Testing Time and Dependencies, as follows.

(1) Perceptual Error: To emphasize the visual effectiveness of the results obtained using our method over other recent approaches, we compare the perceptual error between rendered 3D faces and color 2D face images with MoFA, R-Net, and MOPI. Our results in Tables 1 and 2 demonstrate superior performance compared to recent approaches. A significant improvement of **27.4**% (1.007 → 0.731), **38.2**% (1.296 → 0.801), **40.6**%(1.329 → 0.789), **30.9**%(0.953 → 0.659), **17.7**% (0.785 → 0.646), **24.9**% (1.315 → 0.987), and **17.8**% (1.260 → 1.036) in the perceptual error for VGG-Face, FaceNet, FaceNet-512, OpenFace, DeepFace, ArcFace, and SFace, respectively, is achieved compared to MoFA on the CelebA-test dataset. Similarly, our approach obtains superior performance on the LFW-test set (Table 2) for various methods.

Table 5 A comparison of our method with recent methods MoFA, R-Net and MOPI. It is worth noting that the proposed method poses fewer dependencies and significantly reduces testing time. Moreover, we re-trained MoFA with the same backbone architecture (as ours) to facilitate a fair comparison. Furthermore, FC stands for the last fully-connected layer

Methods	Backbone Exploited	Time Required in m.sec. (↓)	Facial Landmarks Required
MoFA (**TPAMI'18**)	ResNet-50 (FC: 257 nodes)	7.30	✓
R-Net (**CVPRW'19**)	ResNet-50 (FC: 257 nodes)	7.30	✓
MOPI (**WACV'22**)	ResNet-50 (FC: 257 nodes)	7.30	✓
RED-FUSE (Ours)	ResNet-50 (FC: 257 nodes)	**1.85**	×



(a) Impact of photometric loss ($\mathcal{L}_P$)

(b) Impact of landmark loss ($\mathcal{L}_N$)

(c) Impact of perceptual loss ($\mathcal{L}_D$)

Fig. 7 An analysis of the impact of various losses on the training. Our results show that the model drifts the search outside 3DMM space when trained without landmark loss $\mathcal{L}_N$. Besides, the network trained without photometric loss $\mathcal{L}_P$ or deep-feature loss $\mathcal{L}_D$ demonstrates poor visual similarity with the input image

All these results demonstrate that the outputs of the proposed approach are visually more similar to the faces in the unprocessed images, thus establishing the effectiveness of the proposed method.

2) NoW Selfie Challenge: We evaluate our dataset on the standard NoW validation selfie challenge [15]. Our results in Table 3 show that the proposed method outperforms recent methods by a large margin. For example, improvement of **29.6**% (1.99 → 1.40) and **20.5**% (2.54 → 2.02) are obtained in the median and mean, respectively, compared to a monocular 3D face reconstruction method. Moreover, we show the improvement through a cumulative error plot in Fig. 6. In the plot, the curve corresponding to the proposed RED-FUSE is higher than the curves of other approaches, thus validating our method's supremacy. It is worth noting that the evaluation is performed on unprocessed images, i.e., no landmark information is exploited to estimate the meshes.

3) 3D Shape and Color-based Error: Table 4 shows the 3D shape and color-based error comparison of RED-FUSE with regards to MoFA and R-Net. We infer that RED-FUSE improves the shape and color-based RMSE

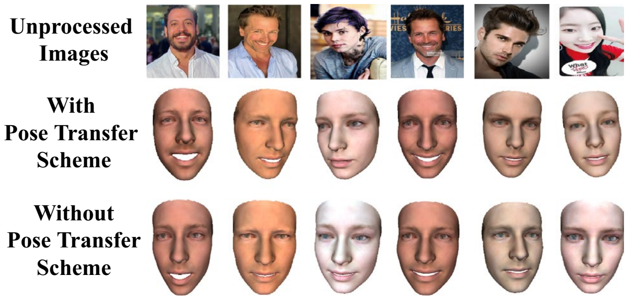
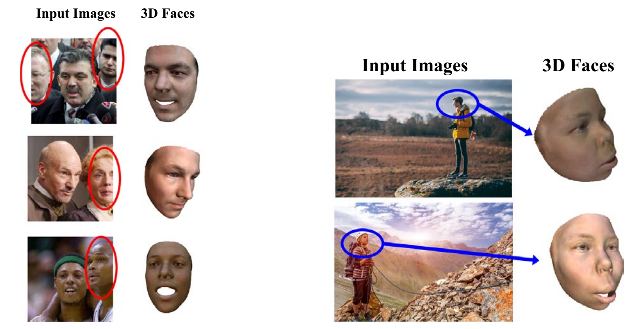


Fig. 8 A qualitative demonstration of the effectiveness of pose transfer module for training the proposed framework

Table 6 A study on the impact of pose transfer module in training the proposed RED-FUSE framework. It is worth noting that the best performance is obtained by exploiting all the components (rotation and translation transfer) of the proposed module

Pose Transfer		Errors	
Rotation	Translation	Shape-based Error (↓)	Color-based.Error (↓)
×	×	4.24 ± 0.15	3.38 ± 0.13
✓	×	3.98 ± 0.14	3.23 ± 0.13
×	✓	3.26 ± 0.12	3.07 ± 0.09
✓	✓	**3.14 ± 0.11**	**2.97 ± 0.09**



(a) Single face reconstruction (b) Undetected far-away faces

Fig. 9 An analysis of the limitations of the proposed model. (Left) The proposed model does not reconstruct the faces in red rings as the network has an upper limit of processing a single face per image. (Right) The faces in the blue ring are not reliably reconstructed, emphasizing the constraint on the area occupied by a face in the captured image

errors by a large margin of **64.2**% (8.78 → 3.14) and **29.8**% (4.23 → 2.97), respectively, compared to MoFA. Also, our method shows a significant improvement of **46.2**% (5.84 → 3.14) and **15.1**% (3.50 → 2.97) for shape and color-based errors, respectively, with respect to R-Net. Furthermore, the improvement of **46.0**% (5.82 → 3.14) and **15.1**% (3.50 → 2.97) is obtained for shape and color-based errors, respectively, compared to MOPI (Table 5).

4) Improved Inference Time: To emphasize the efficacy of the proposed method for real-time applications, we compare our test time with the above-mentioned recent methods MoFA R-Net, and MOPI. These methods require the same test time due to the requirement of processing the raw data during testing. The proposed approach takes **1.85 msec** to generate a 3D face, whereas the above mentioned methods require **7.30 msec** per face, on average, when tested on a Linux platform (Ubuntu 16.04.7) with NVIDIA GK110GL GPU 3D controller card. Therefore, our method improves test time by a large margin of **74.6**% (nearly 4 times faster) compared to the recent approaches. Moreover, unlike various methods, RED-FUSE doesn't require 5 facial landmarks coordinates during the testing, thus eliminating all the test time dependencies.

4.5 Ablation analysis

We present a study on the impact of various losses exploited for the training (Sect. 4.5.1). Moreover, we provide an analysis (qualitative and quantitative) to validate the efficacy of the proposed pose-transferring module for training our model (Sect. 4.5.2).

4.5.1 Impact of losses

We exploit a combination of losses for learning the 3D face representation from unprocessed monocular images in an unsupervised manner. Therefore, we qualitatively demonstrate the effectiveness of each loss in the proposed framework (Fig. 7). In Fig. 7a, the model trained without photometric loss $\mathcal{L}_P$ produces unreliable 3D face texture, i.e., the estimated skin color of the rendered face is inconsistent with the input image. Moreover, the network trained without landmark loss $\mathcal{L}_N$ (Fig. 7b) drags the search outside the 3DMM, thus resulting in a *not-a-human-looking* face. Furthermore, the model trained without deep-feature loss $\mathcal{L}_D$ produces visually less effective 3D faces (Fig. 7c). However, a network trained with all the losses ($\mathcal{L}_P$, $\mathcal{L}_N$, and $\mathcal{L}_D$) demonstrates the best performance, establishing the efficacy of the proposed framework.

4.5.2 Impact of pose transfer module

A critical question arises: *What is the impact of the pose transfer module on the training?* To answer this, we train the model without exploiting the proposed scheme and regress the projection of estimated 3D faces (obtained from unprocessed image and its variants) over the corresponding aligned and cropped face image. Figure 8 shows that the performance of our model degrades when trained without the proposed module, particularly in terms of 3D face shape and expressions. We conjecture that the model trained without our scheme is penalized for estimating poses consistent with the unprocessed image variants, impacting the 3D face shapes and expressions. Therefore, during testing, the model fails to capture accurate 3D face shapes and expressions from unprocessed face images. Besides, the model trained with the pose transfer scheme transfers the estimated 3D face pose of the original unprocessed image to the 3D faces of corresponding variants before penalizing pixel discrepancies. Therefore, the model trained with the pose transfer module learns the correct 3D face shape and expression information from unprocessed face images.

Moreover, we demonstrate the impact of each component of the pose transfer module. Our results in Table 6 show that the model trained without rotation and translation transferring performs poorly on 3D -based errors. However, the accuracy improves by transferring the rotation coefficients ($\boldsymbol{R_O}$) of the 3D mesh ($\boldsymbol{M_O}$) obtained from the unprocessed image to its variants.

Further improvement is observed in transferring the translation coefficient ($\boldsymbol{t_O}$) of $\boldsymbol{M_O}$ obtained from the unprocessed image to its variants. This emphasizes that the translation coefficient is crucial in improving the accuracy of 3D faces. Finally, the model trained with both coefficient transfer, translation, and rotation, demonstrates the best performance, thus, validating the effectiveness of the proposed pose transfer module.

5 Limitations and future work

While RED-FUSE achieves SOTA results for the reconstructed 3D faces (obtained from unprocessed monocular images) and the testing speed, numerous challenges exist. First, RED-FUSE reconstructs only one 3D face irrespective of the number of persons in the image (Fig. 9a). This leads to the need for a more robust network, which divides the images into patches and reconstructs 3D faces based on the face information obtained from each patch. More specifically, the patch size should be small enough to contain a single face only. However, such a network increases computational complexity and poses several dependencies during training, such as prior knowledge of the number of faces in the image. Moreover, RED-FUSE poses challenges in estimating 3D faces from images containing *far-away faces* (Fig. 9b). This suggests the need for more diverse unprocessed training data. Therefore, a face dataset is required for training, consisting of far-away faces, such that the corresponding processed 2D face images do not lose facial information. Finally, details such as makeup, mustaches, etc. (Fig. 9a, row 1) are not reproduced as we exploit BFM, leading to a visual discrepancy between the input image and the corresponding 3D face. BFM spans the range of human facial appearance, thus posing a challenge in reproducing facial accessories such as makeup. Also, BFM contains Principal Component Analysis (PCA) basis vectors (obtained by projecting 3D facial data from high-dimensional space to low-dimensional space) for shape and texture reconstruction, resulting in the loss of fine facial details. A different approach is needed to estimate 3D faces beyond 3DMM.

In future works, we aim to empower our model to tackle the above mentioned issues, including patch-wise 3D face reconstruction, training on expanded face dataset, and reconstruction surpassing the constraints posed by 3DMM.

6 Conclusion

In this work, we proposed a novel ***RE**duced **D**ependency **F**ast **U**nsuperviSEd 3D Face Reconstruction* (***RED-FUSE***) framework to reconstruct 3D faces from unprocessed face images in an unsupervised manner without posing additional dependencies. In particular, RED-FUSE is trained on various 2D face datasets using a multi-pipeline training architecture. A novel pose transfer scheme is exploited to learn the accurate 3D face representation without affecting shape and texture accuracy. This enables lesser dependencies and improved estimation during inference. Our experiments indicate that the proposed model improves the perceptual error, NoW selfie challenge, 3D shape, and color-based error by a large margin of **27.4**%, **29.6**%, **46.2**%, and **15.1**%, respectively, outperforming the current method. Moreover, our approach significantly reduces testing time, i.e., **74.6**%; thus, RED-FUSE not only reduces test time dependencies and improves estimation speed, but also produces reliable 3D faces. Due to the reconstruction accuracy, lower dependencies, and speed, the proposed model is beneficial for real-time applications.

Author Contributions We summarize the author's contributions as (1) HT: conceptualization, data curation, formal analysis, investigation, methodology, project administration, validation, visualization and writing original draft, (2) VKS: analysis, investigation, supervision, validation and draft revision, and (3) Y-SC: supervision and draft revision.

Funding This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of data and materials. All the data and materials are freely available in the public domain.

Declarations

Ethical approval This research uses freely available face datasets. Therefore, ethical approval is not required.

Conflict of interest I declare that the authors have no competing interests as defined by Springer or other interests that might be perceived to influence the results and/or discussion reported in this paper.

References

1. Wang, Y., Liu, J., Tang, X.: Robust 3d face recognition by local shape difference boosting. IEEE Trans. Pattern Anal. Mach. Intell. **32**(10), 1858–1870 (2010)
2. Chen, L., Cao, C., De la Torre, F., Saragih, J., Xu, C., Sheikh, Y.: High-fidelity face tracking for ar/vr via deep lighting adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13059–13069 (2021)
3. Ye, D., Fuh, C.-S.: 3d morphable face model for face animation. Int. J. Image Gr. **20**(01), 2050003 (2020)

4. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, pp. 187–194 (1999)
5. Tewari, A., Zollhoefer, M., Bernard, F., Garrido, P., Kim, H., Perez, P., Theobalt, C.: High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. IEEE Trans. Pattern Anal. Mach. Intell. **42**(2), 357–370 (2018)
6. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0 (2019)
7. Tiwari, H., Chen, M.-H., Tsai, Y.-M., Kuo, H.-K., Chen, H.-J., Jou, K., Venkatesh, K., Chen, Y.-S.: Self-supervised robustifying guidance for monocular 3d face reconstruction. arXiv preprint arXiv:2112.14382 (2021)
8. Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Theobalt, C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1274–1283 (2017)
9. King, D.E.: Dlib-ml: A machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
10. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
11. Tiwari, H., Kurmi, V.K., Venkatesh, K., Chen, Y.-S.: Occlusion resistant network for 3d face reconstruction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 813–822 (2022)
12. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM Trans. Gr. (TOG) **40**(4), 1–13 (2021)
13. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (2015)
14. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition (2008)
15. Sanyal, S., Bolkart, T., Feng, H., Black, M.J.: Learning to regress 3d face shape and expression from an image without 3d supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7763–7772 (2019)
16. Tiwari, H., Subramanian, V.K.: Reduced dependency fast unsupervised 3d face reconstruction. In: 2022 IEEE International Conference on Image Processing (ICIP), pp. 1021–1025 (2022). IEEE
17. Feng, M., Gilani, S.Z., Wang, Y., Mian, A.: 3d face reconstruction from light field images: A model-free approach. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 501–518 (2018)
18. Kemelmacher-Shlizerman, I., Basri, R.: 3d face reconstruction from a single image using a single reference face shape. IEEE Trans. Pattern Anal. Mach. Intell. **33**(2), 394–405 (2010)
19. Zhu, W., Wu, H., Chen, Z., Vesdapunt, N., Wang, B.: Reda: reinforced differentiable attribute for 3d face reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4958–4967 (2020)
20. Tiwari, H., Subramanian, V.K., Chen, Y.-S.: Real-time self-supervised achromatic face colorization. The Visual Computer, 1–16 (2022)
21. Tiwari, H., Subramanian, V.K.: Self-supervised cooperative colorization of achromatic faces. In: 2022 IEEE International Conference on Image Processing (ICIP), pp. 236–240 (2022). IEEE
22. Liu, F., Zhu, R., Zeng, D., Zhao, Q., Liu, X.: Disentangling features in 3d face shapes for joint face reconstruction and recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5216–5225 (2018)
23. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. IEEE Trans. Visual Comput. Graphics **20**(3), 413–425 (2013)
24. Zhu, X., Yang, F., Huang, D., Yu, C., Wang, H., Guo, J., Lei, Z., Li, S.Z.: Beyond 3dmm space: Towards fine-grained 3d face reconstruction. In: European Conference on Computer Vision, pp. 343–358 (2020). Springer
25. Tewari, A., Zollhöfer, M., Garrido, P., Bernard, F., Kim, H., Pérez, P., Theobalt, C.: Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2549–2559 (2018)
26. Sela, M., Richardson, E., Kimmel, R.: Unrestricted facial geometry reconstruction using image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1576–1585 (2017)
27. Tran, A.T., Hassner, T., Masi, I., Paz, E., Nirkin, Y., Medioni, G.G.: Extreme 3d face reconstruction: Seeing through occlusions. In: CVPR, pp. 3935–3944 (2018)
28. Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., Freeman, W.T.: Unsupervised training for 3d morphable model regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8377–8386 (2018)
29. Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S.: Fast-GANFIT: Generative Adversarial Network for High Fidelity 3D Face Reconstruction
30. Tu, X., Zhao, J., Xie, M., Jiang, Z., Balamurugan, A., Luo, Y., Zhao, Y., He, L., Ma, Z., Feng, J.: 3d face reconstruction from a single image assisted by 2d face images in the wild. IEEE Trans. Multimedia **23**, 1160–1172 (2020)
31. Zeng, X., Peng, X., Qiao, Y.: Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2315–2324 (2019)
32. Zhang, R., Tsai, P.-S., Cryer, J.E., Shah, M.: Shape-from-shading: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **21**(8), 690–706 (1999)
33. Yang, X.: Feature sharing attention 3d face reconstruction with unsupervised learning from in-the-wild photo collection. In: Journal of Physics: Conference Series, vol. 2258, p. 012051 (2022). IOP Publishing
34. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 296–301 (2009). Ieee
35. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
36. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 146–155 (2016)
37. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV, pp. 3730–3738. IEEE Computer Society, ??? (2015)
38. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74 (2018). IEEE
39. Amos, B., Ludwiczuk, B., Satyanarayanan, M.: Openface: A general-purpose face recognition library with mobile applications. CMU School Comput. Sci. **6**(2), 20 (2016)

40. Serengil, S.I.: tensorflow-101. https://github.com/serengil/tensorflow-101 (2021)
41. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
42. Zhong, Y., Deng, W., Hu, J., Zhao, D., Li, X., Wen, D.: Sface: Sigmoid-constrained hypersphere loss for robust face recognition. IEEE Trans. Image Process. **30**, 2587–2598 (2021)
43. Chen, D., Hua, G., Wen, F., Sun, J.: Supervised transformer network for efficient face detection. In: European Conference on Computer Vision, pp. 122–138 (2016). Springer
44. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
45. Kingma, D.P., Ba, J.: Adam: A methodfor stochastic optimization. In: International Conference onLearning Representations (ICLR) (2015)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Hitika Tiwari is currently a dual degree Ph.D. research scholar in the Department of Computer Science and the Department of Electrical Engineering at the National Yang Ming Chiao Tung University and Indian Institute of Technology Kanpur, respectively. She received her M.Tech degree from the Indian Institute of Technology Mandi. Her research interests are computer vision, image processing, and machine learning.

Venkatesh K. Subramanian received a B.E degree in electronics from Bangalore University in 1987 and an M.Tech and a Ph.D. in 1989 and 1995, respectively, from the Indian Institute of Technology, Kanpur, India. He worked at IIT Guwahati from 1995 - 1999 in the Department of Electronics and Communication. Currently, he is working as a Professor at the Electrical Engineering department at the Indian Institute of Technology, Kanpur. His research interests include image/video processing and its application in machine vision, vision for robotics, computational photography, light fields, and aerial navigation. He has 150+ publications and 25+ patents.

Yong-Sheng Chen received B.S. and M.S. from National Chiao Tung University in Sep. 1993 and National Taiwan University in 1995, respectively. He completed his Ph.D. at National Taiwan University in 2001. Currently, he is working as an Associate Professor in the Department of Computer Science and Engineering at National Yang Ming Chiao Tung University, Taiwan. His research interests include biomedical, signal processing, medical image processing, and computer vision. He has 99 research outputs, including 8 patents.