Computer Science and Application 计算机科学与应用, 2024, 14(4), 255-267
Published Online April 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.144095

Hans汉斯

由粗到精的高保真单目三维人脸重建

景圣恩，高　添，陶应诚，彭梦昊，侍亚东，王　冠

合肥工业大学计算机与信息学院，安徽　合肥

收稿日期：2024年3月20日；录用日期：2024年4月18日；发布日期：2024年4月26日

摘　要

针对现有单目三维人脸重建方法在细节刻画和身份信息保持方面的不足，本文提出了一种由粗到精的三维人脸重建框架。该框架首先利用从二维人脸图片中提取的特征参数生成初始三维人脸模型，并设计多尺度身份特征提取器捕获个性化特征。然后，通过自适应加权策略筛选对重建任务最具贡献的特征信息。在精细重建阶段，本文关注人脸的几何细节重建，将身份和表情编码融入几何细节生成网络中，以生成具有特定身份和表情信息的几何细节。最后，利用可微分渲染器将三维人脸模型渲染为二维人脸图像，进行自监督训练。在CelebA和AFLW2000-3D数据集上的实验结果表明，本文提出的框架能够从单幅图像中重建出更加真实、自然且具有高度个性化特征的三维人脸模型，在细节刻画和身份信息保持方面均优于现有方法，具有广阔的应用前景。

关键词

三维人脸重建，三维形变模型，自监督学习，人脸渲染

Coarse-to-Fine Monocular 3D Face Reconstruction with High Fidelity

Sheng'en Jing, Tian Gao, Yingcheng Tao, Menghao Peng, Yadong Shi, Guan Wang

School of Computer Science and Information Engineering, Hefei University of Technology, Hefei Anhui

Received: Mar. 20th, 2024; accepted: Apr. 18th, 2024; published: Apr. 26th, 2024

Abstract

Addressing the limitations of existing monocular 3D face reconstruction methods in capturing fine details and preserving identity information, this paper proposes a coarse-to-fine framework for 3D face reconstruction. The framework initially generates a basic 3D face model using feature parameters extracted from a 2D facial image and employs a multi-scale identity feature extractor to capture personalized characteristics. Subsequently, an adaptive weighting strategy is utilized to

文章引用: 景圣恩, 高添, 陶应诚, 彭梦昊, 侍亚东, 王冠. 由粗到精的高保真单目三维人脸重建[J]. 计算机科学与应用, 2024, 14(4): 255-267. DOI: 10.12677/csa.2024.144095

select the most relevant features for the reconstruction task. In the fine reconstruction phase, the focus is on geometric detail reconstruction, integrating identity and expression encodings into a geometric detail generation network to produce detailed geometry specific to the individual's identity and expressions. Finally, a differentiable renderer is employed to convert the 3D face model into a 2D facial image for self-supervised training. Experimental results on the CelebA and AFLW2000-3D datasets demonstrate that the proposed framework can reconstruct more realistic, natural, and highly personalized 3D face models from a single image, outperforming existing methods in terms of detail capture and identity preservation, thus holding promising potential for various applications.

Keywords

3D Face Reconstruction, 3D Morphable Model, Self-Supervised Learning, Face Rendering

1. 引言

近年来，三维人脸重建技术已成为计算机视觉领域的研究热点，其广泛应用于虚拟角色创建[1]、人脸识别系统[2]、情感表达迁移等[3]领域。传统的三维人脸重建方法依赖于结构光相机[4]、三维扫描仪[5]等硬件设备，虽然能够提供高精度的人脸数据，但其高昂的成本和复杂的操作流程限制了在实际生产生活中的应用。为了降低三维人脸重建的门槛，基于数字图像的三维人脸重建技术应运而生。这类技术仅需二维人脸图像作为输入，通过计算机视觉和图形学的方法即可重建出人脸的三维模型，极大地降低了获取数据的难度和成本。其中，单目图像重建技术仅需一张人脸图像即可实现三维重建，无需额外的设备和复杂的操作流程，在实际应用中具有更广阔的前景。

由于单幅图像在深度信息上的缺失，单目图像的三维人脸重建一直是计算机视觉领域中的一个难题。多数现有方法倾向于采用三维形变模型(3D Morphable Model, 3DMM) [6] [7] [8]作为人脸几何形态与外观特性的先验知识。这类方法基于大规模三维人脸数据的统计分析，构建了一套涵盖形状、姿态、表情及纹理等多个主成分的基底。通过这些主成分的线性加权组合，理论上能够重构出完整且逼真的人脸模型。然而，这种方法对于细节信息的捕捉能力严重不足。尤其是皱纹、痣等细微特征。此外，重建后的三维人脸模型在保持原始年龄、性别等身份信息方面也存在较大挑战。

为了解决这一难题，近期的研究工作开始尝试通过生成位移图[9] [10] [11] [12] [13]来精细化人脸模型的细节信息。位移图的应用在一定程度上提升了模型对细节的刻画能力，但这些方法在处理人脸静态与动态特征时仍显得不够成熟。特别是在面对表情幅度较大的人脸图像时，由于缺乏有效的特征区分机制，重建出的模型在关键区域(如眉头、嘴角等)往往无法准确呈现出相应的皱纹和细节特征。这导致重建后的人脸在视觉效果上显得不够自然和真实。

针对上述问题，本文提出了一种由粗到精的三维人脸重建框架，旨在从单幅图像中重建出更加精细和真实的人脸模型。在粗人脸重建阶段，本文利用从二维人脸图片中提取的 3DMM 参数生成一个初始的三维人脸模型，并设计了一个多尺度身份特征提取器来捕获与人脸身份相关的个性化特征。通过自适应加权策略，网络能够筛选出对人脸重建任务最具贡献的特征信息。在精细重建阶段，本文进一步关注人脸的几何细节的重建。本文观察到人脸的身份特征和表情特征分别控制着面部的静态和动态细节，将粗

人脸重建阶段提取的身份和表情编码融入几何细节生成网络中，以生成具有特定身份和表情信息的几何细节。两个阶段的三维人脸模型都由可微渲染器渲染为二维人脸图像，分别与输入图片进行自监督训练。通过结合多尺度身份特征提取以及精细的几何和纹理细节重建技术，本文提出的框架能够从单幅图像中重建出更加真实且具有高度个性化特征的三维人脸模型。

2. 相关工作

2.1. 三维形变模型

1999 年 Blanz 等人[6]开创性地提出了三维形变模型，将三维人脸表示为一组形状和纹理基向量的线性加权组合。这组基向量通过扫描 200 个人脸的三维数据后进行主成分分析得到。Cao 等人[7]提出的 Facewarehouse 模型增加了表情基向量，使人脸模型能够表现出同一个人不同表情的差异。Li 等人[8]进一步提出 FLAME 模型，以形状、姿态和表情作为基向量，能够更加准确地建模头部姿态和面部表情。

早期的方法[14] [15] [16]通常首先初始化三维人脸形状和纹理的基系数，以不断迭代拟合的方式获得 3DMM 参数，具有较高的时间复杂度，且对初始化要求较高，容易陷入局部最优解。随着深度学习技术的发展，一些方法[17] [18] [19] [20] [21]开始使用卷积神经网络对二维图片进行特征提取，回归出表现较好的 3DMM 参数。Ayush 等人[19]采用基于 CNN 的编码器从单幅图片中提取姿态、形状、表情等参数，将 3DMM 模型作为解码器构建三维人脸模型。网络以无监督的方式进行端到端训练，重建出了质量较高的三维人脸。Fan 等人[20]提出一个双神经网络，回归出更具有鲁棒性和个性化的 3DMM 参数，进一步提升了三维人脸的重建效果。Zhu 等人[21]通过 CNN 不仅回归出 3DMM 参数，还得到了人脸局部不同尺度的特征，然后将这些特征和三维人脸模型进行可微渲染得到的二维图像，与输入人脸图像进行误差计算。这种方法能够使投影的脸型更好地与输入人脸图像上每个面部区域的轮廓对齐。

2.2. 人脸几何重建

由于 3DMM 模型本质是线性正交基的低维表达，重建的三维人脸仅仅是粗糙的形状和纹理，面部的皱纹等细节很难重建出来，并且不同个体间的差异也很难表现出来。很多研究者都提出了方法来克服这一局限性。Luan 等人[22]提出了一个非线性的 3DMM 模型，该模型分别使用两个深度神经网络代替线性 3DMM 模型，将形状和纹理参数解码为三维人脸形状和纹理，并设计了可微渲染层来进行端到端的弱监督训练。这种方法与线性 3DMM 相比，明显改善了三维人脸形状和纹理的表示能力。Lee 等人[23]采用 GCN 和 GAN 来分别重建三维人脸的几何网格和面部纹理。为了重建出人脸的几何细节，Matan 等人[24]提出了一个图像到图像网络，从输入的二维图像中生成基于像素的人脸几何表示。一些方法通过生成位移图来表示人脸的面部的几何细节。Yang 等人[11]通过 pix2pixHD 网络[25]预测出不同表情的位移图，并通过权重相加来获得面部的动态细节。

3. 模型框架

图 1 展示了本文提出的框架结构，它由两个阶段组成，红色部分表示粗糙人脸重建阶段，绿色部分表示精细化人脸重建。在粗糙人脸重建阶段，特征提取网络从输入的二维图片中回归出 3DMM 参数，进而构建出三维人脸低频的几何和纹理。在精细化三维人脸重建阶段，人脸几何精细化模块通过图像到图像的生成网络，根据输入图像映射到 UV 空间的纹理图生成人脸的高频几何细节。

3.1. 粗糙人脸重建阶段

粗糙三维人脸重建阶段的目标是对图像中的人脸的低频几何和纹理进行学习和重建。具体来说，给

定一幅二维图像 I 作为输入，编码器回归出一个 468 维的参数空间。该尺寸空间包括面部身份参数 $\beta \in R^{200}$、姿态参数 $\theta \in R^{6}$、表情参数 $\varepsilon \in R^{50}$、相机角度特征参数 $c \in R^{3}$、反射率参数 $\alpha \in R^{200}$ 和照明参数 $\gamma \in R^{9}$。

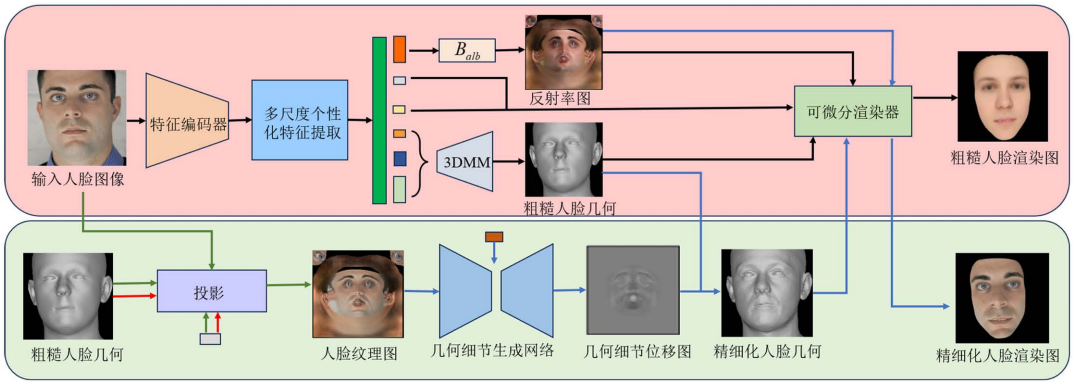

Figure 1. Coarse-to-fine 3D face reconstruction framework
图 1. 由粗到精的三维人脸重建框架

为了能够使编码器提取到图像中人脸区域的关键特征以及与人脸身份相关的个性化特征，本文提出了一个多尺度个性化特征提取模块，从编码特征中有效提取人脸的年龄、性别等个性化信息。图 2 显示了多尺度个性化特征提取模块的结构，该模块分为两个部分。在第一部分中，模块将编码器提取的特征分别通过四种不同尺度的卷积层进行处理。其中 1 × 1 卷积主要负责保留人脸的全局特征，而其余三个尺度的卷积则专注于捕获人脸的局部特征。通过对这些不同尺度的特征进行采样和连接，本文能够获得一个既包含局部细节又兼顾全局信息的密集特征表示。在模块的第二部分中，本文引入一种注意力机制，旨在专注于提取有效信息并有效抑制背景特征的干扰。该注意力机制分为通道注意力和空间注意力两个并行分支。通道注意力分支用于获取不同通道间的特征关系，通过全局平均池化层对每个通道的全局信息进行编码，并利用带有隐藏层的多层感知机来精准估计各通道间的注意力权重。空间注意力用于强调或抑制在不同空间位置的特征，通过两个 1 × 1 卷积来整合和压缩通道维度，再结合两个 3 × 3 空洞卷积来有效聚合具备更大感受野的上下文信息。随后，这两个分支的输出特征经过广播机制进行维度扩展生成注意力映射。最后，本文将生成的注意力映射与原始输入特征进行逐元素相乘，以获得更为精准的人脸个性化特征。通过这种方式，注意力机制能够利用提取的特征自适应地学习权重分布，并将这些学习到的权重与人脸细节轮廓紧密结合，进而强化编码器对关键特征的感知与提取能力。

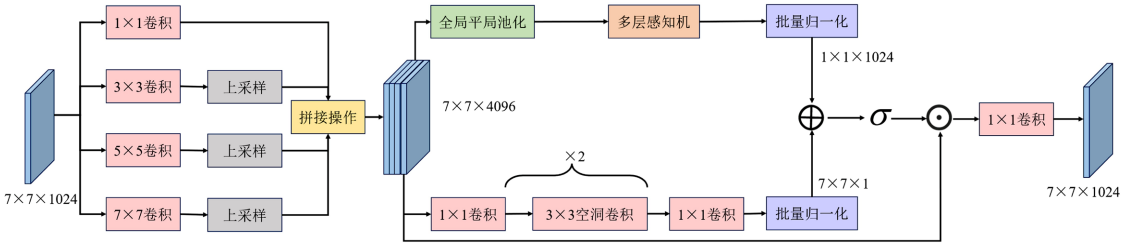

Figure 2. Multi-scale personalized feature extraction module structure
图 2. 多尺度个性化特征提取模块结构

本文参考了 FLAME 模型[8]构建三维人脸的方式，根据面部身份参数 β、表情参数 ε 和姿态参数 θ 构建表示粗糙人脸几何形状 S_{coarse} 与反射率 UV 图 A：

$$S_{coarse}=(\beta,\theta,\varepsilon)=\overline{S}+\beta B_{id}+\varepsilon B_{exp}+\theta B_{pos} \tag{1}$$

$$A(\alpha)=\overline{A}+\alpha B_{alb} \tag{2}$$

其中 $\overline{S}\in R^{3n}$ 表示平均人脸基底，B_{id}、B_{pos}、B_{exp} 和 B_{alb} 分别表示人脸模型的身份、姿态、表情和反射率基底。本文假设人脸为朗伯斯表面，用球面谐波函数来模拟场景光照。假定人脸表面一顶点 p_i，其表面法线和皮肤纹理分别表示为 n_i 和 t_i，该点的光照可以由如下公式计算：

$$C(n_i,t_i|\gamma)=t_i\cdot\sum_{b=1}^{B^2}\gamma_b\Phi_b(n_i) \tag{3}$$

其中 Φ_b 表示球面谐波基函数，γ_b 表示对应的球面谐波参数。参考[26]，B 设置为 3。可微渲染器通过粗糙三维人脸形状、相机参数、光照参数和反射率 UV 图生成二维图像 $I_{r,coarse}$：

$$I_{r,coarse}=\mathcal{R}(S_{coarse},A,\gamma) \tag{4}$$

3.2. 精细化人脸重建阶段

在粗糙人脸重建阶段，三维形变模型的顶点数量有限，无法充分重建出脸部表面的细节信息，如皱纹等。为解决这一问题，本文采用 UV 空间的位移深度图来精细面部人脸的细节几何。位移深度图能够表达沿法线方向的几何变形，打破了基础模型顶点密度的限制。通过将位移深度图逐像素转换为绘制过程中的细节法线，人脸网格能够显示所有细微的细节变化。基于此，本文提出了一个人脸几何细节生成网络。该网络将人脸纹理 UV 图作为输入，通过生成器挖掘纹理图中难以捕捉的细节信息，并将这些细节作为特征图，循环合成位移深度图 $D\in[-0.01,0.01]^{256\times256}$。最后，网络将位移深度图在 UV 域对人脸网格顶点进行叠加，生成具有精细化特征的人脸网格模型。

3.2.1. 投影和纹理采样

首先，本文利用粗糙人脸生成阶段产生三维人脸模型对输入图像 I 和进行投影和纹理采样，使其映射为更适合深度卷积网络处理的 UV 纹理图 T。具体来说，对于每一个三维顶点 p_i，本文使用弱透视模型获得其在输入图像 I 上的对应点 q_i：

$$p_i=f\begin{bmatrix}1&0&0\\0&1&0\end{bmatrix}(Rq_i)+t \tag{5}$$

其中 f 为物体在相机坐标系下和图像坐标系下的转换比例因子，R 为旋转矩阵，t 为偏移向量。然后本文采用双线性采样[27]的方式从输入图像上采集纹理值到三维顶点，得到三维纹理。对于人脸中的遮挡区域，本文使用 z-buffer 算法[28]将不可见区域的像素值设为 0。最后将三维纹理展开到 UV 空间。对于像素值为 0 的区域，本文假设人脸的对称性一致性，通过将输入图像和人脸模型翻转的方式来填补由于遮挡不可见的像素值。

3.2.2. 几何细节生成网络

本文提出的的几何细节生成网络是基于 UNet 网络[29]实现的编码器－解码器结构，旨在捕捉多尺度的特征信息，图 3 显示了几何细节生成网络的结构。对于同一个体的多张人脸图像，可以观察到一些静态细节(如皱纹、眉毛形态等)展现出较强的一致性，这些特征被视为与身份紧密相关且稳定的表征。与此同时，面部的动态细节，如皱眉、张嘴等导致的面部收缩和扩展，与表情有很大关系。鉴于此，本文在生成网络中引入了一种交叉注意力模块，旨在对多尺度特征图进行精细化评估，并筛选出对重建面部细节至关重要的特征。

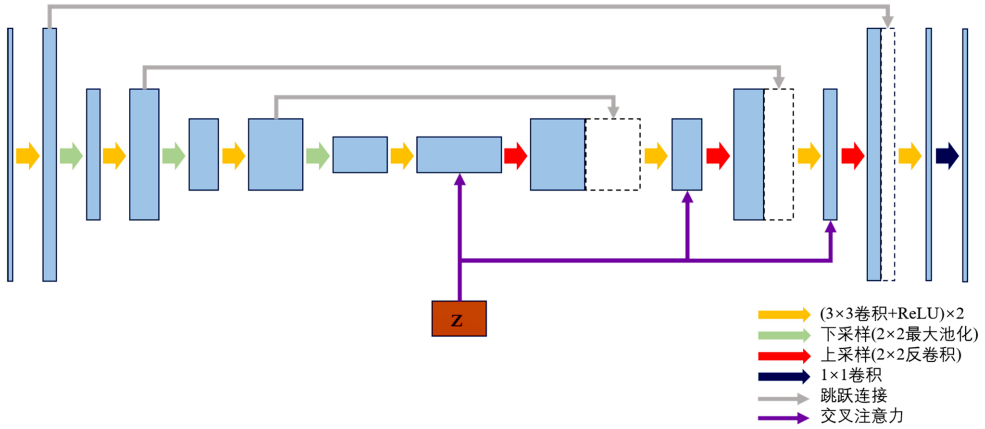

Figure 3. Facial geometric detail generation network structure
图 3. 人脸几何细节生成网络结构

图 4 显示了交叉注意力模块的具体实现，其操作主要在解码阶段。在这一阶段，本文将先前粗糙人脸重建过程中产生的身份编码 β (与静态细节相关)与表情编码 ε (与动态细节相关)进行拼接，形成一个 250 维的条件编码 z。针对每个尺度的特征图，本文将其与先验条件 z 进行交互，其中 z 通过线性变换调整至与特征图相同的维度空间。同时，特征图也分别经过线性映射，生成对应的键向量和值向量。

在交叉注意力的计算过程中，查询向量与键向量进行矩阵乘法运算，生成一个注意力权重矩阵。这个矩阵反映了查询向量中身份、表情和姿态特征与解码阶段各个尺度特征之间的相关性强弱。最后，本文将这个权重矩阵应用于值向量，通过矩阵乘法和 1 × 1 卷积层得到一组加权后的特征图。这一过程实现了对原始特征图中有意义信息的增强和对无关信息的抑制，从而提升了面部细节的重建质量。

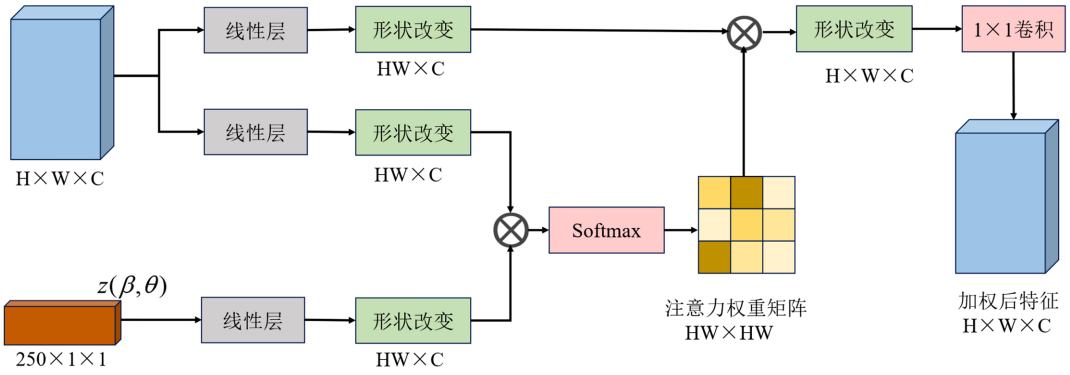

Figure 4. Cross-attention network architecture
图 4. 交叉注意力网络结构

获得位移深度图 D 后，本文能够重建出带有高频细节的人脸几何。具体来说，本文首先将人脸几何 S_{coarse} 和表面法向量 N_{coarse}(即根据粗糙三维人脸模型计算的顶点法线)展开转换到 UV 空间，得到 S'_{coarse} 和 N'_{coarse}，然后通过与位移深度图计算得到 UV 空间下精细化的人脸几何网格 S'_{detail}：

$$S'_{detail} = S'_{coarse} + D \odot N_{coarse} \tag{6}$$

通过这种方式，本文能够重建出精细化的三维人脸，将其渲染后的二维图像可以表示为：

$$I_{r,detail} = \mathcal{R}\left(S_{detail}, A, \gamma\right) \tag{7}$$

3.3. 损失函数

粗糙三维人脸重建阶段的损失函数是在输入图像 I 和渲染图像 $I_{r.coarse}$ 之间计算差异的，可以表示为：

$$L_{coarse} = w_{id}L_{id} + w_{exp}L_{exp} + w_{lmk}L_{lmk} + w_{pho}L_{pho} + w_{reg}L_{reg} \tag{8}$$

包括身份一致性损失 L_{id}、表情一致性损失 L_{exp}、关键点损失 L_{lmk}、光度损失 L_{pho} 和正则化损失 L_{reg}。

身份一致性损失。为了保证输入图像和渲染后的图像要确保看起来是同一个人，本文将 I 和 $I_{r.coarse}$ 分别通过一个预训练的人脸识别网络[30]来提取特征，然后计算两个特征间的余弦相似性来最小化身份损失：

$$L_{id} = 1 - \frac{F_{id}(I)F_{id}(I_{r,coarse})}{\|F_{id}(I)\|_2 \cdot \|F_{id}(I_{r,coarse})\|_2} \tag{9}$$

其中，$F_{id}(\cdot)$ 表示输出的人脸身份特征。

表情一致性损失。为了保证输入图像和渲染后的图像的表情相同，本文将 I 和 $I_{r.coarse}$ 分别通过一个预训练的表情识别网络[31]来提取特征，然后计算两个特征间的余弦相似性来最小化表情损失：

$$L_{exp} = \|F_{exp}(I) - F_{exp}(I_{r,coarse})\|_2 \tag{10}$$

其中，$F_{exp}(\cdot)$ 表示输出的身份特征。

关键点损失。本文采用一个关键点检测网络[32]分别提取 I 和 $I_{r.coarse}$ 上的 68 个人脸关键点 k_i 和 k_i'，然后计算两者的 l_1 损失：

$$L_{lmk} = \sum_{i=1}^{68} \|k_i - k_i'\|_1 \tag{11}$$

光度损失。光度损失计算 I 和 $I_{r.coarse}$ 面部区域间的像素值差异：

$$L_{pho} = V_I \odot (I - I_{r,coarse})_1 \tag{12}$$

其中，$\odot$ 表示对矩阵的逐元素乘法操作，V_I 表示用[33]生成的面部掩码，在面部皮肤区域值为 1，其他区域值为 0。仅计算面部区域的误差可以增强网络对头发、衣服等常见遮挡的鲁棒性。

正则化损失。为了防止面部形状和纹理退化，本文对回归的三维形变模型系数采用正则化损失：

$$L_{seg} = \|\beta\|_2^2 + \|\varepsilon\|_2^2 + \|\alpha\|_2^2 \tag{13}$$

精细化三维人脸重建阶段的损失函数是在输入图像 I 和渲染图像 $I_{r,detail}$ 之间计算差异的，可以表示为：

$$L_{detail} = \omega_{phoD}L_{phoD} + \omega_{smo}L_{smo} + \omega_{regD}L_{regD} \tag{14}$$

包括细节光度损失 L_{phoD}、法线平滑损失 L_{smo} 和细节正则化损失 L_{regD}。

细节光度损失。细节光度损失计算 I 和 $I_{r,detail}$ 面部区域间的像素值差异：

$$L_{phoD} = V_I \odot (I - I_{r,detail})_1 \tag{15}$$

法线平滑度损失。本文在 UV 空间对原始的表面法向量 N'_{coarse} 和叠加了深度位移图后的表面法向量 N'_{detail} 之间计算平滑损失，确保生成的几何细节在相邻像素上的相似性。该损失函数的作用是利用叠加深度位移图前后表面法向量的邻域深度值来约束非边缘区域的法线贴图平滑，增强边缘区域的信息的同时，抑制非边缘区域的信息。损失函数可以表示为：

$$L_{smo} = \sum_{i \in P_{UV}} \sum_{j \in \mathcal{N}(i)} \|\Delta n(i) - \Delta n(j)\|^2 \tag{16}$$

其中 i 表示 UV 空间 P_{UV} 中的每一个像素，$\mathcal{N}(i)$ 表示其临近的像素集和，$\Delta n(i)$ 表示 N_{coarse} 和 N'_{detail} 之间

的像素距离。

细节正则化损失。本文对生成的位移深度图 D 进行正则化计算，以减少生成位移深度图和反射率图的噪声带来的面部失真：

$$L_{regD} = \|D\|_2^2 \tag{17}$$

4. 实验

4.1. 数据集

CelebA 数据集[34]是一个包含 202,599 张二维人脸图像的综合性数据集，涵盖了 10,177 个不同的对象。数据集中每张人脸图像包含 5 个人脸关键点及 40 个边缘点的标注。该数据集涵盖了多种姿态、环境光照、遮挡和表情等变化因素。在本文中，该数据集以 8:1:1 的比例分为训练集、验证集和测试集。

AFLW2000-3D [35]是一个在无约束场景下采集的人脸数据集，包含 2000 幅图像和以及与之对应的三维注释信息。这些注释信息包含了每幅图像的 33DMM 参数，以及 68 个人脸关键点坐标。该数据集中含有多种姿态和光照条件下的人脸图像，在本文中作为三维人脸重建的测试数据集。

4.2. 评价指标

标准化平均误差(Normalized Mean Error, NME)表示人脸关键点坐标预测值和真实值之间的欧式距离平均值，初以标准因子得到的结果。其公式为：

$$NME\left(P, P^*\right) = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|p_i - p_i^*\right\|_2}{d} \tag{18}$$

其中，P 和 P^*分别表示人脸关键点坐标预测值和真实值，p_i 和 p_i^* 分别为重建人脸和真实人脸第 i 个关键点坐标，d 表示标准化因子，通常为两外眼角或两眼瞳孔的距离，N 表示关键点个数。

4.3. 实现细节

本文将输入图像以人脸区域为中心裁剪或缩放为 224 × 224 的大小，特征编码器由 ResNet-50 [36]构成。本文利用 Pytorch3D 可微分光栅器[37]进行三维人脸的渲染。在粗糙三维人脸重建阶段，损失函数参数$\left\{w_{id}, w_{exp}, w_{lmk}, w_{pho}, w_{reg}\right\}$设置为{1,1,0.5,2,1.7e-3}。训练的初始学习率为 1e-4，训练周期为 50。在人脸细节重建阶段，损失函数$\left\{w_{pho}, w_{smo}, w_{regD}\right\}$设置为{2,10,0.5}。训练的初始学习率为 5e-5，训练周期为 30。每个阶段的训练都使用 Adam [38]作为优化迭代器。

4.4. 定量结果

本文在 AFLW2000-3D 数据集上采用 NME 指标与现有方法进行对比，评价重建的三维人脸几何的与原始输入图像的对齐精度。为了进一步研究重建的人脸几何在不同姿态下的对齐准确性，本文将数据集按照偏航角分为三部分，分别表示人脸偏转角度在[0,30)、[30,60)和[60,90)区间的人脸。

表 1 显示了本文的方法相比于其他方法，在[0,30)区间的 NME 最低，在[30,60)和[60,90)区间的 NME 也低于大部分现有方法。MGCNet [44]和 SADRNet [43]在[30,60)区间的 NME 比本文的方法分别低 0.35 和 0.02；在[60,90)区间的 NME 比本文的方法分别低 0.28 和 0.07。这是因为 MGCNet 以多视角人脸图像作为输入的特点，对于人脸大姿态角度图像的三维人脸重建更具有鲁棒性。SADRNet 则依赖于带有三维注释标签的数据集进行训练，而本文的方法采用无三维注释标签的数据集进行自监督训练。AFLW2000-3D 数据集在三个角度区间的分布不均匀，本文在整体均值 NEM 上达到了最低水平。综合分析结果可见，尽管本

文的方法在大姿态角度下的重建精度尚待提升，但在小姿态角度下的重建精度却优于其他现有方法。这一结果证明了本文提出的个性化特征提取模块能有效学习到人脸面部更丰富的身份特征。

Table 1. Quantitative comparison between the proposed method and other methods on the AFLW200-3D dataset
表 1. 本文方法在 AFLW200-3D 数据集上与其他方法的定量对比

方法	NME (mm)			
	[0,30)	[30,60)	[60,90)	平均
SDM [39]	3.67	4.94	9.67	6.12
3DDFA [35]	3.78	4.54	7.93	5.42
3DSTN [40]	3.15	4.33	5.98	4.49
PRNet [41]	2.75	3.51	4.61	3.62
3DDFA-V2 [42]	2.75	3.49	4.53	3.59
SADRNet [43]	2.62	3.44	4.41	3.49
MGCNet [44]	2.72	3.12	3.76	3.20
本文方法	2.51	3.47	4.48	3.17

4.5. 定性结果

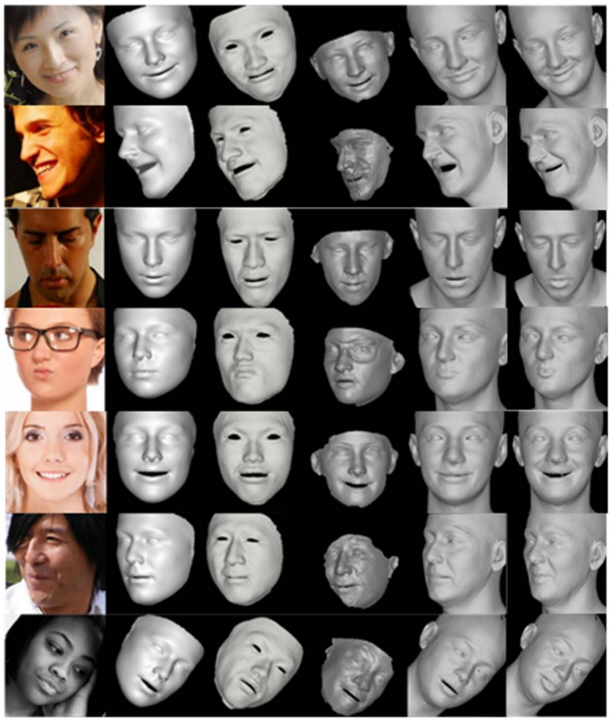

Figure 5. Qualitative comparison between the proposed method and other methods on the AFLW200-3D dataset
图 5. 本文方法在 AFLW200-3D 数据集上与其他方法的定性对比

图 5 显示了本文提出的三维人脸几何重建方法与其他现有方法的定性对比结果。图中(a)列表示 AFLW2000-3D 数据集中的原始输入人脸图像，(b)列至(f)列则分别对应 3DDFA [35]、FaceScape [11]、

Extreme3D [45]、DECA [10]以及本文所提出方法的三维人脸几何重建结果。其中，3DDFA 方法通过人脸标准化投影坐标编码的方式提升了人脸关键点坐标的预测精度。然而，由于仅依赖于 3DMM 参数进行重建，该方法的重建结果受限于低维空间，无法准确还原如面部皱纹等关键细节。FaceScape 和 DECA 方法虽然在重建皱纹等高频细节方面有所突破，但是无法有效恢复出人脸的身份和表情信息。相较于本文方法，Extreme3D 能够重建出更为丰富的细节，但其鲁棒性较差，在脸部存在小部分遮挡(如眼镜、头发等)的情况下易产生伪影，这一点在图示的第二行、第四行和第六行中尤为明显。

相较于其他方法，本文所提出的三维人脸几何重建技术能够生成高保真的三维人脸模型，且细节丰富。从脸部外形的重建效果来看，本文的方法能够更精确地还原面部几何形态和轮廓线条。以图中第三行为例，通过本文方法重建的下巴形状和脸部轮廓与原始输入图像更加吻合。此外，在面部细节的呈现上，本文提出的方法能够准确捕捉与表情变化相关的细微特征。以图中第一行、第二行和第四行为例，通过本文方法重建的模型能够精细地还原由表情引起的嘴部形变以及嘴部和眼部周围的皱纹。总体而言，本文所提出的方法在人脸面部几何的恢复上具有显著优势，重建出的人脸模型在身份和表情的表达上与原始输入图像保持更高的一致性。

4.6. 消融实验

为了验证本文提出的各个模块的合理性和有效性，本文在 AFLW2000-3D 数据集上进行了消融实验。图 6 是消融实验的结果分析，其中(a)列表示输入人脸图像，(b)列至(e)列分别表示基础 3DMM、采用多尺度个性化特征提取模块、采用人脸几何细节重建模块和采用条件编码重建的三维人脸几何的 renlia 你重建情况。

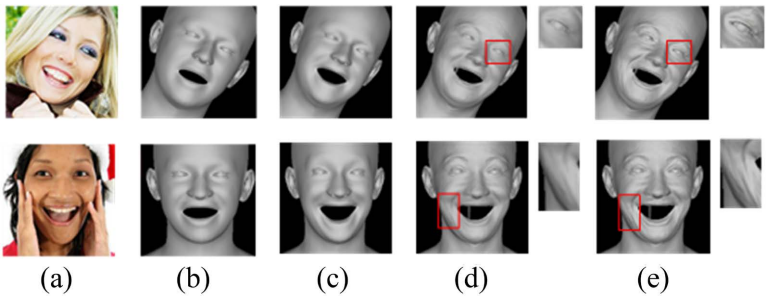

Figure 6. Visualization of ablation experiments
图 6. 消融实验的可视化结果

(1) 多尺度个性化特征提取模块的有效性

此模块旨在有效提取有助于三维人脸重建的个性化特征。表 2 中的数据显示，在未引入多尺度个性化特征提取模块的情况下，NME 提升了 1.28。进一步地，从可视化结果来看，图 6 显示出在缺少此模块的情况下，本文的方法难以精确重建出与原始人脸一致的表情和面部轮廓，特别是在嘴部和面部形状等细节上表现尤为明显。

Table 2. Quantitative results of ablation experiments
表 2. 消融实验的定量结果

方法				平均误差
基础 3DMM	个性化特征提取模块	人脸几何细节重建模块	条件编码	NME
√				5.91

续表

√	√			4.63
√	√	√		3.56
√	√	√	√	3.17

(2) 人脸几何细节重建模块的有效性

此模块主要负责重建人脸表面的皱纹等几何细节。为了验证该模块中身份和表情的联合条件变量对于人脸细节重建的有效性，本文同样进行了消融实验。表 2 的结果显示，在不采用人脸几何细节重建模块的情况下，NME 提升了 1.07；在采用此模块但不引入条件变量的情况下，NME 提升了 0.39。此外，从图 6 中的可视化结果来看，若缺少人脸几何重建模块，本文方法无法有效重建出人脸面部的皱纹等高频细节；而在未采用条件变量的情况下，与表情相关的皱纹细节(如嘴角和眼部周围的皱纹)无法得到清晰的重建。

5. 总结与展望

本文提出了一种由粗到精的高保真三维人脸重建算法，旨在解决基于 3DMM 模型回归的三维人脸重建表达能力不足的问题。该算法将三维人脸的重建分为粗糙和精细化两个过程，从人脸几何轮廓重建到进一步的面部细节生成。为验证所提出方法的有效性，本文在 CelebA 数据集和 AFLW2000-3D 数据集上进行了模型评估。实验结果表明，在小姿态下，本文提出的三维人脸重建方法在几何和纹理重建方面均取得了显著的效果。然而，在中、大姿态下的重建仍存在一定的挑战和不足，这为未来的研究提供了方向。

参考文献

[1] Medin, S.C., Egger, B., Cherian, A., *et al*. (2022) MOST-GAN: 3D Morphable StyleGAN for Disentangled Face Image Manipulation. *Proceedings of the* 36*th AAAI Conference on Artificial Intelligence*, Vol. 36, Online, 22 February-1 March 2022, 1962-1971. https://doi.org/10.1609/aaai.v36i2.20091

[2] Gecer, B., Lattas, A., Ploumpis, S., *et al*. (2020) Synthesizing Coupled 3D Face Modalities by Trunk-Branch Generative Adversarial Networks. *Computer Vision—ECCV* 2020, Glasgow, 23-28 August 2020, 415-433. https://doi.org/10.1007/978-3-030-58526-6_25

[3] Olivier, N., Baert, K., Danieau, F., *et al*. (2023) Facetunegan: Face Autoencoder for Convolutional Expression Transfer Using Neural Generative Adversarial Networks. *Computers & Graphics*, **110**, 69-85. https://doi.org/10.1016/j.cag.2022.12.004

[4] Dipanda, A. and Woo, S. (2005) Towards a Real-Time 3D Shape Reconstruction Using a Structured Light System. *Pattern Recognition*, **38**, 1632-1650. https://doi.org/10.1016/j.patcog.2005.01.006

[5] Lee, H., Song, S. and Jo, S. (2016) 3D Reconstruction Using a Sparse Laser Scanner and a Single Camera for Outdoor Autonomous Vehicle. 2016 *IEEE* 19*th International Conference on Intelligent Transportation Systems* (*ITSC*), Rio de Janeiro, 1-4 November 2016, 629-634. https://doi.org/10.1109/ITSC.2016.7795619

[6] Blanz, V. and Vetter, T. (1999) A Morphable Model for the Synthesis of 3D Faces. *Proceedings of the* 26*th Annual Conference on Computer Graphics and Interactive Techniques*, Los Angeles, 8-13 August 1999, 187-194. https://doi.org/10.1145/311535.311556

[7] Cao, C., Weng, Y., Zhou, S., *et al*. (2013) Facewarehouse: A 3D Facial Expression Database for Visual Computing. *IEEE Transactions on Visualization and Computer Graphics*, **20**, 413-425. https://doi.org/10.1109/TVCG.2013.249

[8] Li, T., Bolkart, T., Black, M.J., *et al*. (2017) Learning a Model of Facial Shape and Expression from 4D Scans. *ACM Transactions on Graphics*, **36**, Article No. 194. https://doi.org/10.1145/3130800.3130813

[9] Chen, A., Chen, Z., Zhang, G., *et al*. (2019) Photo-Realistic Facial Details Synthesis from Single Image. 2019 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, 27 October-2 November 2019, 9428-9438. https://doi.org/10.1109/ICCV.2019.00952

[10] Feng, Y., Feng, H., Black, M.J., *et al*. (2021) Learning an Animatable Detailed 3D Face Model from in-the-Wild Im-

ages. *ACM Transactions on Graphics*, **40**, Article No. 88. https://doi.org/10.1145/3450626.3459936

[11] Yang, H., Zhu, H., Wang, Y., *et al*. (2020) Facescape: A Large-Scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 598-607. https://doi.org/10.1109/CVPR42600.2020.00068

[12] Chen, Y., Wu, F., Wang, Z., *et al*. (2020) Self-Supervised Learning of Detailed 3D Face Reconstruction. *IEEE Transactions on Image Processing*, **29**, 8696-8705. https://doi.org/10.1109/TIP.2020.3017347

[13] Cao, C., Bradley, D., Zhou, K., *et al*. (2015) Real-Time High-Fidelity Facial Performance Capture. *ACM Transactions on Graphics*, **34**, Article No. 46. https://doi.org/10.1145/2766943

[14] Romdhani, S. and Vetter, T. (2005) Estimating 3D Shape and Texture Using Pixel Intensity, Edges, Specular Highlights, Texture Constraints and a Prior. 2005 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR*'05), San Diego, 20-25 June 2005, 986-993. https://doi.org/10.1109/CVPR.2005.145

[15] Paysan, P., Knothe, R., Amberg, B., *et al*. (2009) A 3D Face Model for Pose and Illumination Invariant Face Recognition. 2009 *Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, Genova, 2-4 September 2009, 296-301. https://doi.org/10.1109/AVSS.2009.58

[16] Lee, Y.J., Lee, S.J., Park, K.R., *et al*. (2012) Single View-Based 3D Face Reconstruction Robust to Self-Occlusion. *EURASIP Journal on Advances in Signal Processing*, **2012**, Article 176. https://doi.org/10.1186/1687-6180-2012-176

[17] Daněček, R., Black, M.J. and Bolkart, T. (2022) EMOCA: Emotion Driven Monocular Face Capture and Animation. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 20279-20290. https://doi.org/10.1109/CVPR52688.2022.01967

[18] Deng, Y., Yang, J., Xu, S., *et al*. (2019) Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Long Beach, 16-17 June 2019, 285-295. https://doi.org/10.1109/CVPRW.2019.00038

[19] Tewari, A., Zollhofer, M., Kim, H., *et al*. (2017) MOFA: Model-Based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. 2017 *IEEE International Conference on Computer Vision Workshops* (*ICCVW*), Venice, 22-29 October 2017, 1274-1283. https://doi.org/10.1109/ICCVW.2017.153

[20] Fan, X., Cheng, S., Huyan, K., *et al*. (2020) Dual Neural Networks Coupling Data Regression with Explicit Priors for Monocular 3D Face Reconstruction. *IEEE Transactions on Multimedia*, **23**, 1252-1263. https://doi.org/10.1109/TMM.2020.2994506

[21] Zhu, W., Wu, H.T., Chen, Z., *et al*. (2020) ReDA: Reinforced Differentiable Attribute for 3D Face Reconstruction. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 4957-4966. https://doi.org/10.1109/CVPR42600.2020.00501

[22] Tran, L. and Liu, X. (2018) Nonlinear 3D Face Morphable Model. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7346-7355. https://doi.org/10.1109/CVPR.2018.00767

[23] Lee, G.H. and Lee, S.W. (2020) Uncertainty-Aware Mesh Decoder for High Fidelity 3D Face Reconstruction. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 6099-6108. https://doi.org/10.1109/CVPR42600.2020.00614

[24] Sela, M., Richardson, E. and Kimmel, R. (2017) Unrestricted Facial Geometry Reconstruction Using Image-to-Image Translation. 2017 *IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 1585-1594. https://doi.org/10.1109/ICCV.2017.175

[25] Wang, T.C., Liu, M.Y., Zhu, J.Y., *et al*. (2018) High-Resolution Image Synthesis and Semantic Manipulation with Conditional Gans. 2018 *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 8798-8807. https://doi.org/10.1109/CVPR.2018.00917

[26] Tewari, A., Zollhöfer, M., Garrido, P., *et al*. (2018) Self-Supervised Multi-Level Face Model Learning for Monocular Reconstruction at over 250 Hz. 2018 *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 2549-2559. https://doi.org/10.1109/CVPR.2018.00270

[27] Bas, A., Huber, P., Smith, W.A.P., *et al*. (2017) 3D Morphable Models as Spatial Transformer Networks. 2017 *IEEE International Conference on Computer Vision Workshops*, Venice, 22-29 October 2017, 895-903. https://doi.org/10.1109/ICCVW.2017.110

[28] Catmull, E. (1998) Computer Display of Curved Surfaces. In: Wolfe, R., Ed., *Seminal Graphics*: *Pioneering Efforts that Shaped the Field*, Association for Computing Machinery, New York, 35-41. https://doi.org/10.1145/280811.280920

[29] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention—MICCAI* 2015, Munich, 5-9 October 2015, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[30] Parkhi, O.M., Vedaldi, A. and Zisserman, A. (2015) Deep Face Recognition. *BMVC* 2015—*Proceedings of the British Machine Vision Conference* 2015, Swansea, 7-10 September 2015, 41.1-41.12. https://doi.org/10.5244/C.29.41

[31] Wang, K., Peng, X., Yang, J., *et al*. (2020) Suppressing Uncertainties for Large-Scale Facial Expression Recognition. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 6896-6905. https://doi.org/10.1109/CVPR42600.2020.00693

[32] Yin, X., Tai, Y., Huang, Y., *et al*. (2020) Fan: Feature Adaptation Network for Surveillance Face Recognition and Normalization. *Proceedings of the* 2020 *Asian Conference on Computer Vision*, Kyoto, 30 November-4 December 2020, 301-319. https://doi.org/10.1007/978-3-030-69532-3_19

[33] Nirkin, Y., Masi, I., Tuan, A.T., *et al*. (2018) On Face Segmentation, Face Swapping, and Face Perception. 2018 13*th IEEE International Conference on Automatic Face & Gesture Recognition* (*FG* 2018), Xi'an, 15-19 May 2018, 98-105. https://doi.org/10.1109/FG.2018.00024

[34] Liu, Z., Luo, P., Wang, X., *et al*. (2015) Deep Learning Face Attributes in the Wild. 2015 *IEEE International Conference on Computer Vision*, Santiago, 7-13 December 2015, 3730-3738. https://doi.org/10.1109/ICCV.2015.425

[35] Zhu, X., Lei, Z., Liu, X., *et al*. (2016) Face Alignment across Large Poses: A 3D Solution. 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 146-155. https://doi.org/10.1109/CVPR.2016.23

[36] He, K., Zhang, X., Ren, S., *et al*. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[37] Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.-Y., Johnson, J. and Gkioxari, G. (2020) PyTorch3D. https://github.com/facebookresearch/pytorch3d

[38] Kingma, D.P. and Ba, J. (2014) Adam: A Method for Stochastic Optimization. arXiv Preprint arXiv:1412.6980

[39] McDonagh, J. and Tzimiropoulos, G. (2016) Joint Face Detection and Alignment with a Deformable Hough Transform Model. *Computer Vision—ECCV* 2016 *Workshops*, Amsterdam, 8-10 and 15-16 October, 2016, 569-580. https://doi.org/10.1007/978-3-319-48881-3_39

[40] Bhagavatula, C., Zhu, C., Luu, K., *et al*. (2017) Faster than Real-Time Facial Alignment: A 3D Spatial Transformer Network Approach in Unconstrained Poses. 2017 *IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 4000-4009. https://doi.org/10.1109/ICCV.2017.429

[41] Feng, Y., Wu, F., Shao, X., *et al*. (2018) Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network. *Computer Vision—ECCV* 2018, Munich, 8-14 September 2018, 557-574. https://doi.org/10.1007/978-3-030-01264-9_33

[42] Guo, J., Zhu, X., Yang, Y., *et al*. (2020) Towards Fast, Accurate and Stable 3D Dense Face Alignment. *Computer Vision—ECCV* 2020, Glasgow, 23-28 August 2020, 152-168. https://doi.org/10.1007/978-3-030-58529-7_10

[43] Ruan, Z., Zou, C., Wu, L., *et al*. (2021) SADRNet: Self-Aligned Dual Face Regression Networks for Robust 3D Dense Face Alignment and Reconstruction. *IEEE Transactions on Image Processing*, **30**, 5793-5806. https://doi.org/10.1109/TIP.2021.3087397

[44] Shang, J., Shen, T., Li, S., *et al*. (2020) Self-Supervised Monocular 3D Face Reconstruction by Occlusion-Aware Multi-View Geometry Consistency. *Computer Vision—ECCV* 2020, Glasgow, 23-28 August 2020, 53-70. https://doi.org/10.1007/978-3-030-58555-6_4

[45] Trần, A.T., Hassner, T., Masi, I., *et al*. (2018) Extreme 3D Face Reconstruction: Seeing through Occlusions. 2018 *IEEE*/*CVF* Conference on Computer Vision and Pattern Recognition, Salt Lake City, 18-23 June 2018, 3935-3944. https://doi.org/10.1109/CVPR.2018.00414