Noname manuscript No.
(will be inserted by the editor)

Self-supervised learning for fine-grained monocular 3D face reconstruction in the wild

Dongjin Huang[1] · Yongsheng Shi[1] · Jinhua Liu[1] · Wen Tang[2]

Received: date / Accepted: date

Abstract Reconstructing 3D face from monocular images is a challenging computer vision task, due to the limitations of traditional 3DMM (3D Morphable Model) and the lack of high-fidelity 3D facial scanning data. To solve this issue, we propose a novel coarse-to-fine self-supervised learning framework for reconstructing fine-grained 3D faces from monocular images in the wild. In the coarse stage, face parameters extracted from a single image are used to reconstruct a coarse 3D face through a 3DMM. In the refinement stage, we design a wavelet transform perception model to extract facial details in different frequency domains from an input image. Furthermore, we propose a depth displacement module based on the wavelet transform perception model to generate a refined displacement map from the unwrapped UV textures of the input image and rendered coarse face, which can be used to synthesize detailed 3D face geometry. Moreover, we propose a novel albedo map module based on the wavelet transform perception model to capture high-frequency texture information and generate a detailed albedo map consistent with face illumination. The detailed face geometry and albedo map are used to reconstruct a fine-grained 3D face without any labeled data. We have conducted extensive experiments that demonstrate the superiority of our method over existing state-of-the-art approaches for 3D face reconstruction on four public datasets including CelebA, LS3D, LFW, and NoW benchmark. The experimental results indicate that our method achieved higher accuracy and robustness, particularly of under the challenging conditions such as occlusion, large poses, and varying illuminations.

Keywords 3D face reconstruction · Monocular image · 3DMM · Self-supervised learning · Coarse-to-fine model

1 Introduction

The 3DMM [1] has enabled the use of 3D face reconstruction methods various applications such as face recognition [2, 3], face editing [4], and face animating [5, 6]. In recent years, deep learning methods have become increasingly popular for 3D face reconstruction using images [7]. These methods offer significant advantages over traditional 3DMM models by improving the quality of reconstruction under varying image conditions [8, 9].

Previous methods [10–14] to generate smooth 3D faces through linear parameters regressed by using CNN in the prior 3D face models, since detailed representations of face geometry and texture were lost in latent facial coefficients. Some methods have been proposed to restore detailed face geometry using supervised [15–22] or self-supervised learning [23–26]. Supervised learning methods utilize insufficient ground-truth 3D facial scanned or synthesized training data, which is costly and time-consuming to collect a large number of high-quality 3D face data (unrestricted illumination, expression and pose). Consequently, these methods regress facial parameters of statistical face model through finite linear geometry space, which extracts only low-frequency 3D mesh information. This generates overly smoothed 3D reconstructions that lack fine facial details. Self-supervised methods use face images in-the-wild without any labeled data extensively in order to improve the effectiveness and expressiveness of 3D face reconstructions. This is done by minimizing the pixel-wise errors between the input image and the

✉ Dongjin Huang
djhuang@shu.edu.cn
✉ Yongsheng Shi
yongsheng@shu.edu.cn
1 Shanghai Film Academy, Shanghai University, Shanghai 200072, China

2 Department of Creative Technology, Bournemouth University, Poole BH12 5BB, UK

rendered image by re-projecting the generated 3D face into the 2D space. Although these methods can reconstruct high-fidelity facial geometry and texture information, they are prone to generate magnanimous redundancy and noise, making it difficult to generate 3D faces with rich details that are close to the input.

In this paper, we propose a novel self-supervised learning framework for reconstructing high-fidelity 3D faces from monocular images taken in real-world environments, using a coarse-to-fine scheme. For the coarse stage, we train a regressor using monocular images without ground-truth data, and predict latent representations of the 3DMM model, which include shape, expression, albedo, pose, and illumination coefficients. The latent facial parameters are calculated by increasing the similarity between the rendered image and input image. It is robust for the facial images of occlusion, large pose, and illumination variations. Although the coarse model can accurately construct 3D faces, the resulting geometry and texture are smooth and lack details and wrinkles that change with expression. Therefore, in the fine stage, in order to obtain more abundant facial details, we first use the wavelet transform to extract high-frequency, medium-frequency, and low-frequency features from the input facial images. We then design a depth displacement module based on the wavelet transform perception model to synthesize refined displacement maps without ground-truth displacement labels, using the unwrapped images from the input and rendered images in the coarse stage. The detailed depth displacement map can be used to reconstruct 3D face geometry with expression-dependent details. Furthermore, we propose a novel albedo module of an encoder-decoder structure based on the wavelet transform perception model to generate detailed albedo map that can be used for restoring high-quality facial texture with details. And the resulting faces have consistent illumination with the input images. Our extensive qualitative and quantitative experiments demonstrate the superiority of our method in reconstructing high-fidelity 3D faces with rich details. Our method outperforms other comparative methods on different datasets and achieves state-of-the-art reconstructions with accurate 3D geometry and high-quality texture from a single face image.

The main contributions of this work are summarized as follows:

- For the limitations of traditional 3DMM and the lack of high-fidelity 3D facial data, we propose a novel coarse-to-fine 3D face reconstruction framework using self-supervised learning, where a 3D face generated from a single image in the wild is gradually refined at different reconstruction stages.
- To extract rich details, we design a wavelet transform perception model that can effectively obtain facial features in different frequency domains from input images.
- To reconstruct fine-grained 3D face geometry, a depth displacement module based on the wavelet transform perception model is proposed to robustly generate detailed depth displacement map. The detailed depth displacement map can be used to synthesize 3D face geometry with expression-dependent wrinkles in the fine stage.
- We propose a novel albedo module based on the wavelet transform perception model to generate detailed albedo map that can be used for recovering high-fidelity 3D textures with rich details.

2 Related work

Generally, 3D face reconstruction methods are divided into traditional-based methods and deep learning-based methods. Traditional-based 3D face reconstruction methods [27] are easy to generate 3D faces with artifacts, and they are difficult to reconstruct accurate faces from facial images on in-the-wild. In recent decades, 3D face reconstruction methods based on deep learning have made great progress and attracted much attention. This section reviews the works related to our methods in brief, including supervised reconstruction, self-supervised reconstruction and fine-grained reconstruction.

Supervised reconstruction: Many methods [28–37] perform monocular 3D face reconstruction using 3DMM parameters regressed from paired training data. During the training process, they utilize loss function to minimize the difference between the outputs and ground-truth data. Guo et al. [28] train a unified 3D facial model on different sources from ground-truth 3D scanned data, RGB-D and in-the-wild images. Although this facial model has a powerful ability to generate 3D faces, it comes at great cost to collect and process training data. SADRNet [29] directly regresses 3D geometry from the AFLW2000-3D [30] and the Florence 3D faces [31] to alleviate the limitation of 3DMM, which is capable of handling occluded faces, but only suitable for a set of carefully selected face images. Multi-view images of a person captured under different conditions together with a small number of labeled 3DMM parameters have been used as ground-truth to train a 3D reconstruction model with an proposed encoder-decoder framework [36]. This model can robustly reconstruct 3D faces for a specific-person, but it is not applicable to other people due to the insufficient training data. To solve the problem of monocular 3D face reconstruction when the distance between face and camera is close, PerspNet [37] is proposed to simultaneously reconstruct a 3D face and estimate 6DoF (6 Degrees of Freedom) face pose by using PnP. Moreover, a large-scale 3D dataset with ground-truth 3D face mesh and corresponding 6DoF pose annotations are collected for the PerspNet training. The above supervised-based 3D face reconstruction approaches are trained requiring ground-truth 3D scans, synthesized data or facial parameters by fitting prior 3DMM.

Self-supervised reconstruction: Self-supervised 3D face reconstruction methods make the use of readily available images in-the-wild or synthetic image data without any labels. A non-linear 3D face deformation model has been trained as an improved version of the linear 3DMM via a large number of real-world facial images [38]. This non-linear model has greater capability for representing varied 3D shapes, albedo, and illumination models than the optimization-based linear 3DMM. However, faces generated by this method [38] are not realistic, and the skin colour differs greatly from the input images. NeRF (Neural Radiance Field) [39, 40] also has significantly improved representations of 3D scenes through self-supervised learning, and uses volume rendering to map the latent codes extracted from multi-view images to realistic images. Based on ray casting for volume rendering, NeRF method is high computational cost. RingNet [41] and 3DFFA_V2 [42] can robustly reconstruct 3D face geometries without texture from a single face image in challenge conditions in-the-wild, such as occluded and large poses. MGCNet [43] method improves the realism of reconstructed 3D faces with high-fidelity from a single image by exploiting a new view synthesis algorithm based on occlusion perception. However, its reconstruction results degenerates quickly for large expressions, varied lighting conditions, and head poses. While these self-supervised reconstruction methods do not require paired data, it is difficult for them to generate detailed 3D faces such as wrinkles with expression and high-frequency textures. In this paper, we propose a novel pipeline of fine-grained reconstruction to recover the 3D faces with rich details from the coarse-bench, using a self-supervised learning.

Fine-grained reconstruction: More recently, for many computer generated imaging, such as digital games, movies post-productions and VR/AR applications, fine-grained 3D face reconstructions are essential. FaceScape [18] builds a large-scale 3D faces with rich details to train a detailed 3D face model, which is capable of generating highly realistic 3D face geometries from a single face images under laboratory conditions. FaceVerse [21] uses a hybrid data set of RGB-D images and detailed 3D head scans to learn a fine-grained facial model via a novel coarse-to-fine scheme. This model can recover facial geometry details and appearances based on a conditional StyleGAN network. Deng et al. [17] proposed a fine-scale pipeline to transfer facial wrinkles from the source 3D face to the target 3D faces through a supervised learning. The above supervised learning methods for fine-grained reconstruction, although are capable of reconstructing fine-grained 3D faces under laboratory controlled conditions, can generate unrealistic facial expressions and noise. This is because, when complex face images in the wild are used as input, the limitations of training data sets and the distribution of the collected training data with ground-truth that does not consistent with the face images in-the-wild, leading to degenerated reconstruction results.

Self-supervised learning methods for fine-grained reconstruction with displacement map [16, 19, 23–25] are able to capture facial details using monocular face images in the wild, which greatly extended the application domain capabilities of 3D face reconstructions. In essence, these methods predict the displacement map with high-frequency information from facial images, and combine the information with the coarse-scale 3D shape to generate the fine-grained face. DECA [23] and EMOCA [24] utilize the loss formulations of shape-consistency and detail-consistency to generate 3D facial wrinkles with the change of expressions. However, these methods struggle to reconstruct high-fidelity 3D faces with rich details, as the detail parameters extracted by the detail encoding network only contain a small amount of facial structural information. Moreover, they are unable to generate photo-realistic textures, since the UV albedo map contains fewer texture details. The method [25] also uses displacement maps predicted by the image-to-image translation network to capture facial details without any 2D-3D data as ground truth. However, it lacks robustness under challenging conditions and is prone to generate 3D faces with scratches.

In this paper, we propose a novel self-supervised learning for fine-grained 3D face reconstruction based on coarse-to-fine framework. To reconstruct fine-grained 3D faces, we first utilize the haar-based wavelet transform to extract facial high-frequency features from the input images. Moreover, we propose a novel depth displacement module and an albedo module based on the wavelet transform perceptual model to generate a displacement map and albedo map with rich details from the unwrapped image of the input as well as the rendered coarse image in the UV space. Our proposed approach is robust and accurate for facial images in the wild, even under challenging conditions such as self-occlusions, varying lighting conditions, and extreme poses.

3 Proposed method

We propose a novel fine-grained 3D face reconstruction approach of coarse-to-fine structure from monocular images on in-the-wild. We will explain the specific pipelines of each module in the framework as follows.

3.1 Overview

Given a single facial image, our goal is to reconstruct corresponding fine-grained 3D faces with high-fidelity. To do so, we propose a novel multi-stage 3D face reconstruction architecture by using self-supervised learning, in which the generated face is gradually refined at different stages as depicted in Fig. 1.

Our framework consists of two pipelines: coarse stage and fine-grained stage. In the coarse reconstruction stage, we adopt 3DMM regression model based on VGG-Face network [44] named R-Net to regress facial coefficients of

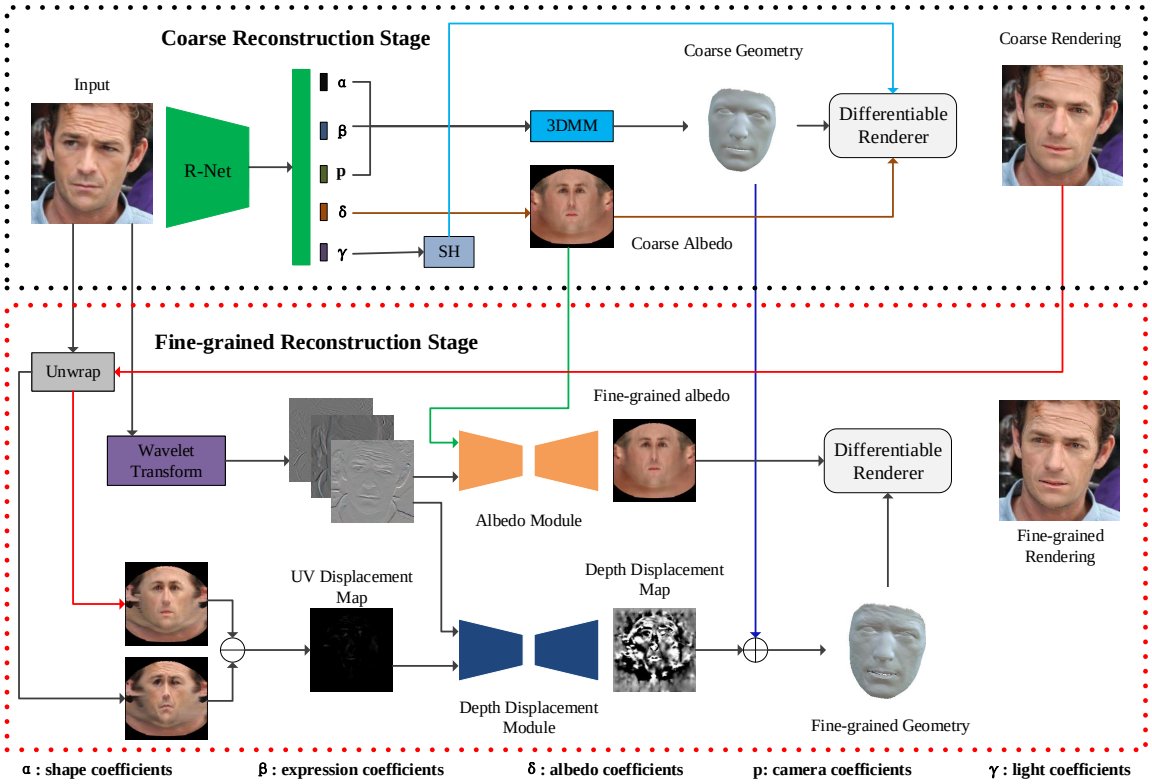

Fig. 1. The overview of the proposed framework for coarse-to-fine 3D face reconstruction: the black pipeline block is Coarse Reconstruction Stage; the red pipeline block is Fine-grained Reconstruction Stage.

shape, expression, albedo, pose, and illumination from input images. Albedo map and face geometry are generated with the prior 3DMM model from the albedo parameters and the facial expression and shape parameters, respectively. The coarse 3D faces are reconstructed using differentiable rendering without any ground-truth labels. For the fine-grained reconstruction pipeline, we propose a separate albedo module and a depth displacement module based on wavelet transform perception model to recover facial details so that we can effectively reduce the limit of the 3DMM model that can only reconstruct smooth 3D faces. Firstly, we exploit a wavelet transform to extract the facial detailed features in different frequency domains from the input images. The high-frequency features and smooth albedo maps as inputs are sent to the albedo module to predict high-quality albedo maps. Then, the inputs and rendered coarse images are unwrapped into UV space in this coarse stage. Furthermore, the depth displacement module is designed to predict depth displacement maps with details from these unwrapped images and high-frequency features extracted by wavelet transform from input images. At the same time, the 3D face geometry with rich details and wrinkles that are changing with expressions can be generated from the detailed displacement maps. Finally, the fine-grained face images are rendered by given the high-quality albedo maps and the 3D face geometry.

3.2 3DMM Model

We use a CNN-based regressor R-Net to regress the facial coefficients of 3DMM model with self-supervised learning. The 3DMM model utilizes the shape, expression and albedo parameters to generate the 3D face shape and the albedo. A differentiable renderer is then used for rendering the 3D shape and the albedo to a 2D face image through the parameters of the illumination and the face pose.

3D Face Model: We take a 3DMM as a surrogate 3D face model to represent the 3D face with the facial identity, expression and albedo PCA bases. Given the parameters of the facial identity, expression and albedo, the 3D facial shape and albedo are represented as follows:

$$S = S\left(\alpha, \beta \mid \bar{S}, B_{id}, B_{exp}\right) = \bar{S} + B_{id}\alpha + B_{exp}\beta \tag{1}$$

$$T = T\left(\delta \mid \bar{T}, B_t\right) = \bar{T} + B_t\delta \tag{2}$$

where $\bar{S} \in \mathbb{R}^{3N}$ and $\bar{T} \in \mathbb{R}^{3N}$ represent the average 3D face shape and albedo, respectively. $B_{id} \in \mathbb{R}^{3N\times|\alpha|}$, $B_{exp} \in \mathbb{R}^{3N\times|\beta|}$ and $B_t \in \mathbb{R}^{3N\times|\delta|}$ indicate PCA bases of the facial identity, expression and albedo, respectively.

$\alpha \in \mathbb{R}^{|\alpha|}$, $\beta \in \mathbb{R}^{|\beta|}$ and $\delta \in \mathbb{R}^{|\delta|}$ refer to the corresponding parameter vectors of the identity, expression and albedo. We utilize $\bar{S}$, $\bar{T}$, B_{id} and B_t are built from the BFM [45], and the expression bases B_{exp} from FaceWarehouse [46]. As a result, the 3D face can be formulated as follows:

$$M = \{S, T\} = \begin{Bmatrix} S = \bar{S} + B_{id}\alpha + B_{exp}\beta, \\ T = \bar{T} + B_t\delta \end{Bmatrix} \tag{3}$$

As in Eq. (3), the 3DMM model finally outputs a 3D face Mesh containing a set of vertices ($N = 53215$).

Camera Model: Since the training data used is from face images in-the-wild, we use the perceptual camera model to project 3D geometry into 2D image space with a fixed scale factor. The face pose parameters $p = \{R, T\}$ can be represented by a rotation matrix $R \in SO(3)$ and a translation vectors $t \in \mathbb{R}^3$. 3D vertices are projected into the 2D space as:

$$v = \prod (fRM_i + t) \tag{4}$$

where $M_i \in \mathbb{R}^3$ indicates the vertex position in the 3D face Mesh M. $\prod \in \mathbb{R}^{2\times 3}$ performs the orthogonal operation as the 3D-2D projection. And v is the projected 2D vertex.

Illumination Model: For shadow shading, assuming Lambertian is used for distant light sources and facial surfaces, and SH (Spherical Harmonics) is used for the environmental illumination model [47]. The shade image can be calculated as:

$$C(\gamma \mid t_i, n_i) = t_i \odot \sum_{b=1}^{B^2} \gamma_b \Phi_b(n_i) \tag{5}$$

where we utilize $b = 3$ SH bands representing the scene illumination [23]. $\gamma \in \mathbb{R}^{27}$ denotes the illumination parameters for RGB face images. t_i represents the skin texture in albedo T. n_i is normal which is built from the 3D face geometry M. $\Phi_b \in \mathbb{R}^3 \to \mathbb{R}$ is SH basis formulation, and $\odot$ represents the operation of the Hadamard product.

Differentiable rendering: Once the facial parameters are given for identity α, expression β, albedo δ, pose p and illumination γ, we can generate a 2D face image I_r by the differentiable rendering [34] as $I_r = R(M(\alpha, \beta, \delta), p, C(\gamma))$, where R represents the rendering operation.

In this paper, we utilize the R-Net regressor to regress these facial parameter vectors $\{\alpha, \beta, \delta, p, \gamma\} \in \mathbb{R}^{222}$, of which 80 is for the face identity, 29 for expression, 80 for albedo, 6 for pose and 27 for illumination by modifying the last fully connection layer from an input image. This regressor is named R-Net for monocular face reconstruction. We describe how to train the R-Net in the following section.

3.3 Coarse reconstruction

In the coarse reconstruction stage, we train a coarse 3D face reconstruction model from a single image in-the-wild with a self-supervised learning. We adopt R-Net based on VGG-Face [44] to encode a face image to the 3DMM coefficients $\{\alpha, \beta, \delta, p, \gamma\} \in \mathbb{R}^{222}$, which can be used to synthesize a coarse face image via differentiable rendering as shown in Fig. 1. R-Net is trained using the following loss functions by minimizing the disparity between the input image and the rendered image.

$$\mathcal{L}_{coarse} = \lambda_{pho}\mathcal{L}_{pho} + \lambda_{lmk}\mathcal{L}_{lmk} + \lambda_{id}\mathcal{L}_{id} + \mathcal{L}_{reg} \tag{6}$$

where $\mathcal{L}_{pho}$, $\mathcal{L}_{lmk}$, $\mathcal{L}_{id}$ and $\mathcal{L}_{reg}$ are the representations for the photometric loss, landmark loss, perceptual identity loss, and regularization term loss, respectively. Coefficients λ_{pho}, λ_{lmk}, λ_{id} denote the weights of these loss terms and are set to constants. The details are shown as follows.

Photometric loss: We adopt this photometric loss to enhance the pixel-level of the rendered image to be similar to the input image. This loss is expressed as:

$$\mathcal{L}_{pho} = \frac{1}{\sum_{(i,j)\in\mathcal{M}} V(i,j)} \sum_{(i,j)\in\mathcal{M}} \left\| V(i,j) * (I(i,j) - I_R(i,j)) \right\|_2 \tag{7}$$

where $\mathcal{M}$ represents the index of an image pixel. I and I_R are the input facial image and the rendered facial image. V is a facial mask obtained by the method [42]. Its value is 1 in the face skin area, and 0 elsewhere.

Landmark loss: Landmark loss measures the error between the 68 landmarks of the input face image and the re-projected 2D locations of 68 key points in the 3D geometry. It can effectively align the pose and expression of 3D faces. The 2D landmarks are detected by the common 3D face alignment method [48]. This loss is defined as:

$$\mathcal{L}_{lmk} = \frac{1}{N} \sum_{i=1}^{N} \left\| q_i - q_i^R \right\|_2^2 \tag{8}$$

where q_i and q_i^R are the i-th landmark location of the input face image and the rendered face image, respectively. $N = 68$ represents the number of face landmarks.

Perceptual identity loss: Recently, perceptual identity loss performs the effectiveness for 3D face reconstruction [23, 24], which measures the identity similarity between the input and the reconstructed images. Inspired by this, we utilize a pre-trained face recognition network VGG-Face [44] to compute the perceptual identity loss. This perceptual identity loss is computed as:

$$\mathcal{L}_{id} = \|\Phi(I) - \Phi(I_R)\|_2^2 \tag{9}$$

where $\Phi(I)$ and $\Phi(I_R)$ represent features extracted by the VGG-Face in the input facial image and the reconstructed facial image.

Regularization: We propose a regularization term for the shape, expression and albedo parameters of human face. It prevents these values from being too large, otherwise the reconstructed 3D geometry and textures will be distorted. This regularization term loss is expressed as:

$$\mathcal{L}_{reg} = \lambda_\alpha \|\alpha\|^2 + \lambda_\beta \|\beta\|^2 + \lambda_\delta \|\delta\|^2 \tag{10}$$

where λ_α, λ_β and λ_δ denotes the weights of these regularization terms.

3.4 Fine-grained reconstruction

The albedos and shapes generated at the coarse stage are constrained by the prior 3DMM model and can only obtain low-frequency components of the skin texture and the geometry. Our goal is to generate high-quality albedo maps with detailed shapes. Therefore, we propose an novel albedo module and a depth displacement module based on the wavelet transform perception, respectively, to generate high-quality albedo maps and depth displacement maps with rich details.

3.4.1 Wavelet transform

Wavelet transform is widely used in image processing to decompose images into different frequency domains. Currently, existing 3D face reconstruction methods do not use it. In this paper, we first introduce a Haar-based wavelet transform to perform low-pass and high-pass filtering from the horizontal and vertical directions, which contains four kernels: $\{LL^T, LH^T, HL^T, HH^T\}$. Where, the low-pass filter $L^T = \frac{1}{\sqrt{2}}[1, 1]$ is to extract low-frequency signals on the smooth surface. The high-pass filter $H^T = \frac{1}{\sqrt{2}}[-1, 1]$ is to capture high-frequency signals in the horizontal, vertical, and diagonal directions of a face image. The first-order Haar-based wavelet transform to decompose different frequency components is illustrated in the Fig. 2. Where, (a) denotes input images; (b) LL represents the low-frequency information of the inputs; (c) LH indicates low-frequency information in the horizontal direction and high-frequency information in the vertical direction; (d) HL denotes high-frequency information in the horizontal direction and low-frequency information in the vertical direction; and (e) HH indicates high-frequency in the diagonal directions.

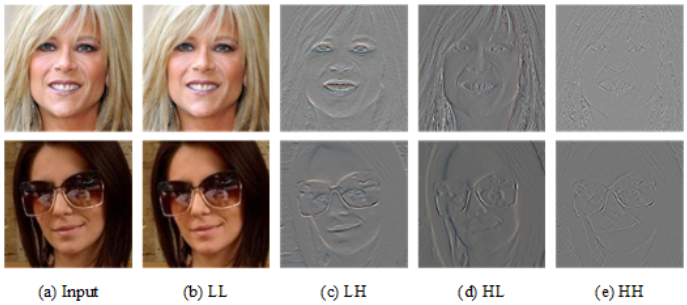

Fig. 2. The demonstration of wavelet transformation.

We can observe that the LL frequency component is close to the input image and contains noise such as background areas except for the face. Almost all facial structural details are stored in the high-frequency components. Therefore, in this paper, we ignore the low-frequency information of LL and adopt the high-frequency components of LH, HL, and HH to capture rich details of face images.

3.4.2 Albedo module

To alleviate the limitations of the prior 3DMM in generating high-fidelity textures with rich details, we propose a novel albedo module with an encoder-decoder structure based on the wavelet transform perception encoder. The bottom wavelet transform perception encoder is used to extract high-frequency texture features and feed them to the top decoder of the albedo module (see Fig. 3). Specifically, we adopt haar-based transform decompose the input image into different frequency components. Given the smooth albedo map and the high-frequency features of LH, HL, and HH, the bottom encoder branch based on wavelet transform perception is used to extract detailed information, and the encoder-decoder branch of the albedo module is dedicated to synthesizing the high-fidelity albedo map by using self-supervised learning. The detailed albedo map can be used to generate facial textures with rich details and realistic expressions, while maintaining the skin color consistent with the input image.

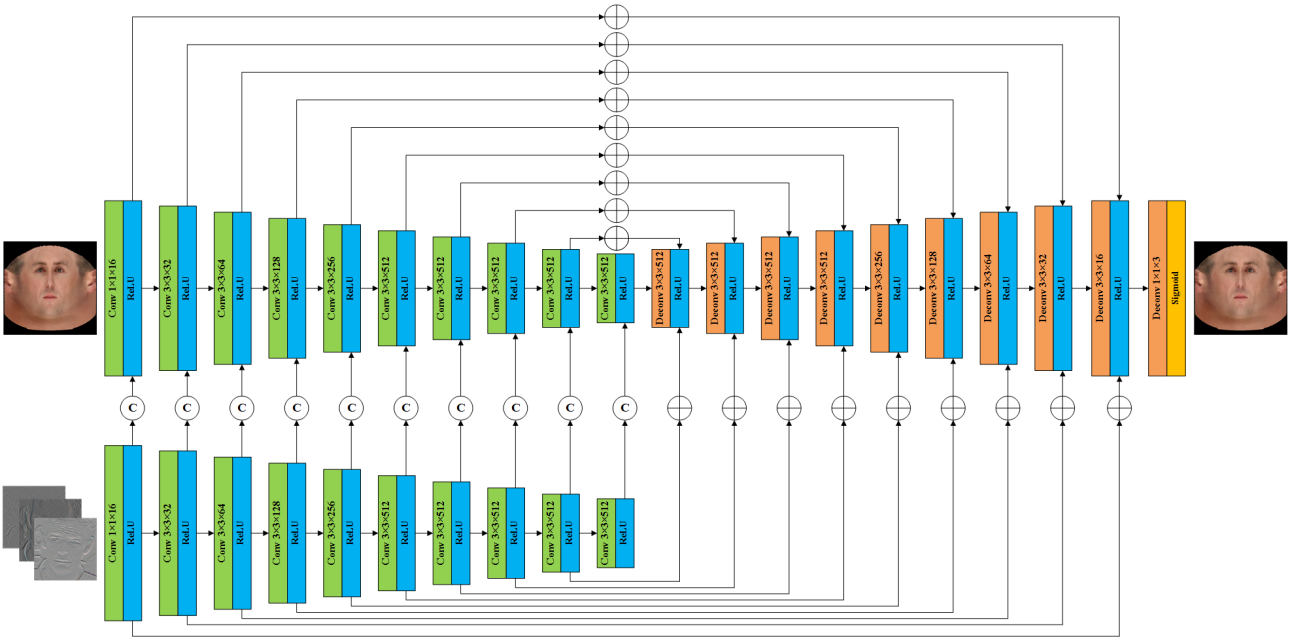

Fig. 3. Architecture of proposed albedo module.

3.4.3 Depth displacement module

To generate fine-grained 3D geometry with rich wrinkles that change with expression, we propose a novel depth displacement module based on the wavelet transform perception with two parallel encoder-decoder branches, as depicted in Fig. 4. Firstly, we unwrap the input image and the rendered face image in this coarse stage into UV space. Then, we obtain the increment image by pixel-level subtraction of these two unwrapped images. We use the bottom encoder-decoder branch to extract depth displacement features from the increment image. Then, we utilize the upper wavelet transform perception-based encoder-decoder branch to generate a depth displacement map with rich details from the high-frequency features of LH, HL, and HH of the input face image. Fine-grained 3D face geometry with structural details can be reconstructed via the depth displacement map. Finally, photo-realistic face images can be rendered from high-quality albedo maps and detailed 3D face geometry.

3.4.4 Loss functions

In the pipeline of fine-grained 3D face reconstruction, we utilize in-the-wild facial images to train the albedo module and the depth displacement module with self-supervised learning by minimizing the following loss functions:

$$\mathcal{L}_{fine} = \lambda_{pho}\mathcal{L}_{pho} + \lambda_{id}\mathcal{L}_{id} + \lambda_{alb}\mathcal{L}_{alb} + \lambda_{smo}\mathcal{L}_{smo} \tag{11}$$

The advantage of the loss function is that it is robust to reduce artifacts and distortions in the reconstruction process. Where, $\mathcal{L}_{pho}$ and L_{id} represent the photometric loss and the perceptual identity loss in this coarse reconstruction stage, respectively. $\mathcal{L}_{alb}$ and $\mathcal{L}_{smo}$ represent the albedo loss and the smoothness loss, respectively. λ_{pho}, λ_{id}, λ_{alb} and λ_{smo} denote the weights of these loss terms and are set to constants. Details for the albedo loss and the smoothness loss are shown as follows.

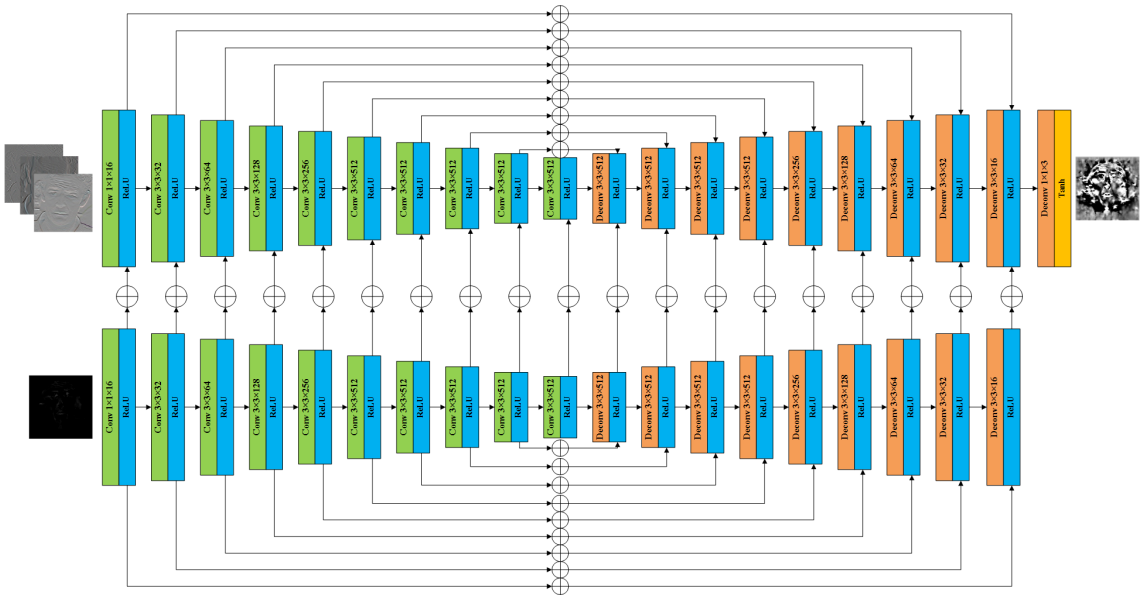

Fig. 4. Architecture of proposed displacement map module.

Albedo loss: $\mathcal{L}_{alb}$ is the consistency regularization term used to prevent baking residual shadows in the detailed albedos, and to avoid the illumination model unable to recover the lighting correctly. This albedo loss is expressed as:

$$\mathcal{L}_{alb} = \left\| \hat{A} - A \right\|_2^2 \tag{12}$$

where $\hat{A}$ represents the detailed albedo map. A represents the smooth albedo map generated by the coarse reconstruction.

Smoothness loss: The smoothness loss measures the similarity of the neighboring pixel-wise on both the normal map and the depth displacement map [25]. Furthermore, the smoothness loss effectively reduces the influence of some face images in the case of occlusions. This smoothness loss is defined as:

$$\mathcal{L}_{smo} = \sum_{i \in \mathcal{V}_{UV}} \sum_{j \in \mathcal{N}(i)} w_n \left\| \Delta n(i) - \Delta n(j) \right\|_2^2 + w_d \left\| \Delta d(i) - \Delta d(j) \right\|_2^2 \tag{13}$$

where $\Delta n(i)$ denotes distance of the i-th pixel between the coarse normal map (i.e., the vertex normals are calculated by the coarse 3D face model) and the refined normal map generated from the displacement map. $\Delta d(i)$ represents distance of the $i-th$ pixel distance between the original depth map obtained by the coarse reconstruction and the detailed depth map generated from the displacement map. The $\mathcal{V}_{UV}$ are vertices in the UV space and $\mathcal{N}(i)$ are the neighbors of the i-th vertex with radius 1. w_n and w_d are the weights of these two loss terms.

4 Experimental results

4.1 Implementation details

We use publicly available CelebA [49] dataset in-the-wild to train our model. Before training, the face detection method [48] is used to detect and align these images. The size of input images is cropped to 300×300. Finally, we collect 191k face images, in which we randomly divide 181k as training and the remaining 10k as testing. To effectively verify the generalization of our model, we also evaluate our method on the LS3D [48], LFW [50], and NoW [41] datasets. We also follow the same protocol of [48] to extract over 5k and 12k face images from LS3D and LFW, respectively.

We train our 3D face reconstruction model separately in two stages. In the coarse reconstruction stage, we initialize all weights randomly and train the face regression model R-Net based on VGG-Face [44]. We adopt Adam optimizer with the batch size of 16, learning rate of $1e$ - 5, and 250000 iterations. Furthermore, we set weight parameters $\lambda_{pho} = 1.0$, $\lambda_{lmk} = 0.5$, $\lambda_{id} = 0.8$ in Eq. (6), and $\lambda_{\alpha} = 3.0$, $\lambda_{\beta} = 10.0$, $\lambda_{\delta} = 1.0$ in Eq. (10). For the fine-grained reconstruction stage, the trained face regression model R-Net is freezed, while the albedo module and the depth displacement module are trained with the same CelebA [49] dataset as the coarse stage. We adopt Adam optimizer with batch size of 12, learning rate of $5e$ - 5, and 200000 iterations. Moreover, we set weight

parameters $\lambda_{pho} = 1.0$, $\lambda_{id} = 0.8$, $\lambda_{alb} = 1.0$, $\lambda_{smo} = 1.0$ in Eq. (11), and $w_n = 0.001$, $w_d = 0.001$ in Eq. (13). Our method is implemented in TensorFlow, using the differentiable rasterizer from 3D mesh renderer [34] to render on NVIDIA TITAN Xp GPU.

4.2 Results

To evaluate our method, we firstly compare the reconstruction results produced by the coarse and the fine-grained models under the challenging conditions in-the-wild for occlusion, large pose and different lighting (as shown in Fig. 5). As can be seen that the fine-grained 3D face reconstruction model can recover more facial structural information such as wrinkles that change with expression, and can generate more realistic facial expressions of individuals compared with that of the coarse 3D face reconstruction model. Our method can reconstruct high-fidelity facial shapes and textures from occluded facial images such as glasses, hat, hands or other objects. For some face images with large poses, our method has the capability of generating 3D faces with fine details and synthesizing vivid expressions that are consistent with the input images. Furthermore, our model can also generate highly realistic textures from the face images under different lighting conditions, while maintaining the same illumination as the input images. We further evaluate the stability and generalization of our method by comparing it with the state-of-the-art 3D face reconstruction methods in terms of quantitative and qualitative analysis.

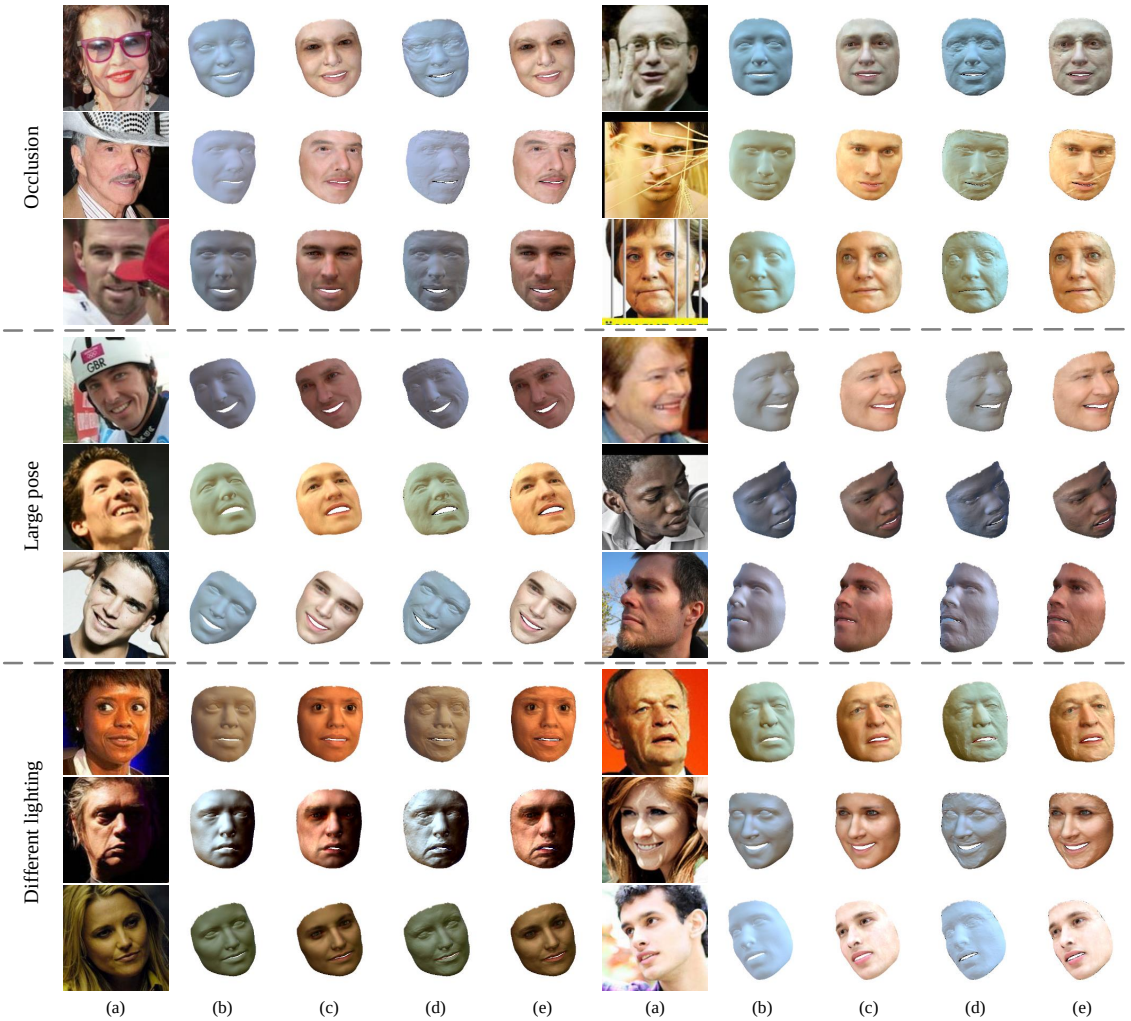

Fig. 5. Reconstruction results of our method for coarse model and fine-grained model on the challenging conditions of occlusion, large pose, and different lighting. (a) Input facial images, (b) Geometry of coarse 3D faces, (c)Texture of coarse 3D faces, (d) Geometry of fine-grained 3D faces, and (e) Texture of fine-grained 3D faces.

4.3 Evaluation on the coarse reconstruction

4.3.1 Qualitative evaluation

For the coarse 3D face reconstruction, we qualitatively compare our method with state-of-the-art coarse 3D face reconstruction methods, namely RingNet [41], Deep3DFace [22], 3DDFA_V2 [42], MGCNet [43], Chen et al. [25] (coarse), DECA [23] (coarse) and EMOCA [24] (coarse) on the test dataset of CelebA [49], and the LS3D [48] and LFW [50] datasets.

We firstly compare our 3D face reconstruction results for geometry with RingNet [41], Deep3DFace [22], 3DDFA_V2 [42], MGCNet [43], Chen et al. [25], DECA [23] and EMOCA [24] on the test dataset of CelebA [49], and the LS3D [48] and LFW [50] datasets in Fig. 6. Our method can reconstruct more accurate face geometries with detailed information such as nasolabial folds. Although RingNet [41], DECA [23] and EMOCA [24] are able to reconstruct more complete head shape, the faces generated by these methods are too smooth and lack realism compared to the input images. The quality of reconstructed 3D geometries by [25] is closer to ours. However, our method is able to reconstruct better mouth shape and more natural facial expressions (as shown in rows 2 and 4).

Since RingNet [41] and 3DDFA_V2 [42] cannot render 3D faces to 2D images, we further compare our 3D face reconstruction results for skin texture with Deep3DFace [22], MGCNet [43], Chen et al. [25], DECA [23] and EMOCA [24] on the test dataset of CelebA [49], and the LS3D [48] and LFW [50] datasets in Fig. 7. Skin textures and expressions of the faces reconstructed by DECA [23] and EMOCA [24] are far from that of the input images, since the FLAME lacks an appearance model. Compared with Deep3DFace [22] and MGCNet [43], [25] can generate more fidelity face images with slightly more details. In general, our method can reconstruct faces with specific characteristics, and the expression is more vivid and realistic.

Fig. 6. Comparison of results of the coarse 3D face reconstruction methods on the test dataset of CelebA [49], and the LS3D [48] and LFW [50] datasets.

4.3.2 Quantitative evaluation

We provide quantitative evaluation on the test dataset of CelebA [49], and the LS3D [48] and LFW [50] datasets on in-the-wild images. After detection and alignment with the face detection method [48], we collect 5572 and 12967

Fig. 7. Comparison results of the coarse 3D face reconstruction methods on the test dataset of CelebA [49], and the LS3D [48] and LFW [50] datasets.

images from the LS3D [48] and LFW [50] datasets, respectively. We compare our 3D face reconstruction model with Deep3DFace [22], MGCNet [43], Chen et al. [25] (coarse), DECA [23] (coarse) and EMOCA [24] (coarse) by evaluating the similarity between our rendered results and input images in terms of PSNR, SSIM and RMSE, duo to RingNet [41] and 3DDFA_V2 [42] cannot generate 2D face image with texture.

Table 1: Quantitative evaluation for the 3D face reconstruction methods on the test dataset of CelebA [49].

Methods	**PSNR ↑**	**SSIM ↑**	**RMSE ↓**
Deep3DFace [22]	23.561	0.821	18.466
MGCNet [43]	23.762	0.831	17.709
Tran et al. [38]	19.009	0.779	28.988
Chen et al. [25](Coarse)	25.611	0.861	13.976
Chen et al. [25](Detail)	25.720	0.867	13.689
DECA [23](Coarse)	21.641	0.783	22.629
DECA [23](Detail)	21.643	0.780	22.626
EMOCA [24](Coarse)	21.139	0.768	23.833
EMOCA [24](Detail)	21.147	0.766	23.813
Ours(Coarse)	25.794	0.863	13.705
Ours(Fine-grained)	**25.982**	**0.874**	**13.329**

In the Table 1, excepting that the SSIM score calculated by the Chen et al. [25] method on the test dataset of CelebA [49] is higher, the PSNR value calculated by our method is better, which is 9.48%, 8.55%, 35.69%, 0.71%, 19.19% and 22.02% higher than that of other state-of-the-art methods. Moreover, the RMSE value obtained by our method is the lowest, which is 25.78%, 22.61%, 52.72%, 1.94%, 39.44% and 42.50% less compared to these methods Deep3DFace [22], MGCNet [43], Chen et al. [25] (coarse), DECA [23] (coarse) and EMOCA [24] (coarse), respectively. We also quantitatively evaluate the 3D face reconstruction results on the LS3D [48] dataset, as shown in Table 2. Our method achieved lower reconstruction results, since the LS3D [48] dataset contains many

Table 2: Quantitative evaluation for the 3D face reconstruction methods on the LS3D [48] dataset.

Methods	PSNR ↑	SSIM ↑	RMSE ↓
Deep3DFace [22]	22.491	0.803	21.184
MGCNet [43]	22.787	0.815	20.095
Tran et al. [38]	19.038	0.772	29.335
Chen et al. [25](Coarse)	23.821	0.840	17.427
Chen et al. [25](Detail)	24.055	0.841	16.848
DECA [23](Coarse)	20.382	0.767	26.303
DECA [23](Detail)	20.391	0.764	26.278
EMOCA [24](Coarse)	19.993	0.755	27.349
EMOCA [24](Detail)	20.022	0.752	27.269
Ours(Coarse)	24.263	0.846	16.639
Ours(Fine-grained)	**24.486**	**0.854**	**16.145**

challenging face images such as occlusions, large poses and different lighting conditions. However, our model still implemented state-of-the-art evaluation results compared to other methods. The PSNR obtained by our method is 7.88%, 6.48%, 27.45%, 1.86%, 10.04% and 21.36% higher than Deep3DFace [22], MGCNet [43], Chen et al. [25] (coarse), DECA [23] (coarse) and EMOCA [24] (coarse), respectively. The SSIM calculated by our method is 5.35%, 3.80%, 9.58%, 0.71%, 10.30% and 12.05% higher than the state-of-the-art methods, respectively. And our method achieves 21.45%, 17.20%, 43.28%, 4.52%, 36.74% and 39.16% lower RMSE than Deep3DFace [22], MGCNet [43], Chen et al. [25] (coarse), DECA [23] (coarse) and EMOCA [24] (coarse), respectively. Furthermore, we achieve quantitative evaluation on the LFW [50] dataset as depicted in Table 3. The results outperform the evaluation results on the test dataset of CelebA [49] and the LS3D [48] dataset, due to the LFW [50] dataset contains less challenging face images. Compared with other methods, our method obtained the highest PSNR and SSIM, and achieved the lowest RMSE. This further demonstrates that our coarse 3D face reconstruction method based on the self-supervised learning is robust and stable for in-the-wild face images.

Fig. 8. Comparison results of the fine-grained 3D face reconstruction methods on the test dataset of CelebA [49], and the LS3D [48] and LFW [50] datasets.

4.4 Evaluation on fine-grained reconstruction

4.4.1 Qualitative evaluation

For the detailed 3D face reconstruction, we qualitatively compare our method with the existing state-of-the-art fine-grained 3D face reconstruction methods, namely Tran et al. [38], Facescape [18], Chen et al. [25], DECA [23] and EMOCA [24] on the test dataset of CelebA [49], and the LS3D [48] and LFW [50] datasets.

Fig. 9. Comparison results of the fine-grained 3D face reconstruction methods on the test dataset of CelebA [49], and the LS3D [48] and LFW [50] datasets.

We first compare our 3D face reconstruction results for geometry with Facescape [18], Chen et al. [25], DECA [23] and EMOCA [24] on the test dataset of CelebA [49], and the LS3D [48] and LFW [50] datasets in Fig. 8. Unlike other methods, Facescape [18] is trained by using detailed 3D face scans as ground-truth. Therefore, the faces reconstructed by it from in-the-wild face images are prone to distortion. The face shapes generated by DECA [23] and EMOCA [24] is very similar. Although they are able to reconstruct a complete head structure with slight details, these reconstructed faces lost a lot of structural details. Chen et al. [25] can reconstruct high-quality faces with rich details, but it will generate a lot of noise such as scratches. Compared to these detailed 3D face reconstruction methods, our method is able to generate fine-grained faces that are more vivid and easily recognized by human eyes.

In Fig. 9, we show the rendering results of Tran et al. [38], Chen et al. [25], DECA [23], EMOCA [24] and our method for the reconstructed 3D faces on the test dataset of CelebA [49], and the LS3D [48] and LFW [50] datasets. The faces reconstructed by DECA [23] and EMOCA [24] are similar to puppets, since they cannot accurately generate facial expression and reconstructed mouth area is distorted (as shown in rows 1, 5 and 6). The textures generated by DECA and EMOCA are not realistic, since the albedo maps they generated do not contain texture details. Although Tran et al. [38] can accurately reconstruct the face images with wrinkles, the skin color generated by it is obviously distorted compared with the input images. Chen et al. [25] has the ability to reconstruct high-fidelity 3D faces with high-realistic texture, but it is not robust and prone to generate a lot of noise. Overall, our method is able to generate face images that closely approximate the input images, and can restore rich details that vary with expression, such as forehead wrinkles, nasolabial folds, details of eyes and mouth, etc.

Fig. 10. Comparison results of the fine-grained 3D face reconstruction methods for normal and shading on the test dataset of CelebA [49], and the LS3D [48] and LFW [50] datasets.

Furthermore, we also represent the reconstruction results of normal and lighting from Tran et al. [38], Chen et al. [25], DECA [23], EMOCA [24] and our method on the test dataset of CelebA [49], and the LS3D [48] and LFW [50] datasets. In Fig. 10, we can see that both DECA [23] and EMOCA [24] cannot accurately capture lighting conditions, and their reconstructed shading images are too dark. Furthermore, the normal maps generated by DECA [23] and EMOCA [24] methods lose too much geometric details, so they are difficult to reconstruct personalized 3D faces with high-fidelity as shown in Fig. 8 and Fig. 9. In contrast to DECA [23] and EMOCA [24], the shading images generated by Tran et al. [38] are too bright (The result of normal map is not open implemented.). Compared with Trans et al. [38], DECA [23] and EMOCA [24], Chen et al. [25] can more accurately capture ambient light and reconstruct more detailed face structure. But it generates a lot of noise such as scratches, which seriously degrades the quality of reconstructed 3D faces as depicted in Fig. 8 and Fig. 9. Compared with these state-of-the-art fine-grained 3D face reconstruction methods, our method can reconstruct better results of lighting and normal maps from in-the-wild facial images under extremely challenging lighting conditions. Moreover, our method is able to significantly recover better expression and wrinkles, especially around the chin, mouth and eyes, where the details are incorrectly reconstructed by other methods.

4.4.2 Quantitative evaluation

To verify the robustness and stability of our fine-grained 3D face reconstruction model, we achieve quantitative evaluation of Deep3DFace [22], MGCNet [43], Tran et al. [38], Chen et al. [25] (detail), DECA [23] (detail), EMOCA [24] (detail) and our method on the test dataset of CelebA [49], and the LS3D [48] and LFW [50] datasets on in-the-wild images.

We firstly display the evaluation results on the test dataset of CelebA [49] in Table 1. Our coarse 3D face reconstruction model achieves significant results than other methods in terms of PSNR, SSIM and RMSE. However, our fine-grained 3D face reconstruction model is better than the coarse reconstruction model, because it can reconstruct more realistic 3D faces. The PSNR and SSIM obtained by our fine-grained 3D face reconstruction model are 0.73% and 1.27% higher than those of the coarse reconstruction model, respectively. Furthermore, the RMSE calculated by our fine-grained 3D face reconstruction model is the lowest, which is 2.74% less than that of the coarse reconstruction model. We also represent the evaluation results on the LS3D [48] dataset in Table 2. Our method is superior to other state-of-the-art methods. Our method calculates 8.87%, 7.46%, 28.62%, 1.79%, 20.08% and 22.30% higher PSNR than Deep3DFace [22], MGCNet [43], Chen et al. [25], DECA [23] and EMOCA [24], respectively. The SSIM obtained by our method is 6.35%, 4.78%, 10.62%, 1.55%, 11.78% and 13.56% higher than the state-of-the-art methods, respectively. And our method achieves 23.78%, 19.66%, 44.96%, 4.17%, 38.56% and

Table 3: Quantitative evaluation for the 3D face reconstruction methods on the LFW [50] dataset.

Methods	PSNR ↑	SSIM ↑	RMSE ↓
Deep3DFace [22]	25.428	0.835	14.518
MGCNet [43]	25.160	0.847	14.816
Tran et al. [38]	19.723	0.786	26.623
Chen et al. [25](Coarse)	26.826	0.870	11.965
Chen et al. [25](Detail)	26.960	0.874	11.715
DECA [23](Coarse)	23.635	0.809	17.657
DECA [23](Detail)	23.662	0.806	17.606
EMOCA [24](Coarse)	22.969	0.794	19.000
EMOCA [24](Detail)	23.043	0.793	18.849
Ours(Coarse)	27.091	0.871	11.616
Ours(Fine-grained)	**27.301**	**0.882**	**11.285**

40.79% lower RMSE than other comparative algorithms, respectively. Furthermore, we provide the reconstruction results on the LFW [50] dataset. As demonstrated in Table 3, we can clearly observe that our method computes higher PSRN and SSIM, and lower RMSE than those on the CelebA [49] and LS3D [48] datasets. Because, the LFW [50] dataset contains fewer face images under challenging conditions. However, our method still outperforms other comparative methods. This fully demonstrates that our method can robustly and stably reconstruct fine-grained 3D faces from single face images on in-the-wild by using self-supervised learning.

Table 4: Reconstruction errors on the NoW [41] benchmark.

Methods	Median (mm)	Mean (mm)	Std (mm)
Facescape [18]	3.92	11.04	16.80
3DDFA_V2 [42]	3.60	10.36	16.25
RingNet [41]	1.36	1.71	1.44
Deep3DFace [22]	1.29	1.87	2.38
MGCNet [43]	1.43	1.99	2.42
Tran et al. [38]	1.48	2.07	2.47
Chen et al. [25]	1.21	1.53	1.35
DECA [23]	1.22	1.55	1.34
EMOCA [24]	1.26	1.55	1.31
Ours	**1.19**	**1.50**	**1.30**

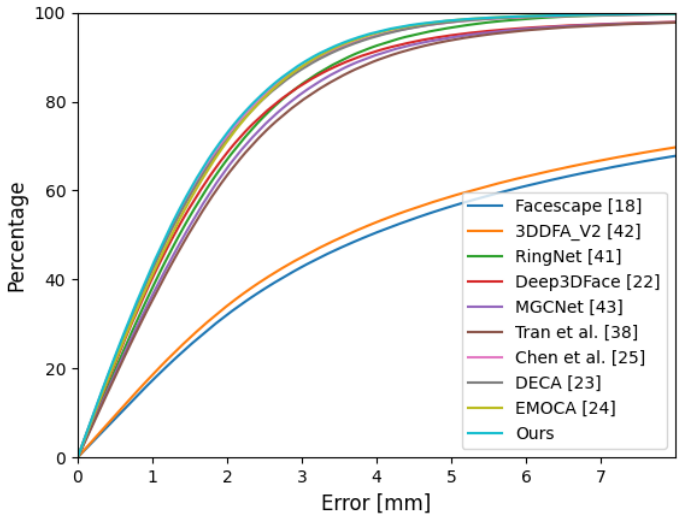

Fig. 11. Cumulative errors on the NoW [41] benchmark.

To further validate the performance of our proposed approach, we quantitatively compare with these state-of-the-art reconstruction methods including Facescape [18], 3DDFA_V2 [42], RingNet [41], Deep3DFace [22], MGCNet [43], Tran et al. [38], Chen et al. [25], DECA [23], and EMOCA [24] on the NoW [41] benchmark. For a fair

comparison to these methods, we first use the method [22] to crop the face images in the NoW validation set to a size of 256 × 256, and further perform a rigid alignment based on the scan-to-mesh distance between the ground truth scan, and the reconstructed 3D mesh. As shown in Table 4 and the cumulative error plot in Fig. 11, our approach achieves state-of-the-art results on the NoW benchmark, providing the lowest reconstruction errors for the mean, median, and standard deviation.

4.5 Ablation study

To evaluation the effectiveness of our fine-grained 3D face reconstruction method, we achieve ablation experiments for the wavelet transform perception model, the depth displacement module and the albedo module. As shown in Table 5, the 3D face reconstruction model is trained without the wavelet transform perception model, the depth displacement module and the albedo module is greatly degrading the quality of reconstructed 3D faces. Our model significantly improves the ability of reconstructing high-fidelity 3D faces with high-frequency details by jointly combining them. As expected, combing these full modules, our method obtains the best results, and can reconstruct more accurate and realistic 3D face from monocular face image in-the-wild.

Table 5: Ablation experiments of 3D face reconstruction model.

Method			**PSNR** ↑	**SSIM** ↑	**RMSE** ↓
Wavelet Transform Perception Model	Depth Displacement Module	Albedo Module			
×	×	×	25.794	0.863	13.705
×	✓	✓	25.805	0.869	13.606
✓	×	✓	25.827	0.870	13.573
✓	✓	✓	**25.982**	**0.874**	**13.329**

Fig. 12 shows the visual results of the ablation experiments. Specifically, (a) represents the input image; (b) indicates the reconstructed results by removing the wavelet transform perception model, depth displacement module, and the albedo module; (c) shows the reconstructed results without the wavelet transform perception model; (d) indicates the reconstructed results without the depth displacement module; (e) represents the reconstructed results with the full models. As shown in the enlarged views of Fig. 12 (b), the generated texture and normal images are smooth and unrealistic due to the limitation of the 3DMM model. When removing the wavelet transform perception model or depth displacement module, the reconstructed 3D face loses some details, and the generated expressions are inconsistent with the input image, as shown in Fig. 12 (c) and Fig. 12 (d). With the full models, we can clearly observe that our approach can reconstruct a high-fidelity 3D face with rich details, vivid expression, and high-realistic textures, as shown in Fig. 12 (e).

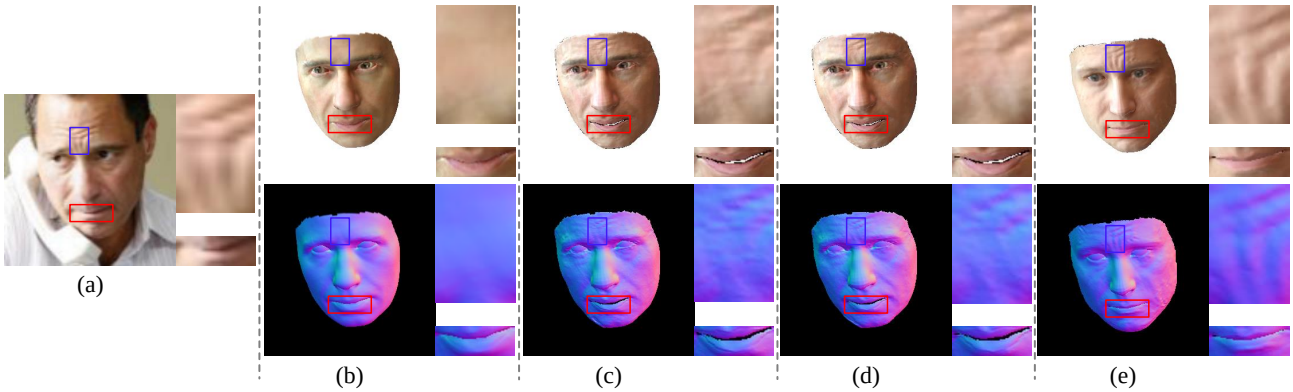

Fig. 12. The comparative and ablation results of the 3D face reconstruction model. Left: Input image. Top: Reconstructed texture images. Bottom: Reconstructed normal images.

5 Conclusion

In this paper, we propose a fine-grained 3D face reconstruction framework with a novel coarse-to-fine scheme which recovers detailed facial geometries and textures from monocular images in-the-wild by using self-supervised

learning. Firstly, we use a face regression network R-Net to regress 3DMM parameters and reconstruct coarse 3D faces by using prior 3DMM model in the coarse pipeline. A novel fine pipeline is designed that consists of a wavelet transform perception model, the albedo module and the depth displacement module, which further generates fine-grained 3D faces with expression-dependent and wrinkles. Extensive experiments show that our method has the ability to reconstruct high-fidelity face geometries and textures with rich details, and demonstrates a significantly improved results compared with the state-of-the-art reconstruction methods on different in-the-wild datasets, both in terms of qualitative and quantitative evaluations. In the future, we want to learn a dynamic-based face model that can reconstruct animated 3D face with fine details vary with expression from single images or video.

Acknowledgements
This work was supported by the Shanghai Talent Development Funding of China (Grant No. 2021016) and the Science and Technology Projects of National Archives Administration of China (Grant No. 2023-X-036).

Authors contributions
Dongjin Huang: wrote the main manuscript and provided funding support. Yongsheng Shi: conducted experiments and the writing of the main manuscript. Jinhua Liu: contributed to conceptualization and methodology. Wen Tang: provided guidance on paper writing and supervision.

Data availability
The data are available from the corresponding author on reasonable request.

Declarations
Conflict of interest The authors declare no competing interests.

References

1. Vetter, T., Blanz, V.: Estimating coloured 3d face models from single images: An example based approach. In: Proceedings of European Conference on Computer Vision (ECCV), pp. 499–513 (1998). https://doi.org/10.1007/BFb0054761
2. Xu, R., Wang, K., Deng, C., et al.: Depth map denoising network and lightweight fusion network for enhanced 3d face recognition. Pattern Recognition **145**, 109936 (2024). https://doi.org/10.1016/j.patcog.2023.109936
3. Shahreza, H.O., Marcel, S.: Comprehensive vulnerability evaluation of face recognition systems to template inversion attacks via 3d face reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(12), 14248–14265 (2023)
4. Huang, Z., Chan, K.C., Jiang, Y., Liu, Z.: Collaborative diffusion for multi-modal face generation and editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6080–6090 (2023)
5. Zhang, L., Qiu, Q., Lin, H., et al.: Dreamface: Progressive generation of animatable 3d faces under text guidance. ACM Transactions on Graphics **42**(4), 1–16 (2023)
6. Zhang, W., Cun, X., Wang, X., et al.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8652–8661 (2023)
7. La Cava, S.M., Orrù, G., Drahansky, M., et al.: 3d face reconstruction: the road to forensics. ACM Computing Surveys **56**(3), 1–38 (2023)
8. Han, Y., Wang, Z., Xu, F.: Learning a 3d morphable face reflectance model from low-cost data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8598–8608 (2023)
9. Bai, H., Kang, D., Zhang, H., Pan, J., Bao, L.: Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 362–371 (2023)
10. Tan, F., Fanello, S., Meka, A., et al.: Volux-gan: A generative model for 3d face synthesis with hdri relighting. In: Proceedings of ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques (SIGGRAPH), pp. 1–9 (2022)
11. Liang, J., Liu, Y., Lu, F.: Reconstructing 3d virtual face with eye gaze from a single image. In: Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (IEEE VR), pp. 370–378 (2022)
12. Li, Y., Hao, Q., Hu, J., Pan, X., Li, Z., Cui, Z.: 3d3m: 3d modulated morphable model for monocular face reconstruction. IEEE Transactions on Multimedia **25**, 6642–6652 (2023)
13. Shang, J., Zeng, Y., Qiao, X., Wang, X., Zhang, R., Sun, G., Patel, V., Fu, H.: Jr2net: joint monocular 3d face reconstruction and reenactment. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2200–2208 (2023)

14. Kumar, R., Luo, J., Pang, A., Davis, J.: Disjoint pose and shape for 3d face reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3115–3125 (2023)
15. Yan, X., Yu, Z., Ni, B., et al.: Cross-species 3d face morphing via alignment-aware controller. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 3018–3026 (2022)
16. Xiao, Y., Zhu, H., Yang, H., et al.: Detailed facial geometry recovery from multi-view images by learning an implicit function. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 2839–2847 (2022)
17. Deng, Q., Ma, L., Jin, A., et al.: Plausible 3d face wrinkle generation using variational autoencoders. IEEE Trans. Vis. Comput. Graph. **28**(9), 3113–3125 (2021)
18. Yang, H., Zhu, H., Wang, Y., et al.: Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 601–610 (2020)
19. Zhu, X., Yu, C., Huang, D., et al.: Beyond 3dmm: Learning to capture high-fidelity 3d face shape. IEEE Trans. Pattern Anal. Mach. Intell. **45**(2), 1442–1457 (2022)
20. Zeng, X., Wu, Z., Peng, X., et al.: Joint 3d facial shape reconstruction and texture completion from a single image. Comput. Vis. Media **8**(2), 239–256 (2022)
21. Wang, L., Chen, Z., Yu, T., et al.: Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20333–20342 (2022)
22. Deng, Y., Yang, J., Xu, S., et al.: 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 (CVPRW), pp. 285–295 (2019)
23. Feng, Y., Feng, H., Black, M., et al.: Learning an animatable detailed 3d face model from in-the-wild images. ACM Trans. Graph. **40**(4), 1–13 (2021). https://doi.org/10.1145/3450626.3459936
24. Danecek, R., Black, M., Bolkart, T.: Emoca: Emotion driven monocular face capture and animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 20311–20322 (2022)
25. Chen, Y., Wu, F., Wang, Z., et al.: Self-supervised learning of detailed 3d face reconstruction. IEEE Trans. Image Process. **29**, 8696–8705 (2020)
26. Yang, M., Guo, J., Cheng, Z., et al.: Self-supervised high-fidelity and re-renderable 3d facial reconstruction from a single image (2021). ArXiv preprint arXiv:2111.08282
27. Egger, B., Smith, W., Tewari, A., et al.: 3d morphable face models—past, present, and future. ACM Trans. Graph. **39**(5), 1–38 (2020)
28. Guo, Y., Cai, L., Zhang, J.: 3d face from x: Learning face shape from diverse sources. IEEE Trans. Image Process. **30**, 3815–3827 (2021)
29. Ruan, Z., Zou, C., Wu, L., et al.: Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction. IEEE Trans. Image Process. **30**, 5793–5806 (2021)
30. Zhu, X., Lei, Z., Liu, X., et al.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 146–155 (2016)
31. Bagdanov, A., Bimbo, A.D., Masi, I.: The florence 2d/3d hybrid face dataset. In: Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding, pp. 79–80 (2011)
32. Guo, L., Zhu, H., Lu, Y., et al.: Rafare: Learning robust and accurate non-parametric 3d face reconstruction from pseudo 2d&3d pairs. In: the AAAI Conference on Artificial Intelligence, pp. 719–727 (2023)
33. Kang, J., Lee, S., Lee, S.: Competitive learning of facial fitting and synthesis using uv energy. IEEE Trans. Syst. Man Cybern. Syst. **52**(5), 2858–2873 (2021)
34. Yin, X., Huang, D., Fu, Z., et al.: Weakly-supervised photo-realistic texture generation for 3d face reconstruction. In: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), pp. 1–8. IEEE (2023)
35. Fan, X., Cheng, S., Huyan, K., et al.: Dual neural networks coupling data regression with explicit priors for monocular 3d face reconstruction. IEEE Trans. Multimedia **23**, 1252–1263 (2020)
36. Gao, Z., Zhang, J., Guo, Y., et al.: Semi-supervised 3d face representation learning from unconstrained photo collections. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops(CVPRW), pp. 348–349 (2020)
37. Kao, Y., Pan, B., Xu, M., Lyu, J., Zhu, X., Chang, Y., Li, X., Lei, Z.: Toward 3d face reconstruction in perspective projection: Estimating 6dof face pose from monocular image. IEEE Transactions on Image Processing **32**, 3080–3091 (2023)
38. Tran, L., Liu, X.: On learning 3d face morphable model from in-the-wild images. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 157–171 (2019)
39. Yao, S., Zhong, R., Yan, Y., et al.: Dfa-nerf: Personalized talking head generation via disentangled face attributes neural rendering (2022). ArXiv preprint arXiv:00791

40. Gafni, G., Thies, J., Zollhofer, M., et al.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8649–8658 (2021)
41. Sanyal, S., Bolkart, T., Feng, H., et al.: 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 (CVPR), pp. 7763–7772 (2019)
42. Guo, J., Zhu, X., Yang, Y., et al.: Towards fast, accurate and stable 3d dense face alignment. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 152–168 (2020)
43. Shang, J., Shen, T., Li, S., et al.: Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 53–70 (2020)
44. Parkhi, O., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proceedings of the British Machine Vision Conference (BMVC), pp. 41.1–41.12 (2015)
45. Paysan, P., Knothe, R., Amberg, B., et al.: A 3d face model for pose and illumination invariant face recognition. In: IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS), pp. 296–301 (2009)
46. Cao, C., Weng, Y., Zhou, S., et al.: Facewarehouse: A 3d facial expression database for visual computing. IEEE Trans. Vis. Comput. Graph. **20**(3), 413–425 (2014)
47. Ramamoorthi, R., Hanrahan, P.: An efficient representation for irradiance environment maps. In: Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), pp. 497–500 (2001)
48. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1021–1030 (2017)
49. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision (CVPR), pp. 3730–3738 (2015)
50. Huang, G., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., University of Massachusetts, Amherst (2008). https://inria.hal.science/inria-00321923