

中国科学技术大学

University of Science and Technology of China

# 硕士学位论文



论文题目 基于单幅人脸图像的高质量三维人脸重建算法研究

作者姓名 武兴

学科专业 计算机应用技术

导师姓名 黄章进 副教授

完成时间 二〇二四年五月

# 中国科学技术大学

# 硕士学位论文



## 基于单幅人脸图像的高质量三维人脸重建算法研究

作者姓名： 武兴

学科专业： 计算机应用技术

导师姓名： 黄章进 副教授

完成时间： 二〇二四年五月二十五日

University of Science and Technology of China  
A dissertation for master's degree



# **Research on High-Quality 3D Face Reconstruction Algorithm Based on Single Face Image**

Author: Wu Xing

Speciality: Computer Application Technology

Supervisor: Associate Prof. Huang Zhangjin

Finished time: May 25, 2024

## 摘 要

从单幅人脸图像重建高质量 3D 形状和纹理是计算机视觉与图形学领域一项长期而富有挑战性的任务。这一技术不仅对于人脸分析、识别与动画等领域具有深远影响，同时也是实现更为真实、自然的人机交互的关键所在。近年来，基于深度学习的重建方法在该领域取得了显著进展，然而，其重建质量往往受限于 3D 扫描数据的缺乏。针对这一难题，本文提出了 PR3D (Precise and Realistic 3D face reconstruction) 方法，它包括基于半监督学习的高精度形状重建算法，以及基于 StyleGAN2 的高保真纹理重建算法。

(1) 在基于半监督学习的高精度形状重建算法中，本文使用自然场景人脸图片及有 3D 标记的人脸图片训练辅助编码器和身份编码器，将输入图像编码为 FLAME（一种参数化的 3D 人脸模型）参数，并设计联合优化过程通过最小化基于可微渲染的能量函数进一步优化这些参数。同时，设计了一种新颖的半监督混合关键点损失，不同于已有的稀疏关键点损失，这一损失函数使得 PR3D 能够从自然场景人脸图片以及有 3D 标记的人脸图片中更有效地学习，减少了基于学习的形状重建方法对真实 3D 扫描数据的依赖。此外，为了满足实际应用中对于实时性的需求，采用师生学习蒸馏得到轻量级形状重建模型，在保持较高精度的同时，显著提升了重建速度，为实时人脸重建任务提供了可行的解决方案。通过在多个基于单幅人脸图像的形状重建基准上进行的评估，表明该方法取得了比前沿方法更高的精度。

(2) 在基于 StyleGAN2 的高保真纹理重建算法中，本文提出了基于 StyleGAN2 风格空间人脸重演的纹理提取方法。首先将输入人脸图像作为源图像，具有互补姿态的渲染人脸作为目标图像，训练反演编码器找到源和目标图像在 StyleGAN2 风格空间的风格编码。其次，训练掩膜网络解耦 StyleGAN2 的风格空间，融合源和目标风格编码，以生成与源人脸具有相同身份同时有着互补姿态的重演人脸图像。最后，从源和重演人脸图像提取纹理信息构成面部纹理贴图。通过在不同身份、姿态等条件下人脸图片作为输入时与其他前沿方法的对比，表明该方法一方面能够解决由于缺乏足够 3D 扫描数据造成的保真度低的问题，另一方面能够有效地解决输入图像中由于自遮挡等因素造成的纹理缺失等问题，生成高保真的重建纹理贴图。

**关键词：**3D 人脸重建；形状；纹理；半监督学习；StyleGAN2



## ABSTRACT

Reconstructing high-quality 3D shape and texture from a single face image is a long-standing and challenging task in the fields of computer vision and graphics. This technology not only has profound implications for facial analysis, recognition, and animation but also serves as a key component for achieving more realistic and natural human-computer interaction. In recent years, significant progress has been made in this field with the advent of deep learning-based reconstruction methods. However, the quality of reconstructions often suffers from the lack of 3D scanning data. To address this challenge, this dissertation proposes the PR3D (Precise and Realistic 3D face reconstruction) method, which includes a high-precision shape reconstruction algorithm based on semi-supervised learning and a high-fidelity texture reconstruction algorithm based on StyleGAN2.

(1) In the high-precision shape reconstruction algorithm based on semi-supervised learning, this dissertation utilizes in-the-wild face images and annotated images to train the auxiliary encoder and the identity encoder. These encoders encode the input images into parameters of FLAME (a parametric 3D face model), and a joint optimization process is designed to further optimize these parameters by minimizing an energy function based on differentiable rendering. Additionally, different from existing sparse landmark loss, a novel semi-supervised hybrid landmark loss is proposed, enabling PR3D to learn from in-the-wild face images and annotated images, effectively reducing the reliance on real 3D scanning data. Furthermore, to meet the real-time requirements of practical applications, a lightweight shape reconstruction model is obtained using teacher-student learning distillation, significantly improving reconstruction speed while maintaining high accuracy, thereby providing a feasible solution for real-time face reconstruction tasks. Evaluation on multiple benchmarks for single-image-based shape reconstruction demonstrates that the proposed method outperforms state-of-the-art methods.

(2) In the high-fidelity texture reconstruction algorithm based on StyleGAN2, this dissertation proposes a texture extraction method based on StyleGAN2's style space face reenactment. Firstly, the input face image is used as the source image, and a rendered image with the complementary pose is used as the target image. An inversion encoder is trained to find the style codes of the source and target images in StyleGAN2's style space. Secondly, a mask network is trained to decouple the style space of StyleGAN2 and merge the source and target style codes to generate reenacted face image

with the same identity as the source face while having the complementary pose. Finally, texture information is extracted from the source and reenacted face images to construct the texture map. Comparative evaluations with other state-of-the-art methods under various conditions, such as different identities and poses, demonstrate that this method not only addresses the issue of poor fidelity due to insufficient 3D scanning data but also effectively solves problems such as texture loss caused by factors such as self-occlusion in the input image. As a result, high-quality and high-fidelity reconstructed texture maps are generated.

**Key Words:** 3D face reconstruction; shape; texture; semi-supervised learning; Style-GAN2

# 目 录

|                                    |    |
|------------------------------------|----|
| 第 1 章 绪论.....                      | 1  |
| 1.1 研究背景及意义.....                   | 1  |
| 1.2 国内外研究现状.....                   | 2  |
| 1.2.1 3D 人脸形状重建研究现状.....           | 2  |
| 1.2.2 3D 人脸纹理重建研究现状.....           | 4  |
| 1.3 研究内容.....                      | 5  |
| 1.4 论文组织结构.....                    | 7  |
| 第 2 章 相关工作 .....                   | 9  |
| 2.1 三维人脸模型.....                    | 9  |
| 2.2 基于单幅人脸图像的 shape 重建 .....       | 13 |
| 2.2.1 弱监督学习人脸重建.....               | 13 |
| 2.2.2 监督学习人脸重建.....                | 14 |
| 2.2.3 半监督学习人脸重建.....               | 16 |
| 2.2.4 实时重建.....                    | 18 |
| 2.3 基于单幅人脸图像的 texture 重建 .....     | 18 |
| 2.3.1 线性纹理空间.....                  | 18 |
| 2.3.2 基于 GAN 的纹理空间 .....           | 19 |
| 2.3.3 原图提取.....                    | 20 |
| 2.4 本章小结.....                      | 21 |
| 第 3 章 基于半监督学习的高精度 shape 重建算法 ..... | 22 |
| 3.1 相关模型.....                      | 22 |
| 3.1.1 三维人脸模型.....                  | 22 |
| 3.1.2 相机模型.....                    | 23 |
| 3.1.3 光照模型.....                    | 23 |
| 3.1.4 可微渲染.....                    | 24 |
| 3.1.5 残差神经网络.....                  | 24 |
| 3.2 算法流程.....                      | 25 |
| 3.2.1 辅助编码器.....                   | 27 |
| 3.2.2 身份编码器.....                   | 29 |
| 3.2.3 联合优化.....                    | 29 |
| 3.3 实验与分析.....                     | 30 |
| 3.3.1 数据集与实现细节.....                | 30 |
| 3.3.2 定性比较.....                    | 30 |
| 3.3.3 定量比较.....                    | 31 |

|                                   |    |
|-----------------------------------|----|
| 3.3.4 消融实验.....                   | 33 |
| 3.4 实时重建.....                     | 34 |
| 3.4.1 轻量级卷积神经网络.....              | 35 |
| 3.4.2 算法流程.....                   | 35 |
| 3.4.3 实验与分析.....                  | 36 |
| 3.5 本章小结.....                     | 37 |
| 第4章 基于 StyleGAN2 的高保真纹理重建算法 ..... | 39 |
| 4.1 StyleGAN2 的潜在空间.....          | 39 |
| 4.2 反演权衡.....                     | 39 |
| 4.3 算法流程.....                     | 41 |
| 4.3.1 掩膜网络.....                   | 43 |
| 4.3.2 反演编码器.....                  | 45 |
| 4.4 实验与分析.....                    | 47 |
| 4.4.1 数据集与实现细节.....               | 47 |
| 4.4.2 实验效果.....                   | 47 |
| 4.4.3 对比实验.....                   | 48 |
| 4.5 本章小结.....                     | 51 |
| 第5章 总结与展望 .....                   | 52 |
| 5.1 研究内容总结.....                   | 52 |
| 5.2 未来工作展望.....                   | 53 |
| 参考文献.....                         | 54 |



## 插图清单

|       |                                             |    |
|-------|---------------------------------------------|----|
| 图 1.1 | 3D 面部分析在医疗整形的应用 <sup>[1]</sup> .....        | 1  |
| 图 1.2 | 获取 3D 面部扫描的场景 .....                         | 2  |
| 图 1.3 | 从单幅人脸图像重建 3D 形状及纹理 <sup>[2]</sup> .....     | 3  |
| 图 1.4 | PR3D 整体框架 .....                             | 7  |
| 图 2.1 | BFM 模型的形状和纹理示意图 <sup>[8]</sup> .....        | 11 |
| 图 2.2 | FLAME 在线交互式查看器 <sup>[9]</sup> .....         | 12 |
| 图 2.3 | Dense landmarks 重建框架 <sup>[16]</sup> .....  | 14 |
| 图 2.4 | 半监督学习人脸重建框架 <sup>[20]</sup> .....           | 16 |
| 图 2.5 | FLAME 线性纹理空间 <sup>[9]</sup> .....           | 18 |
| 图 3.1 | FLAME 人脸模型示意图 <sup>[9]</sup> .....          | 22 |
| 图 3.2 | 本文形状重建框架 .....                              | 25 |
| 图 3.3 | 本文形状重建设计动机示意图 .....                         | 26 |
| 图 3.4 | 辅助编码器的训练示意图 .....                           | 27 |
| 图 3.5 | 身份编码器结构图 .....                              | 29 |
| 图 3.6 | 与其他形状重建方法的定性比较（正视图） .....                   | 31 |
| 图 3.7 | 与其他形状重建方法的定性比较（侧视图） .....                   | 32 |
| 图 3.8 | 师生学习蒸馏得到轻量级人脸重建模型 .....                     | 36 |
| 图 3.9 | 用摄像头实时采集人脸并重建 .....                         | 37 |
| 图 4.1 | StyleGAN2 生成器的层次结构 .....                    | 40 |
| 图 4.2 | 通常的纹理重建流程 .....                             | 41 |
| 图 4.3 | 本文纹理重建框架 .....                              | 42 |
| 图 4.4 | 掩膜网络的架构 .....                               | 43 |
| 图 4.5 | 反演编码器网络架构 .....                             | 46 |
| 图 4.6 | 人脸重演实验效果 .....                              | 48 |
| 图 4.7 | 纹理重建实验效果 .....                              | 49 |
| 图 4.8 | 本文纹理重建与其他方法的定性比较 .....                      | 50 |
| 图 4.9 | 与 DECA <sup>[15]</sup> 在具体部位纹理重建细节的比较 ..... | 51 |

# 表格清单

|       |                                          |    |
|-------|------------------------------------------|----|
| 表 3.1 | 各个面部区域到参考人脸（正面视图图像）的距离 .....             | 26 |
| 表 3.2 | NoW 验证集和测试集上的重建误差 .....                  | 33 |
| 表 3.3 | REALY 基准上正面图像和侧面图像的重建误差 .....            | 33 |
| 表 3.4 | 在 NoW 验证集上对不同模块的消融实验 .....               | 34 |
| 表 3.5 | 在 NoW 验证集上对联合优化过程参数的消融实验 .....           | 34 |
| 表 3.6 | 实时重建方法在重建精度和速度上的比较 .....                 | 37 |
| 表 3.7 | PR3D 与 fast-PR3D 在参数量、运行时间、重建精度的对比 ..... | 37 |
| 表 4.1 | 本文人脸重演与其他重演方法定量比较 .....                  | 50 |

# 第 1 章 绪 论

本章首先在 1.1 节介绍基于单幅人脸图片的三维人脸重建的研究意义及背景，随后在 1.2 节分别介绍基于单幅人脸图像的 shape 和 texture 重建的国内外相关工作，然后在 1.3 节介绍本文研究内容，最后在 1.4 节展示本文的组织结构。

## 1.1 研究背景及意义

面部分析技术在多个领域，如人机交互、安全、动画乃至健康领域，均得到了广泛应用。近年来，该领域的一个显著趋势是整合三维数据，旨在克服传统二维人脸分析固有的局限性。由于人脸具备独特的三维特性，仅仅依赖二维图像难以准确捕捉其完整的几何形状，因为这种二维展示方式压缩了一个维度。而三维成像技术则能提供更为精准的面部表示，且这种表示对姿势和光照变化等因素具有不变性，有效解决了二维成像的不足。

在人脸识别领域，传统的基于二维图像的人脸识别系统容易受到视角、光照变化和欺骗攻击的影响，导致识别准确率下降。而三维面部分析则提供了更为丰富的信息，包括人脸的形状、纹理等，这些信息增强了识别系统对于不同人脸之间的区分度，提高人脸识别的安全性。在医学诊断与治疗领域，如图 1.1 所示，医生可以利用三维面部分析模拟整形效果，以帮助医生更好地规划手术方案、预测手术效果。



图 1.1 3D 面部分析在医疗整形的应用<sup>[1]</sup>

然而，三维面部分析系统所带来的这些优势是以更复杂的成像过程为代价的，这在一定程度上限制了其应用范围的广泛性。目前，通常采用立体视觉系统、3D 激光扫描仪或 RGB-D 相机等工具来捕获面部的 3D 信息。如图 1.2 所示，虽然立体视觉系统和 3D 激光扫描仪能够捕获高质量的面部扫描数据，但它们往往需要特定的环境和昂贵的设备支持。相对而言，RGB-D 相机更为经济实惠、操作简便，但其生成的扫描质量却相对有限。

为了克服这些限制，一种颇具吸引力的替代方法是通过单幅人脸 2D 图像来



图 1.2 获取 3D 面部扫描的场景

重建面部的形状与纹理，如图 1.3 所示。这种从 2D 图像重建 3D 人脸的方法旨在融合捕获 2D 图像的简便性与获取面部 3D 几何表示的优势。从单幅人脸图片重建 3D 人脸在各种下游任务领域有着广泛的应用，如 AR/VR<sup>[2]</sup>，人脸识别<sup>[3]</sup>和人脸动画<sup>[4]</sup>。Hu 等人<sup>[2]</sup>将三维人脸重建技术应用于 VR 虚拟 3D 化身的构建。具体的，他们首先将人脸图片预处理，进行面部和头发区域像素级别的分割。对于面部区域，将基于 PCA 的线性人脸模型的形状和外观拟合到分割的面部区域。对于头发区域，根据分割的轮廓和头发的方向场来从发型库中搜索与输入图像最接近的发型。用于头发属性分类的分类网络还可以识别头发外观属性，如头发颜色、纹理和各种着色器参数等，以便进行适当的渲染，从而完成对用于 VR/AR 的 3D 化身建模。

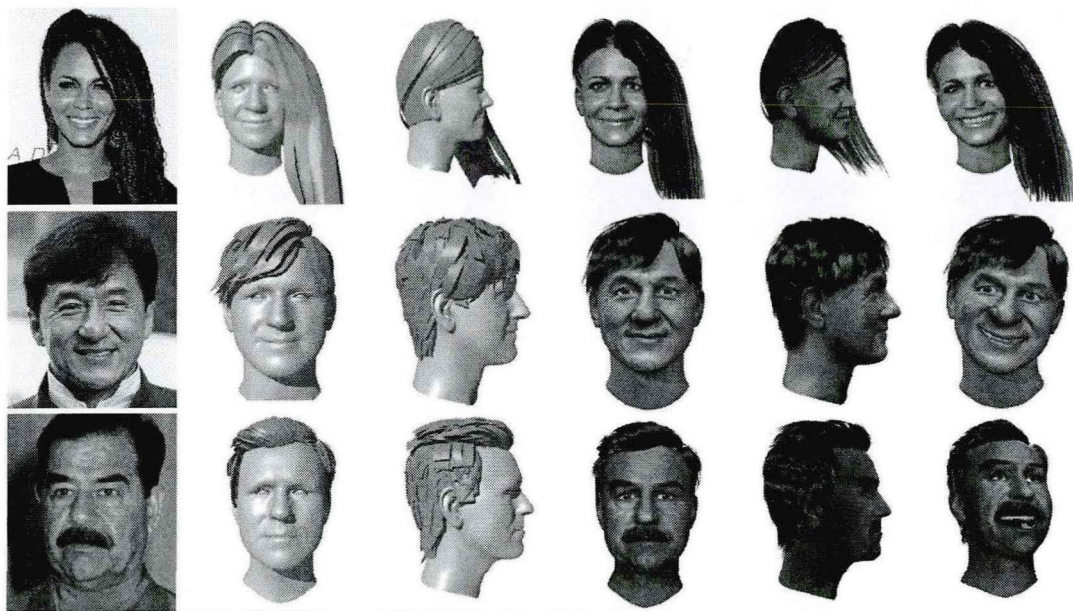
尽管从单幅人脸图像重建三维人脸颇具吸引力，但它本质上是一个不适定问题：即必须从单幅人脸图片中恢复个体的面部几何、头部姿势以及纹理信息，这导致了一个欠定问题的产生。因此，在“3D-from-2D”面部重建的解决方案中存在着一定的模糊性，因为同一张 2D 图片可能由不同的 3D 面部生成，这使得实现高质量的 3D 人脸重建变得相当困难。

## 1.2 国内外研究现状

### 1.2.1 3D 人脸形状重建研究现状

许多基于单幅人脸图像的 3D 形状重建方法都遵循 Vetter 和 Blanz 的方法<sup>[5]</sup>，即通过综合分析的方式估计三维人脸模型（预先计算的统计模型）的系数。此外，还有一些无模型的重建方法直接回归体素<sup>[6]</sup>或网格<sup>[7]</sup>。与无模型工作相比，三维人脸模型能够提供人脸的形状、表情以及纹理的先验知识，可以使用一组系



图 1.3 从单幅人脸图像重建 3D 形状及纹理<sup>[2]</sup>

数向量来表达重建的三维人脸，在时间空间消耗，重建精度等方面更具优势，常见的三维人脸模型有 BFM<sup>[8]</sup>、FLAME<sup>[9]</sup>、HiFi3D++<sup>[10]</sup>等。

综合分析估计三维人脸模型系数的重建方法可以分为基于优化的方法<sup>[11-13]</sup>，和基于学习的方法<sup>[14-17]</sup>。基于优化的方法尽管非常有效，但通常需要在深度、密集光度等约束下，经过长时间的迭代优化过程。这增加了计算的复杂性，并限制了其应用的可行性。近年来，深度学习技术的飞速发展极大地推动了基于学习方法的 3D 形状重建技术的进步。这些方法通常借助深度神经网络 (DNN) 来直接回归三维人脸模型参数，从而实现了更高效和自动化的重建过程。然而，深度神经网络的训练高度依赖于大量带有精确标记的 3D 真实扫描数据。由于现实中获取这类数据的难度和成本较高，因此，数据集的规模和质量成为了限制深度神经网络性能的重要因素。

为了克服这一挑战，研究者们探索了弱监督的训练方法<sup>[14-15,18]</sup>。这些方法通过设计特定的损失函数和优化策略，使得网络能够在没有真实 3D 扫描数据的情况下进行训练。关键点损失是弱监督方法广泛采用的重要损失函数，对于图像可以使用 Bulat 等人<sup>[19]</sup>的关键点检测方法检测图像的 68 个关键点，对于三维人脸模型，设计者往往标注了相应语义的 68 个三维关键点在三维人脸模型的位置，计算将其投影至二维图像后与二维关键点的距离作为关键点损失函数。然而仅仅 68 个关键点难以实现高精度的形状重建，这类弱监督信号本身往往存在误差和数量不足的问题，从而导致形状重建结果的低精度。

为了解决弱监督信号对重建精度的影响，一些工作<sup>[16-17]</sup>构建 2D-3D 数据集以进行监督训练。Wood 等人<sup>[16]</sup>提出了基于合成数据的监督学习人脸重建方法。

通过生成大量合成人脸图像,可以为深度神经网络提供丰富的训练样本。然而,由于合成数据与真实自然场景人脸图像之间存在固有的域差异,这些合成数据无法完全替代真实数据进行训练。因此,虽然合成数据在一定程度上提升了网络的泛化能力,但重建结果的准确性和鲁棒性仍然受到限制。Zielonka 等人<sup>[17]</sup>构建了由成对的 2D/3D 数据构建的 MICA 数据集,它由 8 个较小的数据集组成,包含 FLAME 拓扑下的约 2300 名受试者,但是由于缺乏了表情、姿态相关的训练数据导致无法对表情和姿态进行重建。

陈卓等人<sup>[20]</sup>提出了一种基于 Transformer<sup>[21]</sup>的半监督方法,使用自然场景人脸图像、有 3D 标记人脸图像以及合成人脸图像作为训练数据,直接回归面部顶点,但在面部重建的各种基准测试中表现不佳。本文的形状重建方法采用了类似的半监督学习策略。然而,本研采用不同的框架及损失函数,回归 FLAME 参数而不是直接估计顶点位置。本文使用自然场景人脸图片及有 3D 标记的人脸图片训练辅助编码器和身份编码器,将输入图像编码为 FLAME(一种参数化的 3D 人脸模型)参数,并设计联合优化过程通过最小化基于可微渲染的能量函数进一步优化这些参数。同时,本文设计了一种新颖的半监督混合关键点损失,这一损失函数使得 PR3D 能够从自然场景人脸图像和有 3D 标记的人脸图像中学习,从而有效减轻了对真实 3D 扫描数据的依赖。

大多数现有的形状重建方法侧重于提升重建精度,而相对忽视了重建速度,这在一定程度上限制了这些方法在实时应用场景中的广泛应用。例如,在白浩然等人<sup>[10]</sup>的工作中,每张图像的重建过程在高性能的 NVIDIA Tesla V100 GPU 上测试时,总拟合时间仍然达到了约 60 秒,这显然无法满足实时的需求。为了提升重建速度,研究者们进行了多方面的尝试。郭建珠等人<sup>[22]</sup>通过在网络结构上进行优化,修剪了 77.5% 的通道,以实现在 CPU 上的实时重建速度。然而,这种方法虽然显著提升了速度,但导致了重建精度的明显下降,即错误率显著增加。这表明在网络结构上进行简单的修剪虽然能够减少计算量,但并不能很好地保持重建的精度。另一种策略是朱翔昱等人<sup>[23]</sup>提出的级联结构方法,该方法通过逐步更新一小部分 3DMM 参数来减少每次迭代的计算量。然而,随着级联阶段的数量增加,虽然可以在一定程度上提升重建精度,但总的计算量会线性增加,这也限制了该方法在实时场景中的应用。与以往的方法不同,本文利用 PR3D 获得的知识来指导轻量级模型的训练,在形状重建精度和速度上有了更好的平衡。

### 1.2.2 3D 人脸纹理重建研究现状

除了形状之外,纹理对于确定 3D 人脸重建效果同样重要。纹理重建的保真度直接影响最终重建结果的视觉质量和真实感。大多数基于三维人脸模型的重建方法侧重于提高形状估计精度,而针对纹理重建问题的研究较少。

一些三维人脸模型如 BFM<sup>[8]</sup>、FLAME<sup>[9]</sup>，有自己的线性纹理空间。构建 BFM 线性纹理空间的数据集包括 100 名女性和 100 名男性，从照片中计算得到纹理颜色，修补后通过主成分分析得到 BFM 的线性纹理空间。相似的，FLAME 从 FFHQ 数据集中随机选择了 1500 张图像，通过初始化、模型拟合、纹理补全、纹理空间计算等步骤来计算 FLAME 线性纹理空间。然而这些纹理空间存在质量差、表现能力不足的问题。

为解决上述问题，一些方法致力于构建更高质量的纹理空间。GANFIT<sup>[24]</sup>从 10,000 个 UV 贴图训练生成对抗网络 (GAN<sup>[25]</sup>) 作为纹理解码器，以替换 3DMM 中的线性纹理基底，以增加表现力。然而，他们在训练数据集中的 UV 贴图是从不均匀照明的面部图像中提取的，因此生成的纹理图包含明显的阴影，并不适用于不同光照条件下的渲染。另一项基于 UV-GAN<sup>[26]</sup>的 Lee 等人<sup>[27]</sup>的工作中也存在同样的问题。AvatarMe<sup>[28]</sup>将线性纹理基底拟合与超分辨率网络相结合，从来自 200 个个体的高质量纹理图训练而来。Normalized Avatar<sup>[29]</sup>从一个包含 5000 多个身份的包括高质量的扫描数据和合成数据的纹理图数据集中训练了一个纹理解码器。尽管这些方法生成的纹理图质量很高，但重建的保真度在很大程度上受训练数据集中身份数量的限制。

纹理重建的另一类方法是基于提取的方法，即直接从输入图像中提取纹理信息。例如，Feng 等人<sup>[15]</sup>在 DECA 中采用了一种从输入图像中提取纹理的方法，以实现逼真的外观。然而，从原图提取纹理的方法依赖于高精度的形状重建，由于形状重建精度不足会导致提取纹理出现偏差。另一方面，由于侧视图像中的自遮挡问题，所得到的纹理贴图往往存在缺陷，如部分区域纹理信息缺失或失真。

受到 StyleGAN2 在生成与真实图像难以区分的图像方面出色能力的启发，本文的高保真纹理重建方法结合了前述两类方法的优势。一方面，将 StyleGAN2 的潜在空间作为传统线性空间的替代，以增强表现力。另一方面，本文探索了一种基于人脸重演的方法来提取高质量、高保真的纹理贴图。具体的，利用预训练的 StyleGAN2 生成模型的强大生成能力，通过训练反演编码器和隐空间的掩膜网络，实现人脸重演，然后从源图像和具有互补姿态的重演图像中提取纹理信息，得到高保真的纹理贴图。

### 1.3 研究内容

本文提出一种名为 PR3D(Precise and Realistic 3D face reconstruction) 的方法，旨在从单张人脸图像重建高质量三维人脸，包括基于半监督学习的高精度形状重建算法以及基于 StyleGAN2 的高保真纹理重建算法。

在形状重建方面，本文提出了一种基于半监督学习的高精度形状重建算法。



该重建框架包含两个关键组件：身份编码器和辅助编码器。身份编码器负责从输入图像中提取与个体身份相关的特征，并将其编码为 FLAME 形状参数；而辅助编码器则提供额外的辅助信息，如表情、姿态等。为了进一步优化 FLAME 参数，本算法引入了一个联合优化过程，该过程基于可微分渲染技术，通过最小化渲染结果与输入图像之间的差异来迭代调整参数。本文提出了一种混合关键点损失函数。该函数结合了自然场景人脸图像的稀疏关键点损失和有 3D 标注人脸图像的密集关键点损失，使模型能够充分利用有标注人脸图像的 3D 标注信息，同时从大量自然场景人脸图像中学习到丰富的形状变化模式。这种混合关键点损失的设计使得 PR3D 方法在各种输入图像上表现出强大的鲁棒性，特别是在面对具有严重遮挡或极端姿态变化的挑战性场景时，仍能保持出色的重建性能。此外，为了满足实际应用中对于实时性的需求，本文还使用师生学习蒸馏得到轻量级形状重建模型，在保持较高精度的同时，显著提升了重建速度，为实时人脸重建任务提供了可行的解决方案。

在纹理重建方面，本文提出了一种基于 StyleGAN2<sup>[30]</sup> 风格空间的高保真纹理重建方法。首先将输入人脸图像作为源图像，具有互补姿态的渲染人脸作为目标图像，训练反演编码器找到源和目标图像在 StyleGAN2 风格空间的风格编码。其次，训练掩膜网络解耦 StyleGAN2 的风格空间，融合源和目标风格编码，以生成与源人脸具有相同身份同时有着互补姿态的重演人脸图像。最后，从源和重演人脸图像提取纹理信息构成面部纹理贴图。本方法一方面能够解决由于缺乏足够 3D 扫描数据造成的保真度低的问题，另一方面能够有效地解决输入图像中由于自遮挡等因素造成的纹理缺失等问题，生成高保真的重建纹理贴图。

PR3D 的整体框架如图 1.4 所示，图中所展示的蓝线代表本文中实现的高精度形状重建流程，而绿线则描绘了高保真度纹理重建的详细过程。在形状重建方面，PR3D 以单张人脸图像作为输入，通过两个编码器以及后续的联合优化过程，回归得到 FLAME 参数，从而实现对面部形状的精准重建。随后，将 FLAME 参数解码，生成带有详细姿态和表情信息的面部网格，为后续纹理重建提供了精确的几何基础。在纹理重建部分，训练反演编码器以实现从图像到 StyleGAN2 潜在空间的反转映射。以渲染后的具有互补姿态的人脸图像作为目标图像，通过掩膜网络生成具有互补姿态的面部重演图像。随后，从源图像和重演的面部图像中提取纹理信息，构建出面部纹理贴图，并将其应用于重建的 3D 形状，完成高质量三维人脸重建。

综上所述，本文的主要贡献可概括为以下几个方面：

- 本文提出了一种基于半监督学习的高精度形状重建框架，使用自然场景人脸图片及有 3D 标记的人脸图片训练辅助编码器和身份编码器，将输入图像编码为 FLAME 参数，并设计联合优化过程通过最小化基于可微渲染

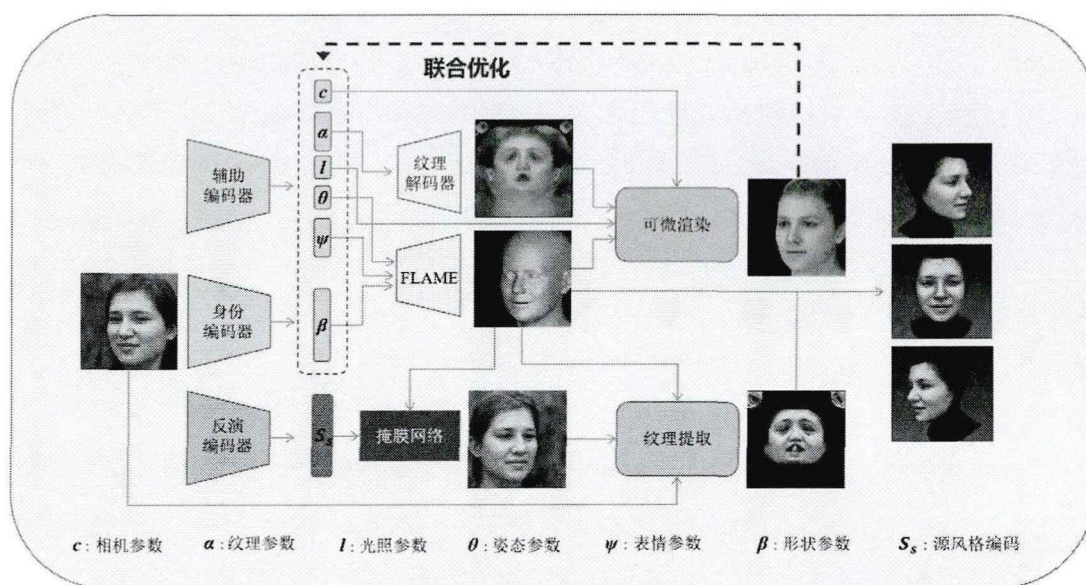


图 1.4 PR3D 整体框架

的能量函数进一步优化这些参数。这一框架的设计显著提高了从单幅人脸图像进行 3D 形状重建的精度和鲁棒性。

- 不同于已有的稀疏关键点损失，本文提出了一种新颖的混合关键点损失函数，旨在从自然场景人脸图片以及有 3D 标记的人脸图片中学习，从而有效减轻了基于学习的形状重建方法对真实 3D 扫描数据的依赖。
- 采用师生学习蒸馏得到轻量级形状重建模型，在保持较高精度的同时，显著提升了重建速度，每张单独图像的推断时间约为 3 毫秒，为实时人脸重建任务提供了可行的解决方案。
- 本文提出了一种基于 StyleGAN2 的高保真纹理重建算法。该算法充分利用了 StyleGAN2 生成模型的生成能力，训练反演编码器与掩膜网络，从源图像和重演图像中提取出高质量、高保真度的纹理信息。
- 在广为认可的 NoW 基准<sup>[31]</sup>和 REALY 基准<sup>[31]</sup>上的定量和定性对比实验表明本文实现了最高的形状重建精度，与其他前沿纹理重建方法的对比实验表明本文实现了更高保真度的纹理重建。

## 1.4 论文组织结构

本论文的组织结构如下所示：

第一章：绪论。首先介绍基于单幅人脸图像的三维人脸重建的研究背景及意义。然后，分别分析了国内外三维人脸形状和纹理重建的研究现状与不足之处，指出高质量重建即提升形状重建精度和提升纹理重建保真度的必要性。最后阐

述了本文的研究内容和论文组织结构。

第二章：相关工作。介绍三维人脸模型，三维人脸形状和纹理重建三个方面的工作。形状重建包括弱监督学习、监督学习和半监督学习三种相关形状重建工作。纹理重建包括线性纹理空间、基于 GAN 的纹理空间以及原图提取三种相关方法。

第三章：基于半监督学习的高精度形状重建算法设计。首先介绍相关模型，包括人脸模型、相机模型等，然后详细介绍本研究的形状重建算法设计动机与框架流程，并实验验证本文形状重建方法的高精度，最后进行在高精度形状重建模型基础上的实时重建工作的介绍。

第四章：基于 StyleGAN2 的高保真纹理重建算法设计。首先介绍 StyleGAN2 的潜在空间，分析真实图像的反演权衡问题，然后详细介绍本文提出的纹理重建算法设计动机与框架流程，最后实验验证本文纹理重建方法的高保真度。

第五章：总结与展望。总结本文在基于单幅人脸图像的高质量三维人脸重建方面做出的贡献以及实际应用，并且对于本文现有的问题提出下一步的研究计划。

## 第2章 相关工作

本章将介绍三维人脸模型，基于单幅人脸图像的形状和纹理重建这三部分的相关工作。在 2.1 节介绍了三维人脸模型的构建方法及表示。在 2.2 节形状重建相关工作中，分别介绍弱监督学习、监督学习和半监督学习形状重建的具体方法，包括流程及损失函数，然后介绍实时形状重建的相关方法。在 2.3 节纹理重建相关工作中，首先介绍线性纹理空间的构建方法，然后介绍基于 GAN 的纹理空间的构建流程及应用，最后介绍从原图提取纹理的重建方法。

### 2.1 三维人脸模型

3DMM，即三维形变模型，是一种经典而重要的参数化的三维人脸模型。最早由 Blanz 和 Vetter<sup>[5]</sup>于 1999 年提出，3DMM 的提出标志着计算机视觉领域在人脸建模方面的重大突破。它被设计为一种统计模型，旨在将人脸的形状和纹理投影到一个线性空间中。这意味着通过一组有限的参数，就可以在一个连续的空间中描述无限多个人脸的形状和纹理变化。3DMM 的核心思想是通过学习来自大量人脸样本的变化，从而捕捉人脸形状和纹理的主要变化模式。通过分析这些变化模式，3DMM 可以生成新的人脸形状和纹理，从而实现对不同人脸的建模和重建。这使得 3DMM 成为了研究人脸识别、人脸动画、虚拟现实和增强现实等领域的重要工具。下文将对两个经典的三维人脸模型 BFM<sup>[8]</sup>和 FLAME<sup>[9]</sup>进行介绍。

#### 1. BFM

##### (1) 3D 人脸扫描

BFM<sup>[8]</sup>的训练数据集包括了 100 名女性和 100 名男性的面部扫描数据，其中大多数是欧洲人。被扫描对象的年龄在 8 岁到 62 岁之间，平均年龄为 24.97 岁，体重在 40 公斤到 123 公斤之间，平均体重为 66.48 公斤。每个人的面部都进行了三次扫描，其中选择了最自然的扫描结果。

##### (2) 注册

为了使原始数据可用，需要将其进行对齐。这意味着扫描需要被重新参数化，以便语义上对应的点（例如鼻尖或眼角）在参数化域中共享相同的位置。注册会为面部的所有点，包括像颊部这样的非结构化区域，建立这种对应关系。在将扫描数据对齐后，再次进行线性组合，得到的仍然是合理的人脸。为了建立对应关系，BFM 使用了改进版的最优步非刚性 ICP 算法<sup>[32]</sup>。注册方法应用于三角化网格的 3D 域中。它会逐渐将一个模板变形成与测量表面相匹配，同时确保平

滑变形。除了建立对应关系外，该方法还通过使用鲁棒的距离度量来填补缺失区域。为了改善模型质量，作者还在嘴唇、眉毛等位置手动添加了标记点。

### (3) 纹理提取与修补

面部反照率由每个顶点的颜色表示，这些颜色是从照片中计算得出的。来自三张照片的信息根据相对可见边界的距离和相对于观察方向的法线方向进行混合。为了改进反照率模型，作者手动去除了头发，并使用修补算法完成了缺失的数据。

### (4) 建模

注册后的面部参数化为  $m = 53490$  个顶点和共享拓扑的三角形网格。顶点  $(x_i, y_i, z_i)^T \in \mathbb{R}^3$  有相关联的颜色  $(r_j, g_j, b_j)^T \in [0, 1]^3$ 。一张人脸使用两个长为  $3m$  的向量表示：

$$\begin{aligned} s &= (x_1, y_1, z_1, \dots, x_m, y_m, z_m)^T \\ t &= (r_1, g_1, b_1, \dots, r_m, g_m, b_m)^T \end{aligned} \quad (2.1)$$

BFM 假设形状和纹理之间独立，构建两个独立的线性模型。利用主成分分析 (PCA) 对数据进行高斯分布拟合，得到一个参数化人脸模型：

$$\begin{aligned} \mathcal{M}_s &= (\mu_s, \sigma_s, U_s) \\ \mathcal{M}_t &= (\mu_t, \sigma_t, U_t) \end{aligned} \quad (2.2)$$

其中  $\mu_{\{s,t\}} \in \mathbb{R}^{3m}$  是均值， $\sigma_{\{s,t\}} \in \mathbb{R}^{n-1}$  是标准差， $U_{\{s,t\}} = [u_1, \dots, u_n] \in \mathbb{R}^{3m \times n-1}$  是形状和纹理的主成分的正交基底。新人脸作为主成分的线性组合从模型中生成：

$$\begin{aligned} s(\alpha) &= \mu_s + U_s \text{diag}(\sigma_s) \alpha \\ t(\beta) &= \mu_t + U_t \text{diag}(\sigma_t) \beta \end{aligned} \quad (2.3)$$

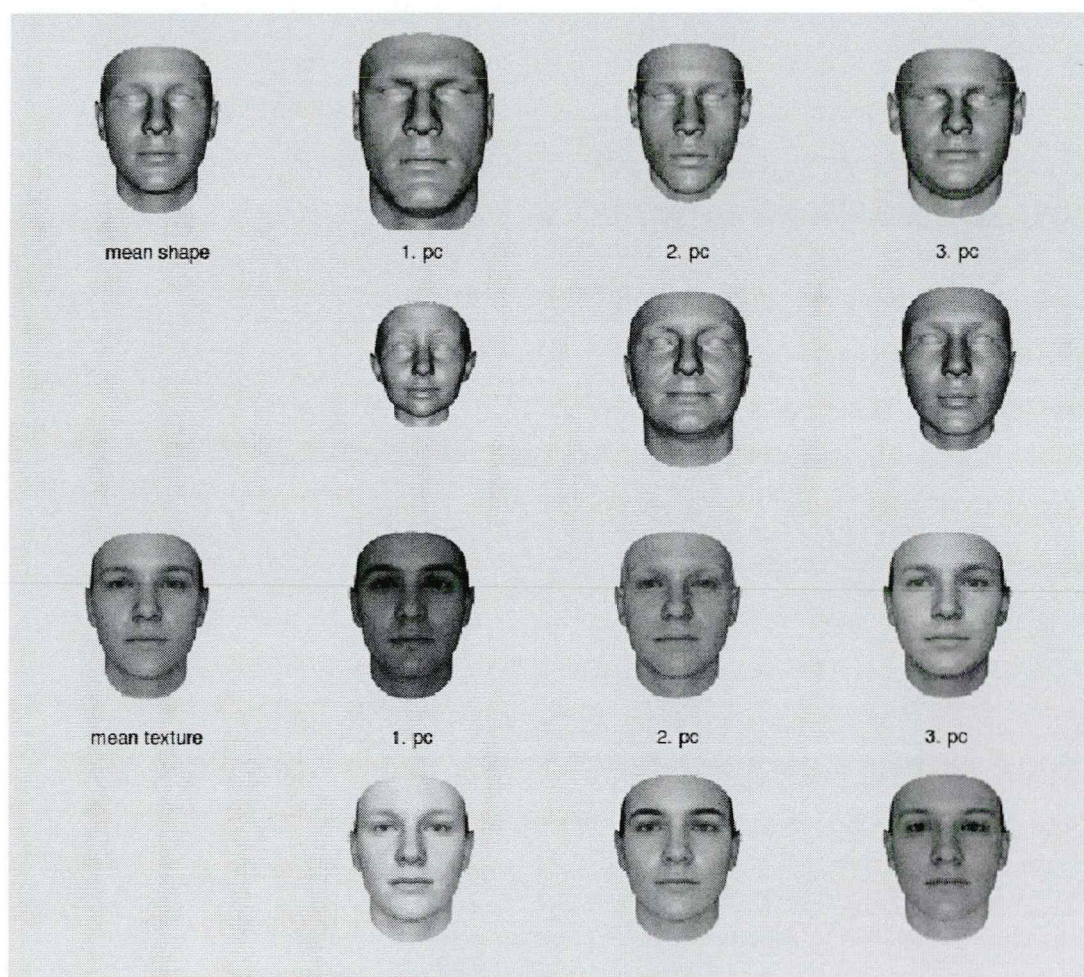
在训练样本正态分布和正确的均值估计的假设下，系数是独立的、单位方差的正态分布。

BFM 模型的形状和纹理示意图如图 2.1 所示，图中表示了 BFM 的平均形状和前三个形状主成分对形状的影响（加减五倍的标准差），以及 BFM 的平均纹理和前三个纹理主成分对纹理的影响（加减五倍的标准差）。

## 2. FLAME

FLAME 模型<sup>[9]</sup> (Faces Learned with an Articulated Model and Expressions) 借鉴了身体模型 SMPL 的表示方式，基于 LBS (linear blend skinning) 并结合 blendshapes 作为表示，包含 5023 个顶点，4 个关节。具体来说，FLAME 将人头拆分成了左眼球、右眼球、下巴、脖子这四个部位，这四个部位都可以绕着自定义的“关节”



图 2.1 BFM 模型的形状和纹理示意图<sup>[8]</sup>

进行旋转形成新的三维表示。通俗来讲，LBS 就是刻画当骨骼发生相对旋转导致皮肤拉伸时，如何计算产生的新“皮肤”的位置。FLAME 等参数化模型可以让人脸的各个属性解耦合，Blendshape 的输入就是每个属性的值，输出就是对应人脸模型怎么发生形变的，具体来说就是 5023 个顶点发生的偏移量。

FLAME 模型的参数有三类：形状，姿态，表情，分别可以表示为  $\beta \in \mathbb{R}^{|\beta|}$ 、姿态  $\theta \in \mathbb{R}^{|\theta|}$  和表情  $\psi \in \mathbb{R}^{|\psi|}$ 。而 FLAME 模型就可以理解为一个输入上述三个参数，输出三维模型的映射  $M(\beta, \theta, \psi) : \mathbb{R}^{|\beta| \times |\theta| \times |\psi|} \rightarrow \mathbb{R}^{3N}$ ，其中  $N$  表示的是顶点数量：5023 个。下面将展开介绍这三个参数的作用逻辑：

假设通过大量的数据得到了一个 FLAME 模型，其中模型的模板（类似于 BFM 中的平均人脸）为  $\bar{T} \in \mathbb{R}^{3N}$ ，它的 pose 参数都为 0。给定  $\beta, \theta, \psi$  三个参数，依次计算三个参数产生的“顶点偏移”，并最后加权所有的参数的影响。对于 shape 参数  $\beta$ ，可以根据公式  $B_S(\beta; S) = \sum_{n=1}^{|\beta|} \beta_n S_n$  计算出其对应的“顶点偏移”，其中  $S_n$  是根据大量数据 PCA 学出来的形状基底，这也是人脸模型 BFM



中经典的做法；对于姿态参数  $\theta$ ，先映射回每个关节的旋转量，然后直接计算出每个顶点的偏移量；对于表情参数  $\psi$ ，可以根据公式  $B_E(\psi; \mathcal{E}) = \sum_{n=1}^{|\psi|} \psi_n E_n$ ，类似于形状参数来计算顶点的偏移量，其中  $E_n$  是根据大量数据 PCA 后学出来的表情基底；得到形状参数、姿态参数、表情参数控制下的人头模型，最后作用上 LBS，就可以得到一个完整、自然的人头模型了。

FLAME 的交互式工具如图 2.2 所示，它用于可视化前 10 个形状参数（共 300 个）和前 10 个表情参数（共 100 个），以及头部和下颌的姿态。通过更改滑块的值，可以在  $\pm 2$  个标准偏差之间变化形状和表情参数。查看器内的下拉菜单允许在女性、男性和通用模型之间进行切换。

FLAME 模型和 BFM 相似，它们的区别如下：FLAME 是一个人头模型，BFM 是一个人脸模型，直观来看 FLAME 的输出是一整个三维的人头结构，包括了后脑勺以及脖子，这是 BFM 模型没有的地方；FLAME 模型通过 LBS 显式地建模了脖子的旋转、眼球的旋转，也是 BFM 所不具备的；FLAME 模型能够表示的表情要比 BFM 模型更加丰富，最原始的 BFM 模型只在 200 个人数据上拟合而成，而 FLAME 模型是在 33000 个人头数据上拟合出来的结果；FLAME 模型没有纹理，BFM 模型有纹理，但有工作将这两者的网格对齐，从而使得在 FLAME 上可以使用 BFM 的纹理空间。

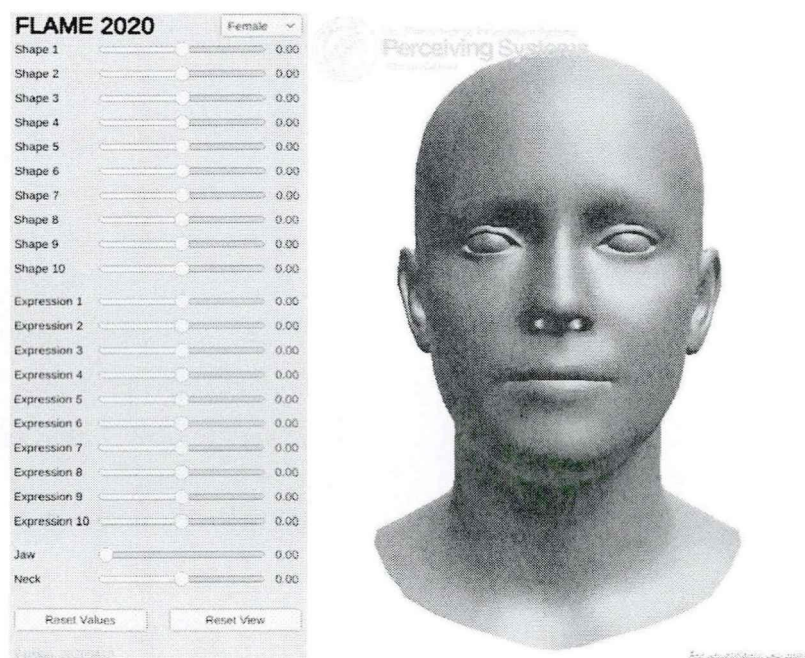


图 2.2 FLAME 在线交互式查看器<sup>[9]</sup>

## 2.2 基于单幅人脸图像的形状重建

### 2.2.1 弱监督学习人脸重建

随着深度神经网络的发展,神经网络可以直接回归 3DMM 参数,无需迭代优化。为了避免网络过拟合,需要大量且多样化的数据集,然而现实中获取真实 3D 扫描的难度和成本较高。一些工作<sup>[14-15]</sup>在无需 3D 信息的情况下以弱监督方式训练网络。他们利用预训练模型检测到的面部关键点和面部掩膜作为计算损失函数的信息。然而,面部关键点的数量有限,且其预测的准确性会限制网络重建高精度面部形状的能力。

以 Deep3D<sup>[14]</sup>为例,给定一个训练的 RGB 图像  $I$ , Deep3D 使用卷积神经网络来回归一个系数向量  $\mathbf{x}$ ,通过一些简单的可微分的数学推导,可以用它来解析生成一个重建的图像  $I'$ 。网络是在没有任何真实标签的情况下进行训练的,它通过评估  $I'$  上的混合级别(图像级别和感知级别)损失并进行反向传播来训练。在图像级别的损失包括光度损失和稀疏的 2D 关键点损失。

#### 1. 光度损失

直接测量原始图像与重建图像之间的密集光度差异是直观的,Deep3D 提出了一种鲁棒的、对皮肤敏感的光度损失,定义为:

$$L_{photo}(\mathbf{x}) = \frac{\sum_{i \in \mathcal{M}} A_i \cdot \|I_i - I'_i(\mathbf{x})\|_2}{\sum_{i \in \mathcal{M}} A_i} \quad (2.4)$$

其中  $i$  表示像素索引,  $\mathcal{M}$  是重投影的人脸区域,  $\|\cdot\|_2$  表示  $L_2$  范数,  $A$  是基于肤色的注意力掩膜。

#### 2. 关键点损失

Deep3D 还使用图像域上的关键点位置作为训练网络的弱监督。使用 Bulat 等人<sup>[19]</sup>的关键点检测方法检测训练图像的 68 个关键点  $\{\mathbf{q}_n\}$ 。在训练过程中,将重建形状的 3D 关键点顶点投影到图像上,得到  $\{\mathbf{q}'_n\}$ ,并计算损失如下:

$$L_{lan}(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N \omega_n \|\mathbf{q}_n - \mathbf{q}'_n(\mathbf{x})\|^2 \quad (2.5)$$

其中  $\omega_n$  是关键点权重,将嘴巴和鼻子上的点设置为 20,其他点设置为 1。

#### 3. 感知级别损失

尽管使用低级别信息来衡量图像差异通常会产生不错的结果,但仅使用它们可能会导致基于卷积神经网络的 3D 面部重建出现局部最小值问题,使重建的 3D 形状不够准确。为了解决这个问题,Deep3D 引入了一个感知级别的损失来进一步引导训练,寻找一个来自预训练深度人脸识别网络的弱监督信号,即提取图

像的深度特征并计算余弦距离:

$$L_{per}(\mathbf{x}) = 1 - \frac{\langle f(I), f(I'(\mathbf{x})) \rangle}{\|f(I)\| \cdot \|f(I'(\mathbf{x}))\|} \quad (2.6)$$

其中  $f(\cdot)$  表示深度特征编码,  $\langle \cdot, \cdot \rangle$  表示向量内积。在此工作中, Deep3D 使用一个内部人脸识别数据集训练了 FaceNet<sup>[33]</sup> 网络结构, 该数据集包含了从互联网上爬取的 50K 个身份的 3M 张人脸图像, 并将其用作 Deep3D 的深度特征提取器。

这些弱监督学习人脸重建方法利用从预训练模型检测到的面部关键点和面部掩膜作为弱监督信号来计算损失函数。广泛使用的关键点检测工具如 FAN<sup>[19]</sup>, 对于给定一个人脸图像, 能够预测 68 个关键点。然而, 有限的关键点数量和预测的准确性限制了网络重建高精度的面部形状的能力。于是本文设计了一种新颖的半监督混合关键点损失, 不同于已有的稀疏关键点损失, 这一损失函数使得 PR3D 能够从自然场景人脸图片以及有 3D 标记的人脸图片中更有效地学习, 减少了基于学习的形状重建方法对真实 3D 扫描数据的依赖。

### 2.2.2 监督学习人脸重建

Wood 等人<sup>[16]</sup>提出 Dense landmarks, 该方法使用合成的 2D-3D 数据集训练网络。然而, 合成和真实面部图像之间存在领域差距, 导致性能次优。Zielonka 等人<sup>[17]</sup>通过统一现有的小型和中型数据集, 构建了 3D 监督数据集, 并关注于可度量面部重建, 然而无法重建人脸的姿态和表情。

如图 2.3 所示, Dense landmarks 包括两个阶段: 首先, 使用传统的卷积神经网络 (CNN) 预测密集的二维关键点  $L$ 。然后, 通过最小化能量函数  $E(\Phi; L)$ , 将参数为  $\Phi$  的三维人脸模型拟合到密集二维关键点上。图像本身不是这个优化过程的一部分, 唯一使用的数据是二维关键点。

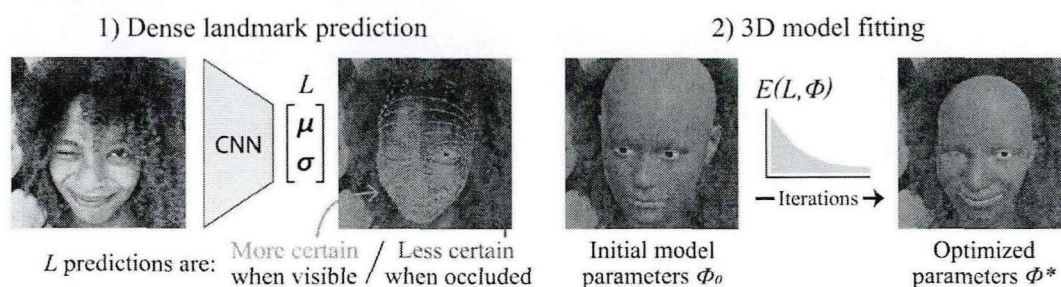


图 2.3 Dense landmarks 重建框架<sup>[16]</sup>

### 1. 预测密集关键点

将每个关键点预测为具有圆形二维高斯概率密度函数的随机变量。因此,  $L_i = \{\mu_i, \sigma_i\}$ , 其中  $\mu_i = [x_i, y_i]$  是该关键点的预期位置,  $\sigma_i$  (标准差) 是不确定性的度量。训练数据包括关键点位置的标签  $\mu'_i = [x'_i, y'_i]$ , 但不包括  $\sigma$ 。网络学习以无监督方式输出  $\sigma$ , 以表明对某些关键点 (例如, 面部正面可见的关键点) 具有确定性, 而对其他关键点 (例如被头发遮挡的关键点) 具有不确定性。这是通过使用高斯负对数似然 (GNLL) 损失进行网络训练来实现的<sup>[34]</sup>:

$$\text{Loss}(L) = \sum_{i=1}^{|L|} \lambda_i \left( \underbrace{\log(\sigma_i^2)}_{\text{Loss}_\sigma} + \underbrace{\frac{\|\mu_i - \mu'_i\|^2}{2\sigma_i^2}}_{\text{Loss}_\mu} \right) \quad (2.7)$$

$\text{Loss}_\sigma$  惩罚网络过度不确定,  $\text{Loss}_\mu$  惩罚网络不准确。 $\lambda_i$  是针对面部特定部位的每个关键点权重, 这是训练过程中唯一使用的损失函数。

### 2. 拟合三维人脸模型

给定密集的二维关键点  $L$ , 第二步的目标是找到使以下能量最小化的最佳模型参数  $\Phi^*$

$$E(\Phi; L) = \underbrace{E_{\text{landmarks}}}_{\text{Data term}} + \underbrace{E_{\text{identity}} + E_{\text{expression}} + E_{\text{joints}} + E_{\text{temporal}} + E_{\text{intersect}}}_{\text{Regularizers}} \quad (2.8)$$

$E_{\text{landmarks}}$  是唯一鼓励三维人脸模型解释预测的二维关键点的项。其他项使用先验知识来正则化拟合。

参数  $\Phi$  被优化以最小化  $E$ 。感兴趣的参数主要控制面部, 但如果相机参数未知会优化它们。

$$\Phi = \underbrace{\{\beta, \Psi_{F \times |\Psi|}, \Theta_{F \times |\Theta|}\}}_{\text{面部}} \underbrace{\{R_{C \times 3}, T_{C \times 3}, f_C\}}_{\text{相机}} \quad (2.9)$$

其中面部身份  $\beta$  在一系列  $F$  帧中共享, 但表情  $\Psi$  和姿态  $\Theta$  每帧都会变化。对于  $C$  中每个相机, 有包括旋转  $R$  和平移  $T$  的六个自由度, 以及一个单独的焦距参数  $f$ , 在单目情况下则只优化焦距。

$E_{\text{landmarks}}$  鼓励三维人脸模型解释预测的二维关键点:

$$E_{\text{landmarks}} = \sum_{i,j,k}^{F,C,|L|} \frac{\|x_{ijk} - \mu_{ijk}\|^2}{2\sigma_{ijk}^2} \quad (2.10)$$

其中, 对于第  $i$  帧中由第  $j$  个摄像头观察到的第  $k$  个关键点,  $[\mu_{ijk}, \sigma_{ijk}]$  是由卷积神经网络预测的密集关键点的二维位置和不确定性, 而  $x_{ijk} = \Pi_j X_j \mathcal{M}(\beta, \psi_i, \theta_i)_k$  则是三维人脸模型上相应关键点的二维投影。



这种监督学习的方法需要通过合成数据集来提供监督信息，然而，合成人脸图像和真实人脸图像之间存在域差距，导致性能不佳。于是本文使用自然场景人脸图像与有 3D 标记的数据集进行半监督训练，以提升模型对于各种实际场景人脸重建的鲁棒性。

### 2.2.3 半监督学习人脸重建

陈卓等人<sup>[20]</sup>提出了一种基于 Transformer 的半监督方法，其网络由用于跨域人脸合成的 cGAN<sup>[35]</sup>和用于重建的 Transformer<sup>[21]</sup>组成，首先使用 cGAN 将真实的人脸图像转换为特定的渲染风格，再将其送入三维人脸重建网络。如图 2.4 所示，其训练数据包含有 3D 标记的人脸图片、合成人脸图片以及自然场景人脸图片。其中有 3D 标记人脸图片和合成人脸图片具有 3D 信息作为监督，而自然场景人脸图片无对应 3D 信息。该方法的流程包含合成人脸生成，跨域人脸生成以及基于 Transformer 的人脸重建。

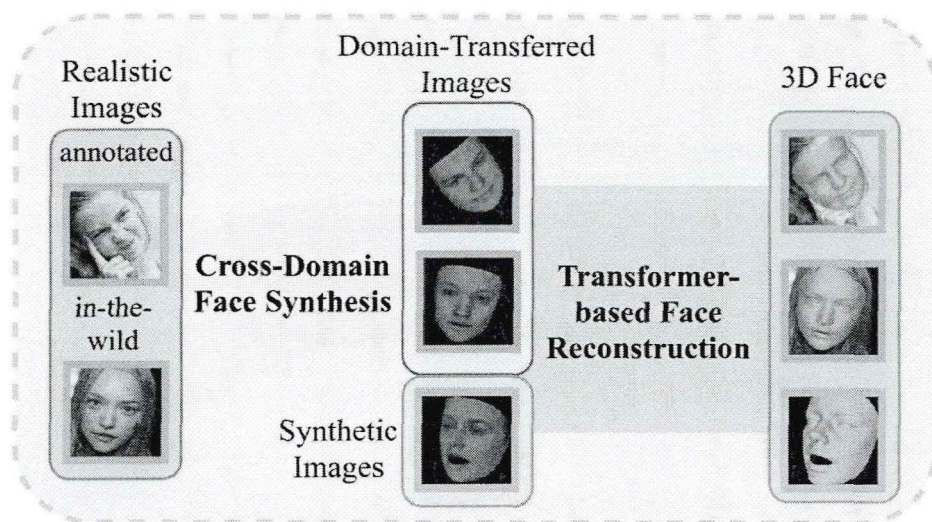


图 2.4 半监督学习人脸重建框架<sup>[20]</sup>

#### 1. 合成人脸图片生成

合成的人脸图像可用于扩展训练数据集，用于域转移 cGAN 和三维人脸重建网络的训练。作者使用 BFM 三维人脸模型生成合成三维人脸  $S \in \mathbb{R}^{3N}$ 。为了扩展训练数据，合成的人脸图像应该在形状、表情、姿态这些基于单幅图像的三维人脸重建的目标参数上有足够的变化。作者通过改变形状和表情的 3DMM 系数随机生成面部几何形状。对于姿态参数，统一采样欧拉角度范围从  $[-90^\circ, 90^\circ]$  俯仰， $[-45^\circ, 45^\circ]$  的偏航和  $[-10^\circ, 10^\circ]$  的滚动。对于纹理和渲染设置，使用固定的 3DMM 纹理系数，并使用相同的环境、漫反射和镜面光照来渲染所有合成人



脸模型, 此外, 将渲染背景颜色设置为全黑, 以在域转移过程过滤掉无用的信息和复杂的背景和光照条件的影响, 从而专注于三维人脸几何的重建。

## 2. 跨域人脸图片生成

跨域人脸图片生成由生成器  $G$  和鉴别器  $D$  组成的域转移 cGAN 实现。生成器  $G$  的目标是学习从真实人脸图像  $r$  到合成人脸图像  $s$  的映射。鉴别器  $D$  的目的是区分渲染合成图像和域转移图像。对于具有三维监督信息的数据集, 训练数据以图像对集合  $(r_i, s_i)$  的形式表示, 其中  $r_i$  为真实人脸图像,  $s_i$  为其对应 3D 信息生成的合成图像。对抗性目标函数为:

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{(r,s)}[\log D(r, s)] + \mathbb{E}_{(r)}[\log(1 - D(r, G(r)))] \quad (2.11)$$

此外, 还采用 L1 损失来最小化输出图像与真实值之间的差异:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{(r,s)}[\|s - G(r)\|_1] \quad (2.12)$$

## 3. 基于 Transformer 的人脸重建

经过跨域人脸图片生成, 无论是自然场景的还是三维标记的真实图像, 都转化为了与合成图像具有相同风格的图像。因此, 基于 Transformer 的三维人脸重建网络可以将各种不存在域间隙的不同人脸图像作为输入, 并预测网格顶点的三维坐标。

对于合成人脸和经过域转移的三维标记人脸图像, 在 Transformer 输出上应用 L1 损失, 以最小化预测值  $\bar{V}_{3D}$  与真实值  $V_G \in \mathbb{R}^{M \times 3}$  之间的误差:

$$L_{vertex}(\bar{V}_{3D}, V_G) = \frac{1}{M} \sum_{i=1}^M \|\bar{V}_{3D}^i, V_G^i\|_1 \quad (2.13)$$

对于自然场景人脸图像, 首先将其域转换为合成样式, 由于合成样式有固定的渲染参数和相同的背景, 图像之间的光度不一致主要是由人脸的几何形状引起的, 这意味着可以利用从 Transformer 输出中获得的输入图像和渲染图像之间的光度损失来自监督三维人脸重建, 即以具有重投影一致性的自监督方式约束重建网络。首先投影重建的人脸, 然后最小化域转移图像像素值和重投影图像像素值之间的 L1 损失。重投影一致性损失为:

$$L_{reproj} = \|I(\bar{V}_{3D}) - I_{in}\|_1 \quad (2.14)$$

其中  $I(\bar{V}_{3D})$  是输出的 3D 人脸的渲染图像,  $I_{in}$  是输入图像。

这种直接回归 3D 顶点的方法有着比基于三维人脸模型的重建方法更强大的表达能力, 但是相应的也需要更多的训练数据及资源消耗, 因为三维人脸模型作为一个统计模型, 本身就包含了人脸的先验知识。并且该方法在各种人脸重建

的测试基准上表现不佳，并没有展现出优越的性能。本文一方面借鉴了该方法半监督学习的思路，另一方面通过先进的三维人脸模型提供先验知识，以实现最先进的形状重建精度。

### 2.2.4 实时重建

现有的大多数 3D 人脸重建方法都专注于准确性，而忽视了速度，这限制了它们的实际应用。例如，在白浩然等人<sup>[10]</sup>的工作中，每张图像的总拟合时间在 NVIDIA Tesla V100 GPU 上测试约为 60 秒。郭建珠等人<sup>[22]</sup>提出了一个回归框架 3DDFA-V2，它在轻量化骨架的基础上，提出了一种元联合优化策略，对一组少量的 3DMM 参数进行动态回归，大大提高了速度和精度。以单张图像作为输入，3DDFA-V2 约需 7.2 毫秒完成重建，并实现了良好的准确性。

Wood 等人<sup>[16]</sup>通过预测 320 个关键点然后拟合 3D 人脸模型，也实现了实时重建。与以往的方法不同，本研究的方法利用 PR3D 获得的知识来指导轻量级模型 fast-PR3D 的训练。这种独特的策略在准确性和速度上都表现出优越性，使 fast-PR3D 成为一种高效且精确的模型。

## 2.3 基于单幅人脸图像的纹理重建

### 2.3.1 线性纹理空间

大多数三维人脸模型都有自己的线性纹理空间，以 FLAME<sup>[9]</sup>为例，它从自然场景人脸图像中构建纹理空间，以涵盖各种族、年龄段等广泛范围。FLAME 从 FFHQ 数据集中随机选择了 1500 张图像来构建纹理空间，如图 2.5 所示。具体步骤如下：

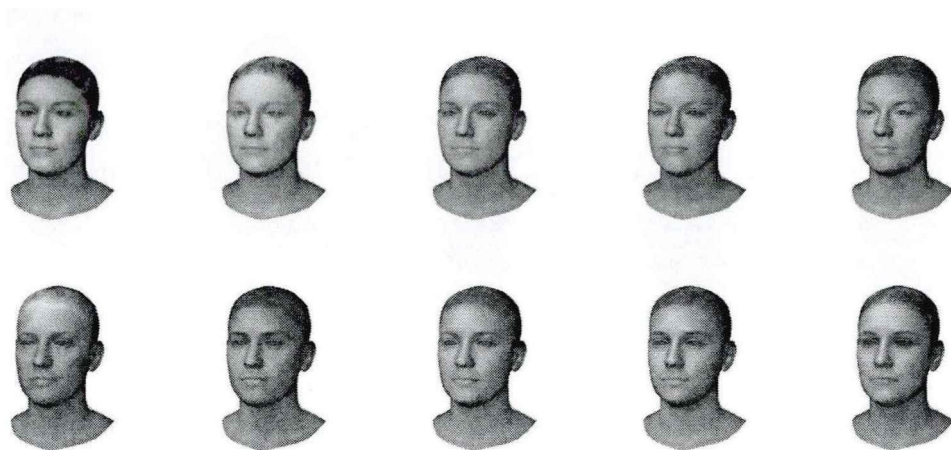


图 2.5 FLAME 线性纹理空间<sup>[9]</sup>

### 1. 初始化

给定一个纹理空间，可以以“analysis by synthesis”的方式将 3D 人脸模型拟合到图像中，然后这些拟合结果可以用于构建纹理空间。为了获得初始纹理空间，将 FLAME 拟合到 BFM 模板，并将 BFM 顶点颜色投影到 FLAME 网格上，以获得初始纹理基底。

### 2. 模型拟合

然后，将 FLAME 拟合到 FFHQ 图像中，优化 FLAME 形状、姿态和表情参数，初始纹理空间的参数，球谐光照的参数（仅优化 9 个球谐系数，共享所有三个颜色通道），以及用于捕捉与初始纹理空间不符的纹理细节的纹理偏移参数。拟合过程最小化了关键点损失、光度损失，以及用于形状、姿态、表情、外观和纹理偏移的正则化项。

关键点损失最小化了从人脸模型面部投影的关键点与预测的 2D 关键点之间的差异（使用 FAN 关键点预测器<sup>[19]</sup>预测）。光度损失仅针对皮肤区域进行优化（由面部分割网络<sup>[36]</sup>提供），以增强对面部遮挡的鲁棒性。

### 3. 纹理补全

模型拟合后，计算的纹理偏移捕捉了每个图像非遮挡皮肤区域的面部外观。为了完成纹理贴图，通过向可见面部区域添加随机笔画（即随机大小和位置的笔画）来逐渐训练自 GMCNN<sup>[37]</sup>调整的一个修补网络，并学习对这些笔画进行补全。训练完成后，使用训练好的修补网络对所有缺失区域进行补全。

### 4. 纹理空间计算

在完成这 1500 个纹理贴图后，使用主成分分析（PCA）来计算一个线性纹理空间。

尽管这种线性纹理空间高度适配于相应的三维人脸模型，但是其纹理质量和保真度不高。从视觉效果来看，基于线性纹理空间的纹理重建和原图有着非常明显的差异。

## 2.3.2 基于 GAN 的纹理空间

为了提升纹理重建的质量与保真度，FFHQ-UV<sup>[10]</sup>首先构建了一个超过 50,000 个纹理贴图的数据集，在此基础上训练了一个基于 GAN 的纹理解码器。

### 1. 构建纹理贴图数据集

数据集创建流程包括三个步骤：人脸图像编辑、人脸 UV 纹理提取和 UV 纹理校正与补全。作者将上述步骤应用于 FFHQ 数据集<sup>[38]</sup>中的所有图像，该数据集包括 70,000 张年龄和种族差异很大的高质量人脸图像。使用基于人脸解析结果的自动过滤器排除无法归一化的面部遮挡图像，最终得到的 UV 贴图经过人工检查和过滤，得到 54,165 张分辨率为  $1024 \times 1024$  的高质量 UV 图，称之为

FFHQ-UV 数据集。

## 2. 基于 GAN 的纹理解码器

首先在上一步构建的 FFHQ-UV 数据集上训练基于 GAN<sup>[25]</sup>的纹理解码器。一张 UV 纹理贴图由  $\tilde{T} = G_{tex}(z)$  生成, 其中  $z \in \mathcal{Z}$  表示  $\mathcal{Z}$  空间中的潜码,  $G_{tex}(\cdot)$  表示纹理解码器。在 3D 人脸重建过程中 UV 纹理恢复的目标是通过纹理解码器  $G_{tex}(\cdot)$  为输入图像找到产生最佳纹理 UV 贴图的潜在代码  $z^*$ 。

三维人脸重建算法包括以下三个阶段: 线性 3DMM 初始化、纹理潜码  $z$  优化和联合参数优化。作者使用了基于 PCA 的形状基底 HIFI3D++<sup>[39]</sup>。关于线性 3DMM 初始化, 使用经过形状基底 HiFi3D++ 训练的 Deep3D 来初始化重建。给定单张输入人脸图像  $I_{in}$ , 预测参数为  $\{p_{id}, p_{exp}, p_{tex}, p_{pose}, p_{light}\}$ , 其中  $p_{id}$  和  $p_{exp}$  分别为 HIFI3D++ 的身份和表情系数,  $p_{tex}$  为纹理系数,  $p_{pose}$  为头部姿态系数,  $p_{light}$  为 SH 光照系数。第二阶段为纹理潜码  $z$  优化, 固定上一阶段初始化的参数  $\{p_{id}, p_{exp}, p_{pose}, p_{light}\}$ , 来寻找到一个能最小化下列损失函数的潜在代码  $z \in \mathcal{Z}$

$$\mathcal{L}_{s2} = \lambda_{lips} \mathcal{L}_{lips} + \lambda_{pix} \mathcal{L}_{pix} + \lambda_{id} \mathcal{L}_{id} + \lambda_{reg}^z \mathcal{L}_{reg}^z \quad (2.15)$$

其中  $\mathcal{L}_{lips}$  为输入图像  $I_{in}$  与被渲染后图像  $I_{re}$  之间的 LIPS 距离,  $\mathcal{L}_{pix}$  是在人脸解析模型预测的人脸区域上计算的  $I_{in}$  和  $I_{re}$  之间的  $L_2$  光度误差,  $\mathcal{L}_{id}$  是基于 Arcface<sup>[40]</sup> 的最后一层特征向量的身份损失,  $\mathcal{L}_{reg}^z$  是关于纹理潜码  $z$  的正则化项。同时, 将潜在码  $z$  约束在  $z' = \sqrt{d}S^{d-1}$  定义的超球上, 以鼓励真实的纹理映射, 其中  $S^{d-1}$  是  $d$  维欧几里德空间中的单位球。

在第三阶段联合参数优化中, 放松了对潜码  $z$  的超球面约束以获得更强的表达能力, 并通过最小化以下损失函数来共同优化参数  $\{z, p_{id}, p_{exp}, p_{pose}, p_{light}\}$

$$\mathcal{L}_{s3} = \lambda_{pix} \mathcal{L}_{pix} + \lambda_{id} \mathcal{L}_{id} + \lambda_{reg}^z \mathcal{L}_{reg}^z + \lambda_{lm} \mathcal{L}_{lm} + \lambda_{reg}^{id} \mathcal{L}_{reg}^{id} + \lambda_{reg}^{exp} \mathcal{L}_{reg}^{exp} \quad (2.16)$$

其中  $\mathcal{L}_{lm}$  为关键点检测器<sup>[19]</sup>得到的二维关键点损失,  $\mathcal{L}_{reg}^{id}$  和  $\mathcal{L}_{reg}^{exp}$  是系数  $p_{id}$  和  $p_{exp}$  的正则化项。

这种基于 GAN 构造的纹理空间相较于线性纹理空间有着更强的表达能力, 但其效果仍受到训练数据集的限制, 并且基于 GAN 的迭代优化的方法时间消耗较久, 这限制了其实际应用的可行性。

### 2.3.3 原图提取

在完成形状重建的基础上, 还可以从原图提取纹理的方式得到重建的纹理贴图, DECA<sup>[15]</sup>就采用了这种方法。因为在完成了形状重建之后, 可以得到 3D 人脸的形状、表情、姿态及相机参数等, 所以可以将 3D 人脸和人脸图片进行对齐, 确保它们在相同的空间位置和朝向上。一旦对齐完成, 就可以将重建的

3D 人脸投影到对应的图像平面上,并根据每个像素点在图像中的位置确定其在 3D 人脸上的对应位置。这个过程称为纹理坐标映射,它确定了每个像素点在 3D 模型上的纹理坐标。然后通过使用纹理坐标,可以从对齐的人脸图像中采样纹理信息,并将其映射到重建的 3D 人脸上。这意味着从图像中提取的颜色信息将被应用到 3D 模型的相应部位上。重建的 3D 人脸模型上的每个顶点都被赋予从对齐图像中提取的纹理信息。这些纹理信息通常以纹理贴图的形式保存,其中每个像素点的颜色值对应于模型上相应顶点的颜色值。最后,可以通过优化算法对纹理贴图进行进一步优化,以提高质量并填补可能存在的缺失区域。

基于原图提取的重建方法避免了基于构造的纹理重建方法受到的数据集的限制,但当输入人脸图片存在因为头发、饰品、姿态等引起的遮挡时会造成纹理缺失等问题。本文提出了基于 StyleGAN2 风格空间人脸重演的纹理提取方法。通过在不同身份、姿态等条件下人脸图片作为输入时与其他前沿方法的对比,表明本研究方法一方面能够解决由于缺乏足够 3D 扫描数据造成的保真度低的问题,另一方面能够有效地解决输入图像中由于自遮挡等因素造成的纹理缺失等问题,生成高保真的重建纹理贴图。

## 2.4 本章小结

本章主要对本文的相关工作进行了介绍。

在形状重建方面,介绍了现在常用的三维人脸模型,包括 BFM, FLAME。随着神经网络的发展,可以训练网络回归三维人脸模型参数。深度神经网络的训练高度依赖大量带有精确标记的 3D 真实扫描数据,而现实中获取这类数据的难度和成本较高,为解决此限制,出现了弱监督训练、合成数据监督训练和半监督训练等方法。本文对这些相关的具体方法分别进行了介绍。

在纹理重建方面。BFM 和 FLAME 模型均有线性纹理空间,并可以使用回归网络预测相应的纹理参数。为了提升纹理空间的表达能力,出现了基于 GAN 的非线性纹理空间。此外,还有基于提取的方法从原图提取纹理实现纹理重建的方法。本文对上述相关的具体方法分别进行了介绍。



## 第3章 基于半监督学习的高精度形状重建算法

本文针对 3D 扫描数据的缺乏, 以及弱监督信号精度及数量不足的问题, 设计了基于半监督学习的高精度形状算法。本章首先在 3.1 节简要介绍了本文使用的三维人脸模型、相机模型以及光照模型等相关模型, 然后在 3.2 节介绍本文的基于半监督学习的高精度形状重建算法设计动机及流程, 详细介绍辅助编码器、形状编码器以及联合优化过程, 随后在 3.3 节通过在多个形状重建基准上的定性和定量对比实验证明本文提出的形状重建方法的高精度, 最后介绍本文在高精度模型基础上蒸馏出的实时轻量级形状重建模型。

### 3.1 相关模型

#### 3.1.1 三维人脸模型

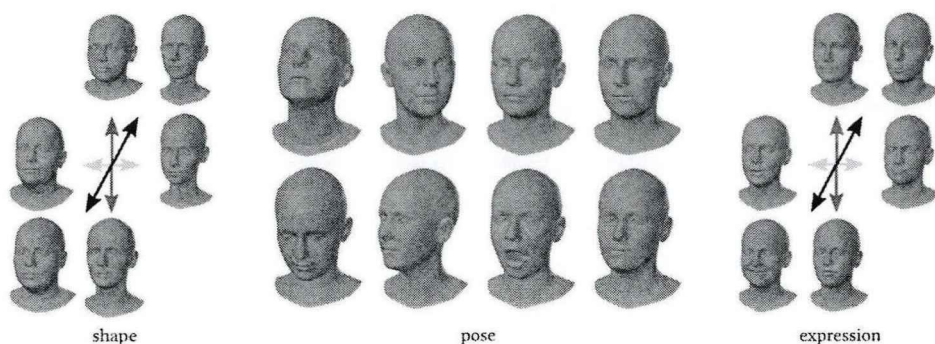


图 3.1 FLAME 人脸模型示意图<sup>[9]</sup>

人脸具有复杂的结构, 包括鼻子、眼睛、嘴巴等丰富的细节信息, 使得直接通过卷积神经网络回归单张人脸图像对应的准确三维人脸变得困难。目前, 由于缺乏带有详细三维人脸标签的大规模数据集, 这一挑战变得更加严峻。为了解决这个问题, 人们普遍采用三维人脸模型进行人脸重建。这种方法不仅确保了在人脸重建过程中不会产生不符合人脸结构的情况, 同时还能够建立不同脸型之间的对应关系。综合而言, 使用三维形变模型是目前处理复杂人脸数据并进行三维人脸重建的有效策略。在众多参数化的 3D 面部模型中, 例如 BFM<sup>[8]</sup>、FLAME<sup>[9]</sup>和 HiFi3D++<sup>[39]</sup>, 本研究选择了 FLAME 作为提供几何先验的模型, 而不是常用的 BFM, 原因如下:

- FLAME 被设计为整个人头的综合模型, 与主要关注面部结构的 BFM 形成对比。在视觉上, FLAME 的输出包括完整的三维头部结构, 包括枕骨和颈

部，这是 BFM 模型中缺少的独特特征。

- FLAME 通过线性混合蒙皮 (LBS) 明确建模了颈部旋转和眼睛旋转的能力，而 BFM 则缺乏这一能力。
- 与 BFM 模型相比，FLAME 在表示各种面部表情方面具有更广泛的能力。原始的 BFM 模型是基于仅有 200 个个体的数据构建的，而 FLAME 模型是使用来自 33,000 个个体头部的数据开发的。

FLAME 具有  $N = 5023$  个顶点， $K = 4$  个关节（颈部、下颌和眼球），并由一个函数  $M(\beta, \theta, \psi) : \mathbb{R}^{|\beta| \times |\theta| \times |\psi|} \rightarrow \mathbb{R}^{3N}$  定义。该函数接受描述形状  $\beta \in \mathbb{R}^{|\beta|}$ 、姿势  $\theta \in \mathbb{R}^{3K+3}$  和表情  $\psi \in \mathbb{R}^{|\psi|}$  的系数，并返回  $N$  个顶点。FLAME 的纹理空间是从 BFM<sup>[8]</sup> 转换而来，由函数  $A(\alpha) : \mathbb{R}^{\alpha} \rightarrow \mathbb{R}^{d \times d \times 3}$  定义，其中反照率参数  $\alpha \in \mathbb{R}^{|\alpha|}$ 。

### 3.1.2 相机模型

现有自然场景人脸数据集中的照片通常是从远处拍摄的。因此，本研究使用正交相机模型  $c$  将 3D 网格投影到图像空间中。

正交相机模型是一种简化的摄影模型，假设光线在穿过透镜后是平行的，从而使投影过程更加直观和可控。这个模型在计算机图形学和计算机视觉中广泛应用，特别是在需要简化投影计算的情境下。正交相机模型基于平行投影的原理，即相机的投影光线是平行的。这样的简化假设使得相机模型的数学推导更为直观。3D 面部顶点投影到图像中为

$$v = s\Pi(M_i) + t \quad (3.1)$$

其中， $M_i \in \mathbb{R}^3$  表示  $M$  中的一个顶点， $\Pi \in \mathbb{R}^{2 \times 3}$  是正交的三维到二维投影矩阵， $s \in \mathbb{R}$  和  $t \in \mathbb{R}^2$  分别表示各向同性的缩放和二维平移，参数  $s$  和  $t$  构成相机模型  $c$ 。投影矩阵表示为：

$$P = \begin{bmatrix} s & 0 & 0 & t_x \\ 0 & s & 0 & t_y \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

### 3.1.3 光照模型

人脸重建任务中的光照模型是至关重要的，因为照明条件的变化会直接影响图像中物体表面的外观。在实际拍摄中，由于不同环境和拍摄条件，人脸图像可能面临各种光照情况，包括阴影、高光和其他照明效果。为了准确地还原人脸的三维形状，需要考虑这些光照因素，因为它们提供了有关人脸表面细节和几何

特征的重要信息。因此,光照模型在人脸重建任务中是不可或缺的,能够帮助提高对人脸外观的还原准确性。

对于人脸重建,最常用的照明模型是基于球面谐波 (Spherical Harmonics)<sup>[41]</sup>的模型。假设光源是遥远的,并且面部表面的反射是 Lambertian 反射,即无论观察者的视角如何,表面对观察者表现的亮度都是相同的。着色的面部图像被计算为:

$$B(\alpha, l, N_{uv})_{i,j} = A(\alpha)_{i,j} \odot \sum_{k=1}^9 l_k H_k(N_{i,j}) \quad (3.2)$$

其中反照率  $A$ 、表面法线  $N$  和着色的纹理  $B$  在 UV 坐标系中表示,其中  $B_{i,j} \in \mathbb{R}^3$ ,  $A_{i,j} \in \mathbb{R}^3$ ,  $N_{i,j} \in \mathbb{R}^3$  表示 UV 坐标系中像素  $(i, j)$ , SH 基函数和系数定义为  $H_k: \mathbb{R}^3 \rightarrow \mathbb{R}$  和  $l = [l_1^T, \dots, l_9^T]^T$ , 其中  $l_k \in \mathbb{R}^3$ ,  $\odot$  表示 Hadamard 积。

### 3.1.4 可微渲染

可微渲染是一种新兴的图形渲染技术,它结合了深度学习和传统的图形学方法。与传统的渲染技术不同,可微渲染将整个渲染过程表述为可微分的函数,使得渲染过程中的每个步骤都能够被导数化。这意味着可以直接在渲染过程中应用梯度下降等优化方法,从而实现对渲染结果的端到端优化。这种方法不仅可以用于生成逼真的图像,还可以用于图像编辑、图像合成和三维重建等任务。

PyTorch3D<sup>[42]</sup>是一个基于 PyTorch 框架的深度学习库,专门用于处理三维图形。PyTorch3D 提供了一系列可微渲染的工具和模块,使得用户能够轻松地构建和训练可微分的渲染模型。这些模型包括可微分的相机模型、光照模型和材质模型,能够实现高效、灵活的图形渲染,并且可以直接集成到深度学习模型中进行端到端的训练。PyTorch3D 的出现极大地推动了可微渲染技术在深度学习领域的发展,并为图形生成、三维重建和虚拟现实等领域提供了新的研究思路和解决方案。

### 3.1.5 残差神经网络

ResNet-50<sup>[43]</sup>是一种经典的深度卷积神经网络模型,由微软研究院提出。作为 ResNet 系列的一部分,ResNet-50 在图像分类、目标检测和语义分割等计算机视觉任务中表现出色,并在多个国际竞赛中取得了优异的成绩。

ResNet-50 的核心创新是引入了残差学习模块,通过跨层连接(即跳跃连接)将输入直接添加到神经网络的某一层输出,从而解决了深度卷积神经网络训练过程中的梯度消失和梯度爆炸问题。这种设计使得网络更易于训练,能够训练更深更复杂的模型,并且在一定程度上提高了模型的准确性。

ResNet-50 具有 50 层的深度,其中包含了多个残差块 (Residual Blocks)。每

个残差块由几个卷积层和跳跃连接组成，可以有效地学习输入和输出之间的残差信息，从而提高了模型的特征提取能力和泛化能力。

由于其卓越的性能和可训练性，ResNet-50 已成为图像分类领域的重要基准模型之一，并广泛应用于各种计算机视觉任务和实际应用中。其在训练和推断过程中的高效性和稳定性，使其成为研究和工程实践中不可或缺的工具。

## 3.2 算法流程

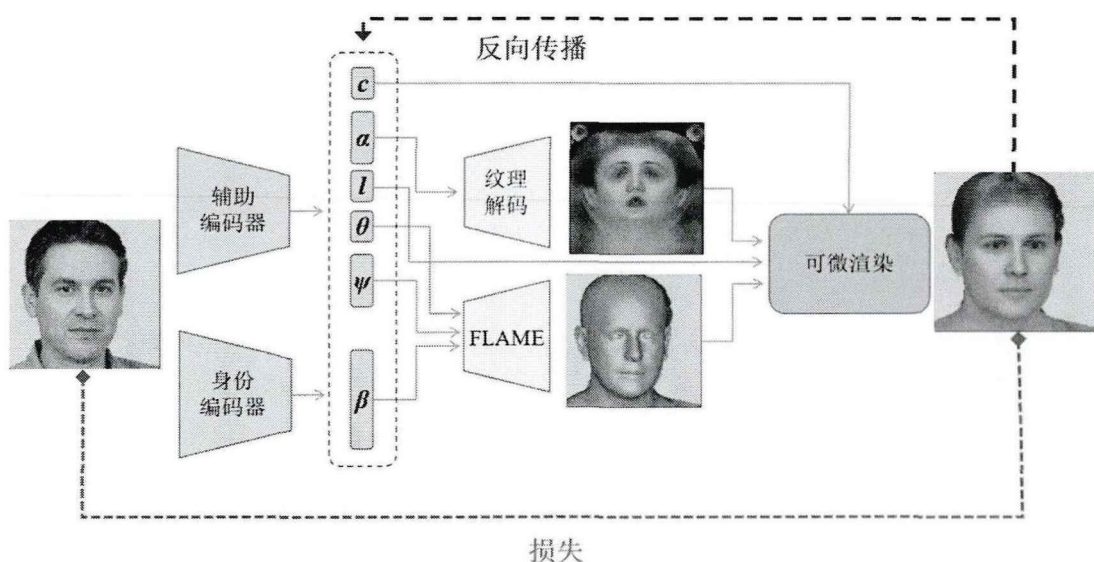


图 3.2 本文形状重建框架

本文提出了一种高精度的形状重建方法，如图 3.2 所示。它包含两个编码器，辅助编码器和身份编码器。辅助编码器负责编码姿态参数  $\theta$ 、表情参数  $\psi$ 、相机参数  $c$ 、光照参数  $l$  和外观参数  $\alpha$ ，身份编码器负责编码形状参数  $\beta$ 。本设计的动机如下：

对于同一身份的图像，姿势、配饰、光照和表情等因素不会影响它们的中性形状。以前的工作如 Deep3D<sup>[14]</sup> 和 DECA<sup>[15]</sup> 使用单个编码器来回归所有参数，旨在将重建的 3D 人脸与输入图像（无论是在关键点上还是在渲染像素上）尽可能对齐。然而，它们忽视了身份特征在确定 3D 形状中的关键作用，往往导致同一身份的不同照片的重建结果差异显著。以 DECA<sup>[15]</sup> 为例，如图 3.3 所示，当使用同一身份的不同照片作为输入时，重建结果的姿态和表情参数被置为零，仅留下 FLAME 的形状参数。理论上，这应该产生几乎相同的 3D 人脸，但实际结果显示出显著差异。将第一张图像（正面视图图像）的重建结果作为参考，计算每个重建结果与其相应参考点的距离，并根据在图 3.3 中使用的 plasma 颜色映

射将不同的顶点分配给不同的颜色。Plasma 渐变条是一种常用于数据可视化的颜色映射方法。它的颜色从深紫色逐渐过渡到明黄色，呈现出明亮而生动的色彩，特点是颜色变化平滑，能够清晰地显示数据的变化趋势。当 plasma 颜色映射从深紫色到明黄色变化时，表示与参考形状的差异越来越大。本研究还计算了面部各个区域的平均距离，如表 3.1 所示，其中输入序号按从左到右，从上到下的输入图片顺序确定。

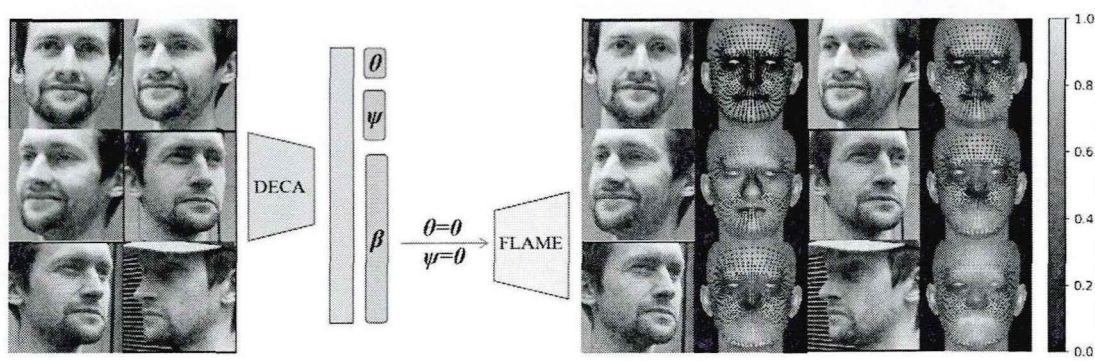


图 3.3 本文形状重建设计动机示意图

表 3.1 各个面部区域到参考人脸（正面视图图像）的距离

| 输入 | 眼部   | 前额   | 鼻子   | 嘴唇   | 面部整体 |
|----|------|------|------|------|------|
| 1  | 0.30 | 0.23 | 0.45 | 0.22 | 0.38 |
| 2  | 0.64 | 0.72 | 0.70 | 0.54 | 0.62 |
| 3  | 1.20 | 0.66 | 0.79 | 1.77 | 1.23 |
| 4  | 1.25 | 0.58 | 0.84 | 2.29 | 1.35 |
| 5  | 1.98 | 2.25 | 1.81 | 3.46 | 2.14 |

受此启发，本文使用单独的身份编码器来重建中性的 3D 面部形状，同时将姿势和表情等信息视为辅助数据，以微调 3D 形状。这确保了与输入图像更好地对齐，同时保持对身份的忠实，从而实现了最先进的结果。在具体设计如图 3.2 所示。本文形状重建的创新点可概括为以下三个方面：

- 本文提出了一种基于半监督学习的高精度形状重建框架，使用自然场景人脸图片及有 3D 标记的人脸图片训练辅助编码器和身份编码器，将输入图像编码为 FLAME 参数，并设计联合优化过程通过最小化基于可微渲染的能量函数进一步优化这些参数。这一框架的设计显著提高了从单幅人脸图像进行 3D 形状重建的精度和鲁棒性。
- 不同于已有的稀疏关键点损失，本文提出了一种新颖的混合关键点损失函数，旨在从自然场景人脸图片以及有 3D 标记的人脸图片中学习，从而有效减轻了基于学习的形状重建方法对真实 3D 扫描数据的依赖。



- 采用师生学习蒸馏得到轻量级形状重建模型，在保持较高精度的同时，显著提升了重建速度，每张单独图像的推断时间约为3毫秒，为实时人脸重建任务提供了可行的解决方案。

### 3.2.1 辅助编码器

如图3.4所示，本文使用可微渲染来训练辅助编码器（ResNet-50<sup>[43]</sup>）。本文使用FLAME的前100个形状参数，前50个表情参数和前50个纹理参数。值得注意的是，本文没有使用全部的300个形状参数因为在此的目标是协助其他参数的回归，100个形状参数在后续的步骤会被身份编码器的输出所替代。用于训练的数据集包含用于半监督训练的有3D标记的人脸图片和自然场景人脸图片。神经网络通过反向传播输入图像  $I$  与渲染输出  $I'$  之间的误差来进行更新，其中使用了混合关键点损失  $L_{landmark}$ 、光度损失  $L_{photometric}$ 、感知级别损失  $L_{perception}$  和正则化损失  $L_{reg}$ 。总损失函数为：

$$L = \omega_{lan} L_{landmark} + \omega_{pho} L_{photometric} + \omega_{per} L_{perception} + \omega_{reg} L_{reg} \quad (3.3)$$

在预训练阶段，仅使用自然场景面部图像的关键点损失和正则化损失进行良好的初始化，其中  $\omega_{lan} = 1e^{-4}$ ， $\omega_{reg} = 1e^{-4}$ 。在正式训练期间， $\omega_{lan} = 1.0$ ， $\omega_{pho} = 2.0$ ， $\omega_{per} = 0.2$ ， $\omega_{reg} = 1e^{-4}$ 。与之前邓誉等人<sup>[14]</sup>和Feng等人<sup>[15]</sup>的工作不同，本研究设计的混合关键点损失对于精确的形状重建更为有效。下面解释了每个损失函数的具体含义。

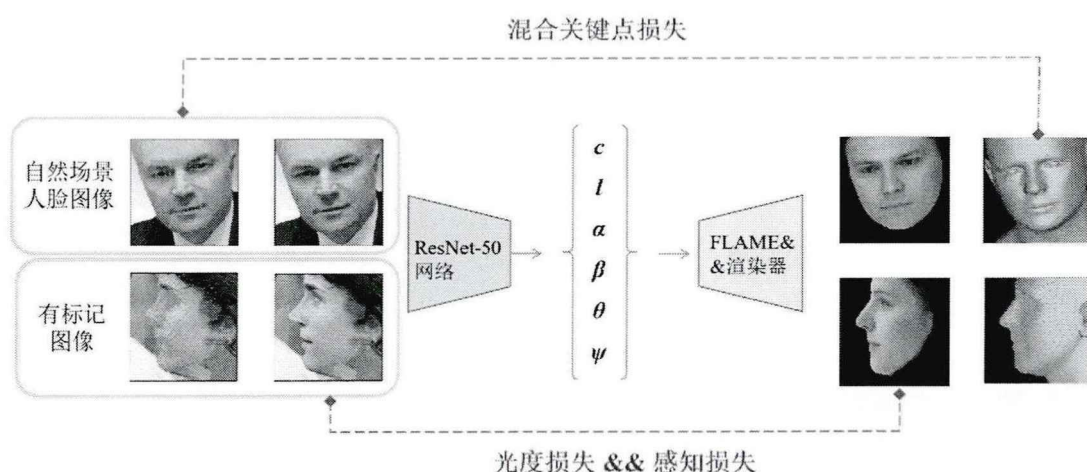


图 3.4 辅助编码器的训练示意图

#### 1. 混合关键点损失

Wood 等人<sup>[16]</sup>证明了在面部分析任务中使用更多关键点的有效性。因此，对于自然场景面部图像，本文不像邓誉等人<sup>[14]</sup>和Feng等人<sup>[15]</sup>的弱监督学习人脸



重建那样使用 FAN<sup>[19]</sup> 检测 68 个标记点。不同之处在于本文使用 122 个关键点 (105 个来自 mediapipe<sup>[44]</sup>, 17 个来自 FAN<sup>[19]</sup>)。对于来自 300W-LP 数据集<sup>[23]</sup> 的包含 3D 标记的图像时, 因为其 3D 信息是用 BFM 表示的, 并且与图像不对齐, 所以并不能直接使用。为解决这个问题, 设计以下步骤:

- 将 BFM 网格转换为 FLAME 拓扑结构。
- 根据检测到的关键点来优化相机模型参数, 从而将 FLAME 网格与图像对齐。
- 将对齐的 FLAME 网格面部 1787 个顶点的 2D 投影作为密集关键点。

关键点损失计算了标记的 2D 面部关键点  $k_i$  与从估计的 FLAME 网格顶点  $M_i$  通过正交相机模型  $\Pi$  投影到 2D 空间的相应关键点之间的 L1 损失。具体而言, 混合关键点损失定义为:

$$L_{landmark} = \lambda \sum_{i=1}^{122} \|k_i - s\Pi(M_i) + t\|_1 + (1 - \lambda) \sum_{j=1}^{1787} \|k_j - s\Pi(M_j) + t\|_1 \quad (3.4)$$

其中, 对于自然场景人脸图像,  $\lambda = 1$ , 对于有 3D 标记的图像,  $\lambda = 0$ 。  $s$  代表缩放,  $t$  代表平移。

## 2. 光度损失

为了增强本研究框架对遮挡和面部配饰 (如眼镜或头发) 的鲁棒性, 本文采用了一个面部解析模型<sup>[45]</sup>。该模型使本文能够获得一个面部掩膜, 表示为  $V_I$ , 其中为面部皮肤区域分配值 1, 其他地方分配值 0。通过使用这个面部掩膜, 可以在重建过程中专注于面部区域。光度损失度量了输入图像  $I$  与重建图像  $I'$  之间的差异。这个损失的计算方式是:

$$L_{photometric} = \|V_I \times (I - I')\|_{1,1} \quad (3.5)$$

## 3. 感知损失

邓誉等人<sup>[14]</sup>发现仅使用上述提到的低级信息可能导致基于 CNN 的 3D 面部重建陷入局部最小值问题, 因此他们引入了来自面部识别网络的弱监督信号。类似地, 本文使用了 Cao 等人<sup>[46]</sup> 的面部识别网络预训练模型, 计算输入图像  $I$  和重建图像  $I'$  的深度特征  $f(\cdot)$  的余弦距离。

$$L_{perception} = 1 - \frac{\langle f(I), f(I') \rangle}{\|f(I)\|_2 \cdot \|f(I')\|_2} \quad (3.6)$$

## 4. 正则化损失

为了避免网络退化, 本研究中的  $L_{reg}$  包括外观  $\|\alpha\|_2^2$ 、形状  $\|\beta\|_2^2$  和表情  $\|\psi\|_2^2$ 。

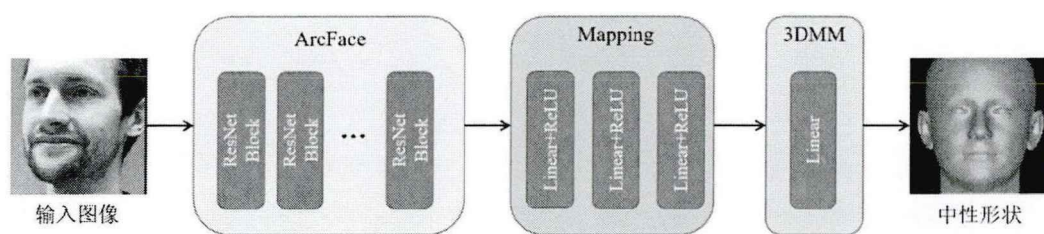


图 3.5 身份编码器结构图

### 3.2.2 身份编码器

如图 3.5 所示, 根据单个输入的 RGB 图像  $I$ , 身份编码器旨在预测呈现中性表情和姿态的人脸形状。于是本文利用了 *Glnt360K*<sup>[47]</sup> 上预训练的 *ArcFace*<sup>[40]</sup> 架构。这个基于 *ResNet100* 的网络使用附加的角度边缘损失在 2D 图像数据上进行了训练, 以获得用于人脸识别的高度区分特征。它对光照、表情、旋转、遮挡和相机参数具有不变性, 这对于鲁棒的形状预测非常理想。本文通过一个小的映射网络  $\mathcal{M}$ <sup>[17]</sup> 扩展了 *ArcFace* 架构, 来将 *ArcFace* 特征映射到 *FLAME* 的形状参数空间中

$$\mathbf{z} = \mathcal{M}(\text{ArcFace}(I)) \quad (3.7)$$

其中  $\mathbf{z} \in \mathbb{R}^{300}$ 。映射网络  $\mathcal{M}$  由三个全连接的线性层组成, 使用 *ReLU* 激活函数, 并带有最终的线性输出层。 $\mathcal{M}$  使用 *MICA* 数据集<sup>[17]</sup> 中成对 2D/3D 数据进行训练。*MICA* 数据集由 8 个较小的数据集组成, 包含 *FLAME* 拓扑下的约 2300 名受试者。在映射网络  $\mathcal{M}$  训练期间固定了预训练的 *ArcFace* 网络的大部分, 只调整最后 3 个 *ResNet* 块, 这是因为 *ArcFace* 是在大量的身份数据上训练的, 对更多层进行微调会导致过拟合, 从而导致预测结果变差。

$$\sum_{(I, \mathcal{G}) \in \mathcal{D}} |\kappa_{mask}(\mathcal{G}_{3DMM}(\mathcal{M}(\text{ArcFace}(I))) - \mathcal{G})| \quad (3.8)$$

其中  $\mathcal{G}$  为真实的 3D 网格,  $\kappa_{mask}$  为区域依赖权重 (面部区域的权重为 150.0, 头部后部为 1.0, 眼睛和耳朵为 0.01)。

### 3.2.3 联合优化

现在, 从身份编码器获得形状编码  $\beta$ , 并从辅助编码器获得姿势编码  $\theta$ 、表情编码  $\psi$ 、相机编码  $c$ 、光照编码  $l$  和外观编码  $\alpha$ 。然后, 本文使用两阶段拟合来优化上述参数。在第一阶段, 通过最小化能量函数来优化  $\Phi_1 = \{\theta, c, l, \alpha\}$ , 使得:

$$E(\Phi_1) = E_{landmark} + E_{photometric} + E_{regularizers} \quad (3.9)$$

其中

$$E_{landmark} = \sum_{i=1}^{68} \|k_i - s\Pi(M_i) + t\|_1 \quad (3.10)$$

在这里使用由 FAN<sup>[19]</sup>检测到的 68 个关键点, 因为 mediapipe 检测到的关键点在大角度姿势下通常不准确。 $E_{photometric}$  与公式 (3.5) 相同。 $E_{regularizers}$  包括  $\|\theta\|_2^2$ ,  $\|c\|_2^2$ ,  $\|l\|_2^2$  和  $\|\alpha\|_2^2$ 。

在第二阶段, 本研究冻结在第一阶段优化的参数, 仅通过最小化公式(3.9)来优化  $\Phi_2 = \{\beta, \psi\}$ 。 $E_{landmark}$  和  $E_{photometric}$  与第一阶段相同, 但  $E_{regularizers}$  不同, 包括  $\|\beta\|_2^2$  和  $\|\psi\|_2^2$ 。

### 3.3 实验与分析

#### 3.3.1 数据集与实现细节

关于辅助编码器, 使用自然场景人脸图像数据集和有 3D 标记的人脸图像数据集上训练。前者包括 VGGFace2<sup>[46]</sup>、CelebA<sup>[48]</sup> 和 FFHQ<sup>[38]</sup>, 涵盖了头部姿态、年龄和身份方面的多样性。本文使用 Bulat 等人<sup>[19]</sup>的方法检测人脸并将图像裁剪为 224×224。它们经过 Mediapipe<sup>[44]</sup>设置的最小检测置信度为 0.9 进行筛选, 得到了约 1000K 个自然场景下的人脸图像。带标注的数据集是从 300W-LP<sup>[23]</sup> 构建而成的, 包含约 120K 个带标注的合成人脸图像。使用 PyTorch<sup>[49]</sup>实现, 使用 PyTorch3D 的可微分光栅化器<sup>[42]</sup>进行渲染。使用 Adam 优化器<sup>[50]</sup>训练, 批量大小为 10, 学习率为  $1e^{-4}$ , 总迭代次数为 500K。

关于身份编码器, 使用 MICA 数据集<sup>[17]</sup>训练。MICA 数据集<sup>[17]</sup>中包含成对的 2D/3D 数据, 由 8 个较小的数据集组成, 包含 FLAME 拓扑下的约 2300 名受试者。使用 PyTorch<sup>[49]</sup>实现, 使用 AdamW 优化器<sup>[51]</sup>训练, 批量大小为 8, 学习率为  $1e^{-5}$ , 权重衰减为  $2e^{-4}$ , 总迭代次数为 160k。联合优化的第一阶段和第二阶段分别包含 400 和 100 次迭代, 学习率为 0.001, 权重衰减为  $1e^{-4}$ 。

#### 3.3.2 定性比较

图 3.6 与图 3.7 分别从正视图与侧视图两个方面将 PR3D 与其他的人脸重建方法进行视觉上的比较, 包括 Deep3D<sup>[14]</sup>, FFHQ-UV<sup>[10]</sup>, MICA<sup>[17]</sup> 和 DECA<sup>[15]</sup>。在不同身份、姿态、光照等条件下的人脸图片作为输入时, 本文的方法展示出较好的鲁棒性和准确性。由于不准确的弱监督信号, 基于自监督的方法像 Deep3D<sup>[14]</sup> 和 DECA<sup>[15]</sup> 在输入图片和重建结果之间有肉眼可见的不匹配性。基于优化的方法如 FFHQ-UV<sup>[10]</sup> 需要长时间的优化时间, 且仍然无法避免弱监督信号的影响。MICA<sup>[17]</sup> 使用有监督学习以提升重建精度, 但是缺乏了表情、姿态的训练数据导

致了来自输入图片信息的不充分利用。

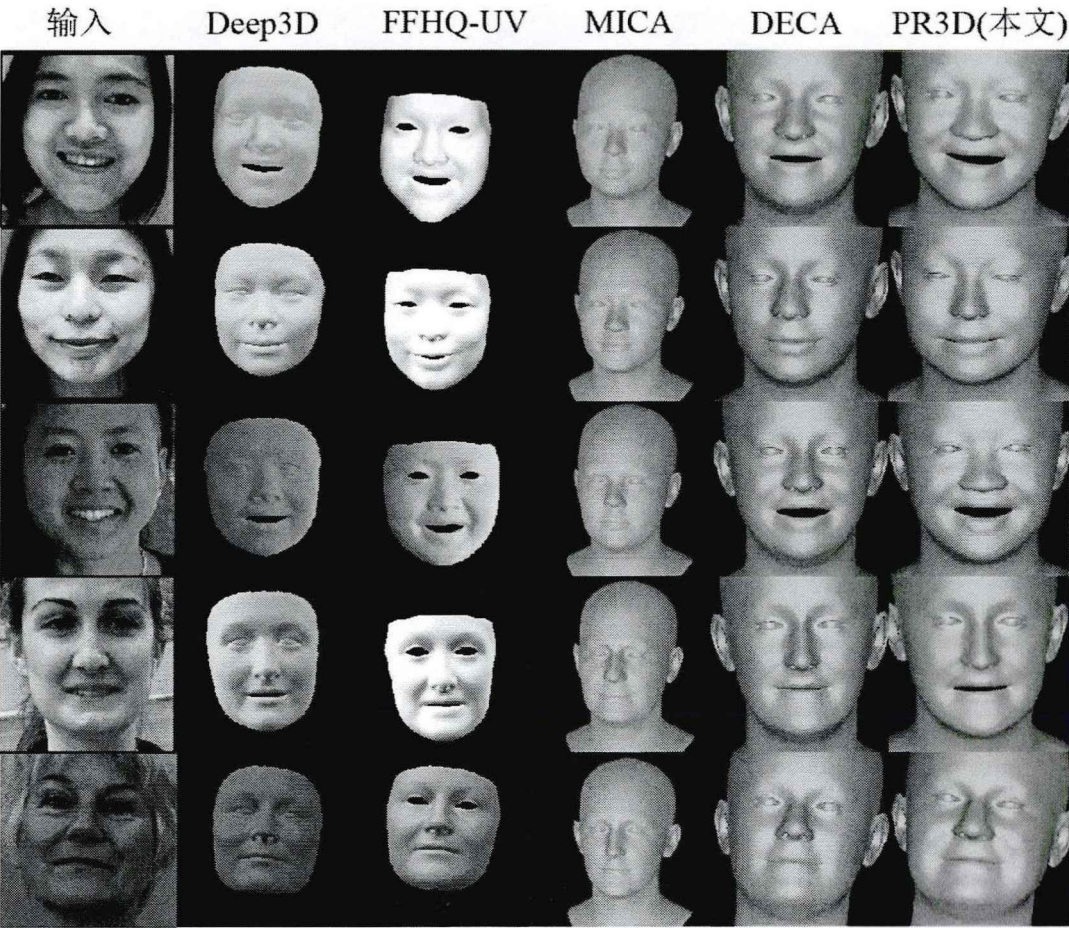


图 3.6 与其他形状重建方法的定性比较（正视图）

3.3.3 定量比较

1. NoW 基准

NoW 数据集<sup>[31]</sup> 包括来自 100 名测试对象的 2054 张图像，涵盖了各种表情、姿势和遮挡情况。该数据集被分为验证集和测试集。NoW 为每个测试对象提供了 3D 真实扫描，使用的评估指标是在刚性对齐后的扫描到重建网格的距离。表 3.2 展示了本文的方法与其他最先进方法的性能比较。本文的方法在验证集和测试集上都实现了最低的平均、中位数和标准差误差，超过了包括最近提出的 AlbedoGAN<sup>[52]</sup> 等在 NoW 基准上的所有已发布方法。这表明本文的基于半监督学习的形状重建方法具有最高的重建精度，并且在各种表情、头部姿态和遮挡情况下都具有鲁棒性。

2. REALY 基准

REALY 数据集<sup>[39]</sup> 专注于计算真实扫描和预测网格在四个特定的面部区域（鼻子、嘴巴、额头和脸颊）之间的相似度。该数据集分为两个子集：正面视图和侧



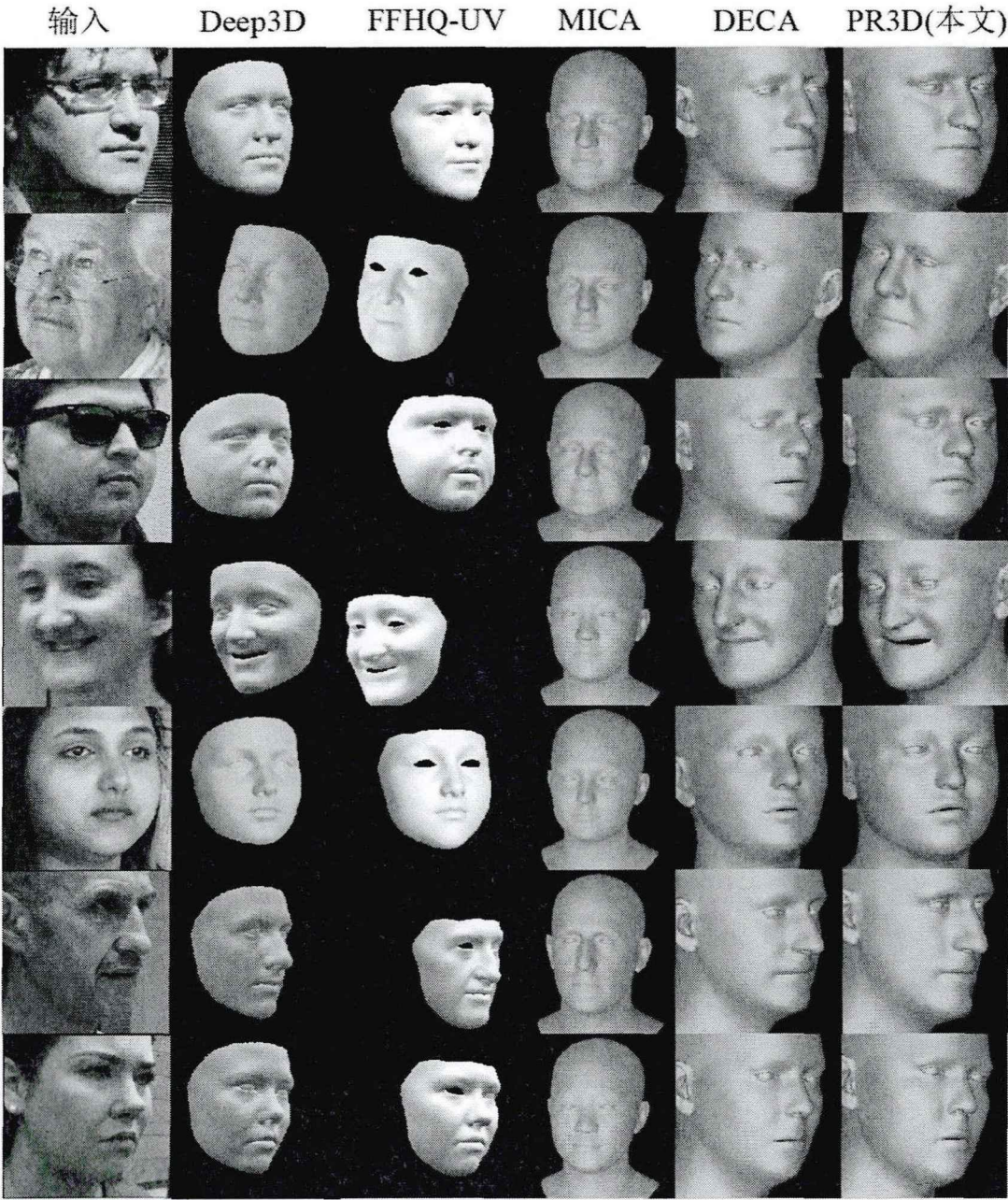


图 3.7 与其他形状重建方法的定性比较（侧视图）

面视图。对四个面部区域的平均归一化均方误差 (NMSE) 进行计算以进行排序。如表 3.3 所示，本文的形状重建方法在正面视图和侧面视图上分别比 DECA<sup>[15]</sup> 模型提升了 14.6% 和 13.8%，并分别比 MICA<sup>[17]</sup> 模型提升了 19.5% 和 15.3%。这表明虽然本研究采用了与 DECA 和 MICA 类似的模块，但本研究的设计有效地结合了两种方法的优点，并极大地提高了形状重建的准确性。考虑到正面视图和侧面视图的性能，PR3D 表现出与 AlbedoGAN<sup>[52]</sup> 相当的性能。REALY<sup>[39]</sup> 中的测试图像是从实验室环境的参与者收集而来的。PR3D 在 REALY<sup>[39]</sup> 的侧面视图图像上没有表现出卓越的性能可能的原因是缺乏来自实验室环境的侧面视图训



表 3.2 NoW 验证集和测试集上的重建误差

| 方法                              | 验证集         |             |             | 测试集         |             |             |
|---------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                                 | 中位数         | 平均数         | 标准差         | 中位数         | 平均数         | 标准差         |
| Deep3D <sup>[14]</sup>          | 1.09        | 1.37        | 1.16        | 1.11        | 1.41        | 1.21        |
| DECA <sup>[15]</sup>            | 1.18        | 1.46        | 1.25        | 1.09        | 1.38        | 1.18        |
| Dense landmarks <sup>[16]</sup> | 0.95        | 1.17        | 0.97        | 1.02        | 1.28        | 1.08        |
| MICA <sup>[17]</sup>            | 0.91        | 1.13        | 0.94        | 0.90        | 1.11        | 0.92        |
| AlbedoGAN <sup>[52]</sup>       | 0.90        | 1.12        | 0.96        | 0.98        | 1.21        | 0.99        |
| PR3D(本文)                        | <b>0.89</b> | <b>1.10</b> | <b>0.92</b> | <b>0.89</b> | <b>1.10</b> | <b>0.92</b> |

练图像。

表 3.3 REALY 基准上正面图像和侧面图像的重建误差

| 方法                        | @ 鼻子  | @ 嘴部  | @ 前额  | @ 脸颊  | @ 全部         |
|---------------------------|-------|-------|-------|-------|--------------|
| <b>正面视图</b>               |       |       |       |       |              |
| PR3D(本文)                  | 1.569 | 2.211 | 1.966 | 1.122 | <b>1.717</b> |
| AlbedoGAN <sup>[52]</sup> | 1.656 | 2.087 | 2.102 | 1.141 | 1.746        |
| DECA <sup>[15]</sup>      | 1.697 | 2.516 | 2.394 | 1.479 | 2.010        |
| PRNet <sup>[53]</sup>     | 1.923 | 1.838 | 2.429 | 1.863 | 2.013        |
| EMOCA <sup>[54]</sup>     | 1.868 | 2.679 | 2.426 | 1.438 | 2.103        |
| MICA <sup>[17]</sup>      | 1.585 | 3.478 | 2.374 | 1.099 | 2.134        |
| RingNet <sup>[31]</sup>   | 1.934 | 2.074 | 2.995 | 2.028 | 2.258        |
| <b>侧面视图</b>               |       |       |       |       |              |
| AlbedoGAN <sup>[52]</sup> | 1.576 | 2.218 | 2.142 | 1.112 | 1.762        |
| PR3D(本文)                  | 1.465 | 2.482 | 2.247 | 1.075 | <b>1.817</b> |
| PRNet <sup>[53]</sup>     | 1.868 | 1.856 | 2.445 | 1.960 | 2.032        |
| DECA <sup>[15]</sup>      | 1.903 | 2.472 | 2.423 | 1.630 | 2.107        |
| EMOCA <sup>[54]</sup>     | 1.867 | 2.636 | 2.448 | 1.548 | 2.125        |
| MICA <sup>[17]</sup>      | 1.525 | 3.567 | 2.379 | 1.109 | 2.145        |
| RingNet <sup>[31]</sup>   | 1.921 | 1.994 | 3.081 | 2.027 | 2.256        |

### 3.3.4 消融实验

为了验证本文提出的方法的有效性，进行了消融实验以探索不同的设计变化。基准模型采用了简单的编码器-解码器架构，使用由 FAN<sup>[19]</sup>检测的 68 个关键点损失以及光度损失、感知级损失和正则化损失进行训练。在消融实验中，对基准模型进行了几项修改。这些变体在 NoW 验证集上进行了评估。

- 122 个关键点: 首先，将 68 个关键点替换为 122 个关键点，其中使用了 Mediapipe<sup>[44]</sup>提供的 105 个关键点和 FAN<sup>[19]</sup>提供的 17 个关键点。如表 3.4 所示，检测更多高质量的关键点有助于实现更好的性能。
- 标记损失: 尽管增加了关键点的数量，但它们仍然不足以恢复高精度的 3D 人脸，因此引入了来自有 3D 标记数据集的密集关键点损失。值得注意的是，有标记数据集的大部分图像是大角度姿态图像，因此这弥补了大姿态情况下关键点检测工具的低精度问题。
- 身份编码器: 最后，添加了身份编码器来提取一对姿态或光照等其他因素

不变的身份特征，模型实现了表 3.4 中报告的最终最佳结果。

表 3.4 在 NoW 验证集上对不同模块的消融实验

| 辅助编码器  |         |      | 身份编码器 | 中位数  | 均值   | 标准差  |
|--------|---------|------|-------|------|------|------|
| 68 关键点 | 122 关键点 | 标记损失 |       |      |      |      |
| ✓      |         |      |       | 1.20 | 1.50 | 1.26 |
|        | ✓       |      |       | 1.14 | 1.42 | 1.19 |
|        | ✓       | ✓    |       | 1.07 | 1.34 | 1.13 |
|        | ✓       | ✓    | ✓     | 0.89 | 1.10 | 0.92 |

为探究联合优化过程中的参数对重建精度、时间消耗的影响，本文分别改变联合优化过程第一阶段和第二阶段的步数，在 NoW 验证集上进行实验，结果如表 3.5 所示。从表中可以看出，固定阶段二步数为 100，阶段一步数从 0 开始，以 100 为增量递增，重建误差呈现先降低后增加的变化趋势，在阶段一步数为 500 时取得最低值；固定阶段一步数为 400，阶段二步数从 0 开始，以 50 为增量递增，重建误差呈现先降低后增加的变化趋势，在阶段二步数为 100 时取得最低值。从时间消耗变化可以得出，步数越多时间消耗越久，而重建精度不一定提高，实验中最低时间消耗为 4.8 秒，最高为 39.7 秒。综上，可以根据实际需求，在时间消耗和精度提升方面做好权衡，选择适合的步数设置。

表 3.5 在 NoW 验证集上对联合优化过程参数的消融实验

| 实验序号 | 步数设置 |     | NoW 验证集 |        |        | 运行时间（秒） |
|------|------|-----|---------|--------|--------|---------|
|      | 阶段一  | 阶段二 | 中位数     | 平均数    | 标准差    |         |
| 1    | 0    | 100 | 0.9113  | 1.1278 | 0.9415 | 4.8     |
| 2    | 100  | 100 | 0.9101  | 1.1251 | 0.9352 | 9.7     |
| 3    | 200  | 100 | 0.9033  | 1.1166 | 0.9290 | 14.6    |
| 4    | 300  | 100 | 0.8963  | 1.1084 | 0.9236 | 18.8    |
| 5    | 400  | 100 | 0.8903  | 1.1025 | 0.9207 | 23.6    |
| 6    | 500  | 100 | 0.8876  | 1.1005 | 0.9204 | 28.2    |
| 7    | 600  | 100 | 0.8883  | 1.1020 | 0.9228 | 35.9    |
| 8    | 700  | 100 | 0.8916  | 1.1065 | 0.9257 | 39.7    |
| 9    | 400  | 0   | 0.9099  | 1.1252 | 0.9391 | 19.5    |
| 10   | 400  | 50  | 0.8954  | 1.108  | 0.9261 | 21.7    |
| 11   | 400  | 100 | 0.8903  | 1.1025 | 0.9207 | 23.6    |
| 12   | 400  | 150 | 0.8945  | 1.1079 | 0.9228 | 26.6    |
| 13   | 400  | 200 | 0.9020  | 1.1186 | 0.9295 | 29.7    |

3.4 实时重建

不同于郭建珠等人<sup>[22]</sup>和 Wood 等人<sup>[16]</sup>等人的实时重建工作，本文基于高精度的形状重建模型，采用了师生学习蒸馏出轻量级实时重建模型 fast-PR3D。将 PR3D 的广泛知识和复杂性提炼为更简化和高效的架构，在保持高准确性的同时

显著提高计算速度，使 fast-PR3D 成为一种用于高效人脸重建的实时模型。

### 3.4.1 轻量级卷积神经网络

Fast-PR3D 使用 MobileNetV2<sup>[55]</sup> 作为骨架。MobileNet<sup>[56]</sup> 旨在在移动设备和嵌入式设备上实现高效的图像识别和分类任务，通过精心设计的架构和参数数量的减少，实现了在资源受限的设备上高效地运行。与传统的深度神经网络相比，MobileNet 具有更小的模型尺寸和更少的计算量，同时保持了良好的准确性和性能。

MobileNet 的设计基于两个关键思想：深度可分离卷积和轻量级的网络结构。深度可分离卷积是 MobileNet 中的核心组件，通过将标准的卷积操作分解为深度卷积和逐点卷积两个步骤，从而降低了参数量和计算量。此外，MobileNet 采用了轻量级的网络结构，包括深度可分离卷积层、批量归一化层和 ReLU 激活函数，以实现高效的特征提取和模型压缩。

MobileNetV2 是 MobileNet 的进化版本，进一步提升了性能和效率。MobileNetV2 在 MobileNet 的基础上引入了一系列创新的设计，包括倒残差结构、线性瓶颈和扩张卷积等。这些改进使得 MobileNetV2 具有更高的准确性、更快的收敛速度和更低的计算开销。MobileNetV2 的倒残差结构采用了一种新型的连接方式，将低维度的特征图直接与高维度的特征图进行连接，从而提高了信息的传递效率和特征的重用性。此外，线性瓶颈结构和扩张卷积等技术进一步提升了模型的表达能力和泛化能力。

MobileNetV2 在多个图像分类和目标检测任务上取得了优异的性能，在保持高效性的同时实现了与传统深度神经网络相媲美的准确性。由于其出色的性能和高效的设计，MobileNetV2 已经成为移动端和嵌入式系统中广泛应用的深度学习模型之一。

### 3.4.2 算法流程

作为教师模型，PR3D 进行高维 FLAME 参数回归，其中包括  $\beta \in \mathbb{R}^{300}$ 、 $\psi \in \mathbb{R}^{50}$ 、姿势编码  $\theta \in \mathbb{R}^6$  和相机编码  $c \in \mathbb{R}^3$ 。作为学生模型，fast-PR3D 使用轻量级的 MobileNetV2<sup>[55]</sup> 骨干网络，仅回归了  $\beta$  的前 40 个维度和  $\psi$  的前 10 个维度，而保持了姿势和相机编码的维度不变，因为  $\beta$  和  $\psi$  的最后部分对面部形状和表情的影响较小。图 3.8 中展示了说明这个过程的示意图。本文使用简单的均方差（MSE）损失函数来评估教师模型 PR3D 和学生模型 fast-PR3D 之间的差异。该损失函数的计算方式如下：

$$L_{distill} = \sum_{i=1}^{40} \|\beta_i^* - \beta_i\|_2^2 + \sum_{i=1}^{10} \|\psi_i^* - \psi_i\|_2^2 + \sum_{i=1}^6 \|\theta_i^* - \theta_i\|_2^2 + \sum_{i=1}^3 \|c_i^* - c_i\|_2^2 \quad (3.11)$$

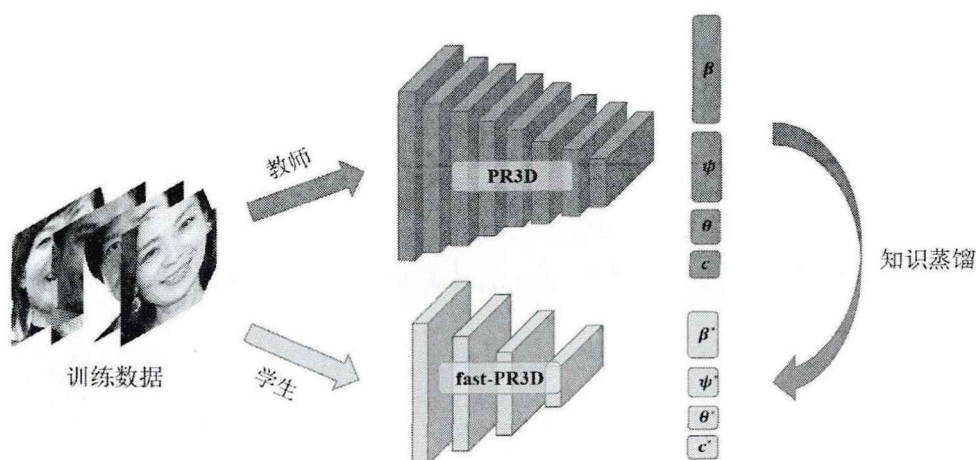


图 3.8 师生学习蒸馏得到轻量级人脸重建模型

其中,  $\beta_i^*$ 、 $\psi_i^*$ 、 $\theta_i^*$  和  $c_i^*$  表示从 fast-PR3D 回归得到的参数,  $\beta_i$ 、 $\psi_i$ 、 $\theta_i$  和  $c_i$  来自 PR3D。训练后, 本研究使用 onnxruntime 库<sup>[49]</sup> 进一步加速 MobileNetV2 网络。

### 3.4.3 实验与分析

用于实时重建的训练集包含来自 VGGFace2<sup>[46]</sup>、CelebA<sup>[48]</sup> 和 FFHQ<sup>[38]</sup> 的约 200K 张自然环境人脸图像, 并使用 Bulat 等人<sup>[19]</sup> 的方法将它们裁剪为  $120 \times 120$ 。本研究使用 Adam 优化器<sup>[50]</sup> 对 fast-PR3D 进行训练, 批量大小为 10, 初始学习率为  $1e^{-4}$ , 总迭代次数为 100K。在单个 CPU 线程 (AMD Ryzen 7 5800H) 上运行, fast-PR3D 以单张图像作为输入花费约 3 毫秒来回归 FLAME 参数。

由于与实时重建相关的工作较少, 本文仅将 fast-PR3D 与两种最近的实时重建工作 3DDFA-V2<sup>[22]</sup> 和 Dense landmarks<sup>[16]</sup> 在 NoW<sup>[31]</sup> 基准上进行比较。为了有效比较三种方法的速度, 本研究设计了以下实验: 对于 fast-PR3D 和 3DDFA-V2<sup>[22]</sup>, 在相同的 CPU (AMD Ryzen 7 5800H) 上对它们进行测试。对于 Dense landmarks<sup>[16]</sup>, 由于缺乏开源可用性以及实验平台之间的可比性能, 采用了他们论文中在 CPU 平台 (i5-11600K) 上的测试结果。

如表 3.6 所示, 在相近的实时速度下, fast-PR3D 与 3DDFA-V2<sup>[22]</sup> 相比, 准确性明显提高, 这是因为 PR3D 的高准确性在知识蒸馏过程中被大部分保留了下来。3DDFA-V2<sup>[22]</sup> 在快速推理速度方面表现出色, 但在准确性方面表现不佳。Wood 等人<sup>[16]</sup> 预测了 320 个关键点, 然后将 3D 面部模型拟合到这些关键点上, 实现了较高的重建精度, 但代价是相对较长的时间。与 Dense landmarks<sup>[16]</sup> 相比, fast-PR3D 具有可比的重建精度, 但由于其更简单、更直接的推理过程, 速度快了 2.4 倍。

表 3.6 实时重建方法在重建精度和速度上的比较

| 方法                              | NoW 验证集 |      |      | 推理时间（毫秒） |
|---------------------------------|---------|------|------|----------|
|                                 | 中位数     | 平均数  | 标准差  |          |
| 3DDFA-V2 <sup>[22]</sup>        | 1.21    | 1.54 | 1.31 | 2.9      |
| Dense landmarks <sup>[16]</sup> | 1.00    | 1.24 | 1.02 | 6.5      |
| Fast-PR3D(本文)                   | 1.03    | 1.29 | 1.08 | 2.7      |

当使用笔记本摄像头捕捉实时面部图像时，重建结果和比较如图 3.9所示，表明本文的实时重建方法更加贴近捕捉到的面部形状，特别是在眼睛和鼻子区域。



图 3.9 用摄像头实时采集人脸并重建

表 3.7将 PR3D 与 fast-PR3D 的资源消耗（时间与空间）与重建精度，即投入与产出进行对比。由于 PR3D 的辅助编码器采用了 ResNet-50 架构，身份编码器采用了 ArcFace 架构，参数量之和为 118M。由于联合优化过程的步数设置，时间消耗在几秒到几十秒间，当阶段一步数为 400，阶段二步数为 100 时，时间消耗约为 20 秒，在 NoW 验证集的重建误差的中位数，平均数和标准差分别为 0.89，1.10 和 0.92。由于 fast-PR3D 采用了 MobileNetV2 这一轻量级的架构，参数量为 3.2M，时间消耗约为 3 毫秒，在 NoW 验证集的重建误差的中位数，平均数和标准差分别为 1.03，1.29 和 1.08。综上，fast-PR3D 并不追求重建的高精度，采用了轻量级的网络架构，大幅降低了资源消耗，在保证一定重建精度的同时实现了实时重建。

表 3.7 PR3D 与 fast-PR3D 在参数量、运行时间、重建精度的对比

| 模型        | 参数量  | 时间消耗 | 重建误差 |      |      |
|-----------|------|------|------|------|------|
|           |      |      | 中位数  | 平均数  | 标准差  |
| PR3D      | 118M | ~20s | 0.89 | 1.10 | 0.92 |
| fast-PR3D | 3.2M | ~3ms | 1.03 | 1.29 | 1.08 |

3.5 本章小结

本章介绍了本文提出的基于半监督学习的高精度形状重建算法。首先介绍了与算法相关的模型，包括人脸模型、相机模型等。然后介绍了半监督形状重建算法的提出动机与具体设计。接下来通过实验与分析来说明本文提出的形状重



建方法比现有的其他前沿方法达到了更高的重建精度。最后介绍了实时重建的方法和实验分析,说明本文的实时形状重建效果在精度和速度上达到了更好的平衡。

## 第4章 基于 StyleGAN2 的高保真纹理重建算法

本文针对由于缺乏足够 3D 扫描数据造成的保真度低的问题,以及由于输入图片自遮挡等因素造成的纹理缺失等问题,设计了基于 StyleGAN2 的高保真纹理重建算法。本章首先在 4.1 节介绍了 StyleGAN2 的潜在空间和 StyleGAN2 生成图片的过程。其次,在 4.2 节分析了在真实图片作为输入,进行反演时在重构性和可编辑性的权衡。接下来,通过分析通常的纹理重建流程存在的问题,提出了本文的基于 StyleGAN2 的高保真纹理重建方法,分别对反演编码器和掩膜网络的训练方法和损失函数进行详细介绍。最后,在 4.4 节介绍数据集与实现细节,实验效果及对比实验,证明了本文提出的纹理重建方法的高保真度。

### 4.1 StyleGAN2 的潜在空间

StyleGAN2<sup>[30]</sup>是一种用于生成对抗网络(GAN)的先进架构,旨在生成高质量的逼真图像。它是对先前版本 StyleGAN 的改进和扩展,在生成逼真图像方面取得了巨大的进步,并在多个领域,如计算机视觉和艺术创作中引起了广泛关注。

如图 4.1 所示,StyleGAN2 提供了一个基本潜在空间  $\mathcal{Z} \subset \mathbb{R}^{512}$ ,其中样本从多元标准高斯分布  $\mathcal{N}(0, I)$  中提取,然后通过多层感知器(MLP)生成中间潜在表示  $w \in \mathcal{W} \subset \mathbb{R}^{512}$ 。然后,将  $w$  输入合成网络的  $N_l$  层。 $\mathcal{W}$  空间已被证明比  $\mathcal{Z}$  空间更解纠缠<sup>[30]</sup>,使其成为合成图像处理的最常见选择<sup>[57]</sup>。Wu 等人<sup>[58]</sup>为 StyleGAN2 引入了一个能够更好地对合成图像进行语义编辑的潜空间,称为样式空间  $\mathcal{S}$ 。即在 StyleGAN2 中,每个潜码  $w \in \mathcal{W}$  都通过仿射变换  $A$  转换为基于通道的样式向量  $s \in \mathcal{S}$ ,然后传递到生成器的不同层中(如图 4.1 所示)。

### 4.2 反演权衡

对于真实人脸图像作为输入,需要将其映射到 StyleGAN2 的隐空间中。已经被证明不是每个真实的图像都可以反演进入 StyleGAN 的潜在空间<sup>[59]</sup>。为了缓解这个限制,可以通过输入  $k$  个不同的风格编码而不是单个向量来增加 StyleGAN 生成器的表现力。将这个扩展空间表示为  $\mathcal{W}^k \subseteq \mathbb{R}^{k \times 512}$ ,其中  $k$  是生成器的风格输入数量。例如,一个能够在  $1024 \times 1024$  分辨率下合成图像的生成器在扩展的  $\mathcal{W}^{18}$  空间中操作,对应着 18 个不同的风格输入。通过输入不一定来自  $\mathcal{W}$  真实分布的风格编码,即 StyleGAN 映射函数范围之外的编码,可以实现更多的表现力。这种扩展可以通过复制单个风格编码或者使用  $k$  个不同的风格编码来应用。

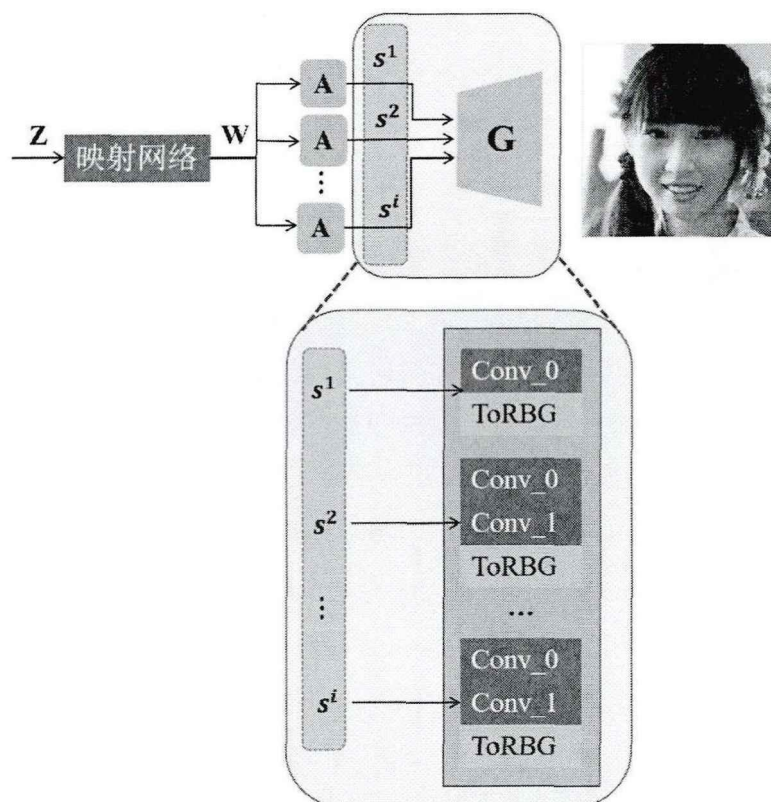


图 4.1 StyleGAN2 生成器的层次结构

本研究将这些扩展分别表示为  $\mathcal{W}_*$  和  $\mathcal{W}_*^k$ 。

在 GAN 反演任务时，仅仅进行重构是有限的，因为最多只能将反演的图像重构为原始图像。正如以前 Abdal 等人<sup>[59]</sup>的工作所提到的，反演的中心动机是为了进一步进行潜在编辑操作。也就是说，成功重构应该使得对真实图像进行广泛和多样化的潜在编辑更加容易。具体来说，对评估 GAN 反演方法应该基于重构性和可编辑性两方面进行评估。重构包括两个独立属性—失真度和感知质量。形式上，失真度被定义为  $E_{x \sim p_X}[\Delta(x, G(w))]$ ，其中  $p_X$  是真实图像的分布， $\Delta(x, G(w))$  是图像  $x$  和  $G(w)$  之间的图像空间差异度量。感知质量衡量重构图像的真实性，与任何参考图像都没有关系。对于可编辑性，预期是给定反演的潜在编码，可以找到许多潜在空间方向，对应于图像空间中解耦的语义编辑。

$\mathcal{W}_*^k$  与  $\mathcal{W}$  有两个不同之处。首先， $\mathcal{W}_*^k$  可能在不同的风格调制层包含不同的风格编码。其次，每个风格编码不受  $\mathcal{W}$  真实分布的限制，而是可以从  $\mathbb{R}^{512}$  中取任何值，所以  $\mathcal{W}_*^k$  达到了比  $\mathcal{W}$  更低（即更好）的失真度。另一方面，由于 StyleGAN 最初是在  $\mathcal{W}$  空间中训练的，因此  $\mathcal{W}$  比其  $\mathcal{W}_*^k$  对应部分具有更好的感知质量和可编辑性。综上，在  $\mathcal{W}_*^k$  空间中，接近  $\mathcal{W}$  时，失真度会变差，而可编辑性和感知质量会提高。

### 4.3 算法流程

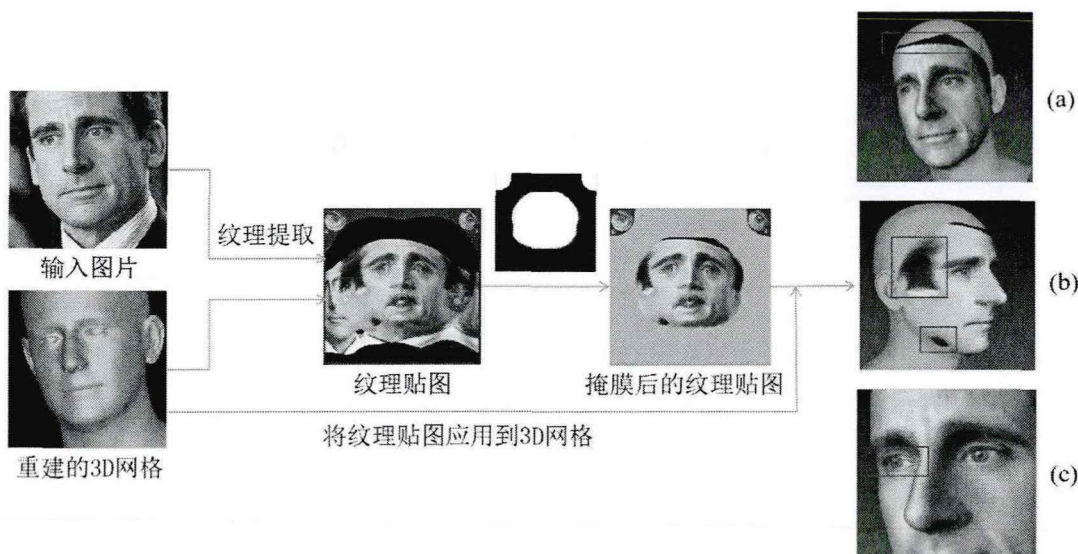


图 4.2 通常的纹理重建流程

如图4.2所示，在完成上章的形状重建流程后，重建的 3D 网格与输入图像对齐。3D 网格由一系列三角形面构成，而图像则是由离散的像素点组成。为了建立两者之间的精确对应关系，采用了双线性插值方法，该方法能够在连续的几何空间与离散的像素空间之间建立平滑的映射关系。

值得注意的是，FLAME 模型为其拓扑结构中的每个三角形面片提供了 UV 坐标。UV 坐标是计算机图形学中用于表示二维纹理映射到三维物体表面的常见方式。通过利用这些 UV 坐标，能够根据对齐后的网格和图像，从图像中准确地提取出对应的纹理信息。

然而，尽管这种方法在一定程度上能够实现纹理的提取与映射，但仍存在一些固有的问题。如图4.2(a)所示，在额头区域存在明显的像素丢失现象。这是由于在提取纹理时，如果原始图像过度裁剪，那么与面部对应的一些像素将无法被正确提取，从而导致像素信息的丢失。

此外，如图4.2(b)所示，一些面部区域包含了错误的像素颜色。这主要是由于姿态变化引起的遮挡效应。在某些姿态下，部分面部区域可能会被遮挡，导致这些区域的像素颜色与实际的面部颜色不匹配。

最后，如图4.2(c)所示，眼部区域的纹理存在明显的不对齐现象。这主要是由于形状重建的精度不足所致。如果形状重建的结果与真实面部形状存在较大偏差，那么在提取纹理时，纹理与网格之间的对应关系就会出现错误，从而导致纹理的对齐问题。

综上所述，尽管基于投影和双线性插值的纹理提取方法在一定程度上能够

实现纹理的重建,但仍存在一些固有的问题,如像素丢失、颜色错误和纹理不对齐等。

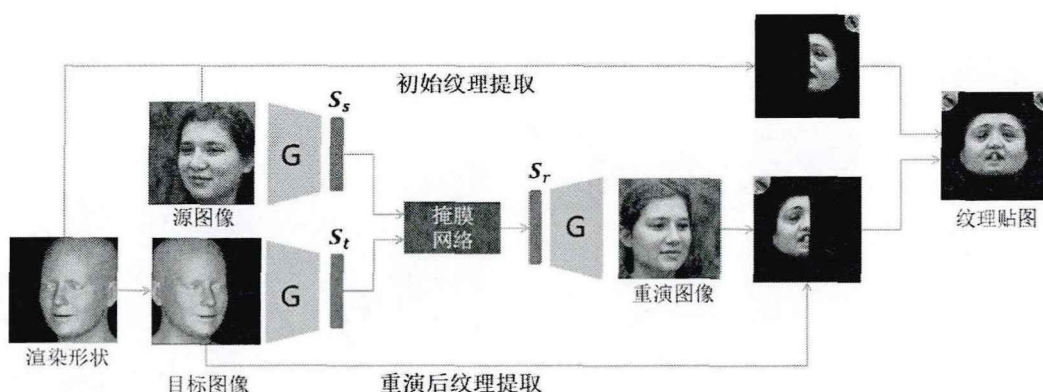


图 4.3 本文纹理重建框架

为了克服传统纹理重建方法在细节表现和逼真度方面的局限性,本文深入探索了 StyleGAN2 这一强大生成式对抗网络(GAN)的卓越生成能力。StyleGAN2 不仅具备生成高质量、高保真度图像的能力,更关键的是,其生成机制为精细而逼真的纹理重建提供了可能。作为一个先进的 GAN 模型,StyleGAN2 通过引入一个用于合成图像语义编辑的潜在空间,极大地拓宽了其在人脸图像生成和编辑领域的应用边界。

本文充分利用 StyleGAN2 的潜在空间特性,将人脸重演技术巧妙地引入纹理重建的流程之中。人脸重演技术能够有效地将目标图像的姿势和表情信息精确地转移到源图像上,同时确保源图像的身份特征得以完整保留。与人脸重演工作 StyleMask<sup>[60]</sup>相似,本研究设计了掩膜网络,对源风格编码  $S_s$  和目标风格代码  $S_t$  之间的差异进行处理。掩膜网络通过学习并输出重演的风格编码  $S_r$ ,实现了源图像和目标图像之间风格信息的有效融合。通过这种方式,能够生成既保持源图像身份特征,又融合了目标图像表情和姿势信息的重演图像。

与先前的工作(如<sup>[8,10,15]</sup>)不同,本文在纹理重建领域取得了显著的创新和突破。通过将人脸重演方法应用于纹理重建,本研究成功地补充了源图像中由于自遮挡引起的缺失纹理信息。具体而言,通过调整重建的 3D 形状的姿势或表情参数,能够得到渲染的目标图像。随后,利用 Tov 等人提出的反演方法<sup>[61]</sup>,可以得到图像对应的风格编码,在经过掩膜网络的处理后,得到了具有与源图像具有互补姿态的重演面部图像。掩膜网络很好地解耦了 StyleGAN2 的风格空间,因此只有源图像的姿势根据目标姿势进行更改,使得重演图像与具有目标姿势的 3D 网格对齐。源图像和重演图像分别与相应的 3D 网格对齐,因此可以使用专门设计用于提取目标区域纹理的 UV 掩膜从不同姿势的面部提取纹理,获得



重建的高质量和高保真度纹理贴图。

#### 4.3.1 掩膜网络

给定  $\mathcal{Z}$  空间中的一对随机潜码，一个针对源图像，一个针对目标图像，即  $\mathbf{z}_s, \mathbf{z}_t$ ，然后输入 StyleGAN2 的映射网络在  $\mathcal{W}$  空间中得到相应的编码。将得到的编码  $\mathbf{w}_s$  和  $\mathbf{w}_t$ ，用仿射模块 A 计算样式编码  $\mathbf{s}_s$  和  $\mathbf{s}_t$ ，最终分别生成源 ( $I_s$ ) 和目标 ( $I_t$ ) 图像。给定源和目标风格编码  $\mathbf{s}_s$  和  $\mathbf{s}_t$ ，输入掩膜网络计算一个掩码向量  $\mathbf{m}$ ，使重演的风格码  $\mathbf{s}_r \in \mathcal{S}$  包含  $\mathbf{s}_t$  对应面部姿态 (即头部姿态和表情) 的通道和  $\mathbf{s}_s$  对应面部身份特征 (即面部形状、头发颜色等) 的通道。也就是

$$\mathbf{s}_r = \mathbf{m} \odot \mathbf{s}_t + (1 - \mathbf{m}) \odot \mathbf{s}_s \quad (4.1)$$

其中  $\odot$  表示基于元素的乘法。本文通过将两个输入样式代码的差即  $\Delta \mathbf{s} = \mathbf{s}_s - \mathbf{s}_t$  输入掩膜网络来得到  $\mathbf{m}$ 。

一个风格编码  $\mathbf{s} \in \mathcal{S}$  由  $N_l$  个风格向量  $\mathbf{s}^i, i = 1, \dots, N_l$  组成，每一个对应于合成网络  $\mathcal{G}$  的不同层。遵循 StyleGAN 的基本层次结构，其中第一层控制粗细节 (如头部姿态)，中间层控制语义属性 (如表情或发型)，最后一层控制输出图像的细粒度细节，本研究以类似的方式设计掩膜网络。如图 4.4 所示，本研究对每一输入层优化一个单独的掩码  $\mathbf{m}_i, i = 1, \dots, N_l$ 。掩膜网络的每个子网络由两个全连接层和一个 ReLU 激活层构成，最后一层的输出经过 sigmoid 激活函数。为了获得最终的掩码  $\mathbf{m}$ ，将每个子网络的输出  $\mathbf{m}_i$  连接起来。

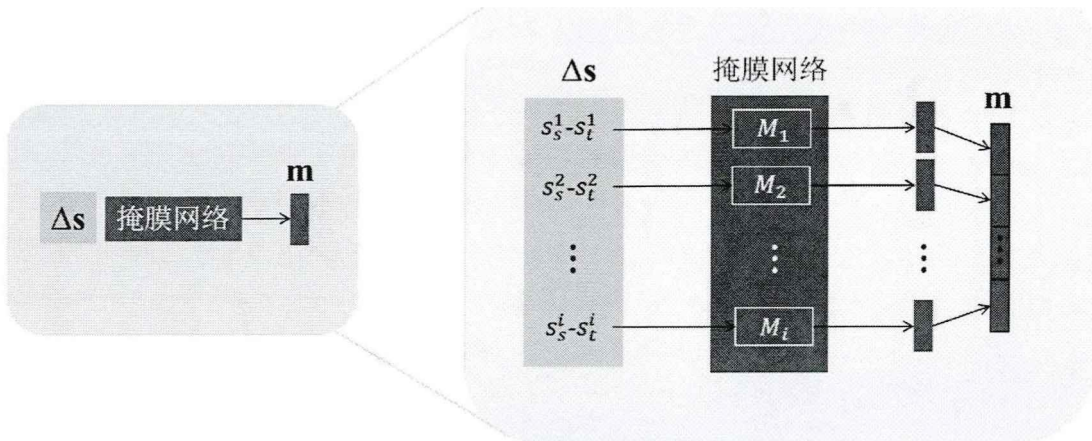


图 4.4 掩膜网络的架构

为了训练掩膜网络来解耦面部姿态和身份特征，本文利用了 3D 形变模型 (3DMM) 的解纠缠特性。其中，三维形状模型  $\mathbf{X} \in \mathbb{R}^{3N}$ ，其中  $N$  为顶点个数，定义为：

$$\mathbf{X} = \bar{\mathbf{X}} + \mathbf{U}_s \mathbf{a}_s + \mathbf{U}_e \mathbf{a}_e \quad (4.2)$$

其中,  $\bar{X} \in \mathbb{R}^{3N}$  表示平均三维形状模型,  $U_s \in \mathbb{R}^{3N \times m_s}$  和  $U_e \in \mathbb{R}^{3N \times m_e}$  表示形状和表情的标准正交基,  $a_s \in \mathbb{R}^{m_s}$  和  $a_e \in \mathbb{R}^{m_e}$  分别表示相应的形状和表情系数。本研究使用<sup>[15]</sup>中提出的方法计算  $X_s$  和  $X_t$ 。此外, 通过式 (4.2), 可以利用源人脸形状系数  $a_s$  和目标人脸姿态系数  $a_e$  及姿态得到真实的重演三维人脸  $X_t^{gt}$ 。

训练掩膜网络的损失函数之一为重演损失:

$$\mathcal{L}_r = \lambda_x \mathcal{L}_x + \lambda_{id} \mathcal{L}_{id} \quad (4.3)$$

其中  $\lambda_x, \lambda_{id}$  是损失系数,  $\mathcal{L}_x = \|X_r - X_t^{gt}\|_1$  代表形状损失,  $\mathcal{L}_{id}$  是身份损失。具体的, 形状损失 ( $\mathcal{L}_x$ ) 定义为利用重演图像计算出的重演三维形状  $X_r$  与重演真实三维形状  $X_t^{gt}$  之间的  $L_1$  距离。此外, 本文使用 ArcFace<sup>[40]</sup> 计算源图像  $I_s$  和重演图像  $I_r$  之间的身份保持损失。

为了进一步增强源和重演图像之间的一致性, 本文使用循环一致性损失<sup>[62]</sup>。具体来说, 采样得到一个新的随机目标图像  $I_{r'}$ , 并使用它来重演  $I_s$  和  $I_r$ , 如下所示:

$$\begin{aligned} s_r^1 &= m^1 \odot s_{r'} + (1 - m^1) \odot s_s \\ s_r^2 &= m^2 \odot s_{r'} + (1 - m^2) \odot s_r \end{aligned} \quad (4.4)$$

其中, 由风格编码  $s_r^1$  和  $s_r^2$  生成的两幅重演图像被鼓励传递源身份信息 and 第二个目标图像  $I_{r'}$  的面部表情和姿态信息。因此, 循环一致性损失定义为

$$\mathcal{L}_{cycle} = \lambda_x \mathcal{L}_{cx} + \lambda_{id} \mathcal{L}_{cid} \quad (4.5)$$

式中循环形状损失  $\mathcal{L}_{cx}$  表示为:

$$\mathcal{L}_{cx} = \|X_{r^1} - X_{r'}^{gt}\|_1 + \|X_{r^2} - X_{r'}^{gt}\|_1 \quad (4.6)$$

其中,  $X_{r^1}$  和  $X_{r^2}$  表示使用两幅重演图像  $I_r^1$  和  $I_r^2$  计算出的三维形状,  $X_{r'}^{gt}$  表示使用源人脸形状系数  $a_s$  和从  $I_{r'}$  得到的目标人脸姿态系数  $a_e$  及姿态来计算出的真实三维形状。定义循环身份损失定义为:

$$\mathcal{L}_{cid} = \text{sim}(\mathcal{F}(I_s), \mathcal{F}(I_r^1)) + \text{sim}(\mathcal{F}(I_s), \mathcal{F}(I_r^2)) + \text{sim}(\mathcal{F}(I_r), \mathcal{F}(I_r^2)) \quad (4.7)$$

其中  $\text{sim}(\cdot)$  表示了源 ( $I_s$ ) 与两个新的重演图像 ( $I_r^1$  和  $I_r^2$ ) 在 ArcFace<sup>[40]</sup> 特征上计算的余弦相似度。

最后, 掩膜网络在训练过程中最小化的总损失为

$$\mathcal{L} = \mathcal{L}_r + \mathcal{L}_{cycle} \quad (4.8)$$

式中  $\mathcal{L}_r$  为式 (4.3) 给出的重演损失,  $\mathcal{L}_{cycle}$  为式 (4.5) 给出的循环一致性损失。

### 4.3.2 反演编码器

基于 4.2 节的介绍, 需要控制反演编码器映射到  $\mathcal{W}_*^k$  中接近  $\mathcal{W}$  的区域。第一种方法是鼓励推断出的  $\mathcal{W}_*^k$  潜空间编码更接近  $\mathcal{W}_*$ , 即最小化不同风格编码之间的方差, 或者等效地, 鼓励风格编码相同。为此, 采用一种“渐进”训练方案。

设  $E(x) = (w_0, w_1, \dots, w_{N-1})$  表示反演编码器的输出, 其中  $N$  是风格调制层的数量。常见的编码器直接训练到  $\mathcal{W}_*^k$  中, 即分别同时学习每个  $w_i$ 。不同地, 本文推断出一个潜在编码  $w$ , 以及一组相对  $w$  的偏移量。更正式地说, 学习的输出形式为  $E(x) = (w, w + \Delta_1, \dots, w + \Delta_{N-1})$ 。

具体地, 在训练开始时, 设置  $\forall i: \Delta_i = 0$ , 并且编码器被训练为推断出一个  $\mathcal{W}_*$  编码。然后, 逐步允许网络按顺序为每个  $i$  学习一个不同的  $\Delta_i$ 。这样做有助于编码器逐渐从  $\mathcal{W}_*$  扩展到  $\mathcal{W}_*^k$ 。这种方案与 StyleGAN 特定输入层的语义含义结合在一起, 使得可以首先学习输入图像的粗略重构, 然后逐渐添加细节, 同时仍保持接近  $\mathcal{W}_*$ 。低频细节极大地控制着失真度。因此, 渐进式训练方案首先专注于通过调整粗略级别的偏移量来改善低频失真。然后, 编码器逐渐补充由更精细级别的偏移量引入的高频细节来逐渐补充这些偏移量。为了明确地强制接近  $\mathcal{W}_*$ , 添加了一个关于  $\Delta$  的正则化损失:

$$\mathcal{L}_{d\text{-reg}}(w) = \sum_{i=1}^{N-1} \|\Delta_i\|_2. \quad (4.9)$$

第二种接近  $\mathcal{W}$  的方法是鼓励反演编码器获得的  $\mathcal{W}_*^k$  潜在编码更接近  $\mathcal{W}^k$ 。也就是说, 鼓励各个风格编码位于  $\mathcal{W}$  的实际分布内。为此, 本文采用了一个潜空间鉴别器<sup>[63]</sup>, 它以对抗的方式训练, 以区分来自  $\mathcal{W}$  空间的真实样本 (由 StyleGAN 的映射函数生成) 和反演编码器学习的潜空间编码。需要注意的是, 这样的潜空间鉴别器解决了学习推断属于一个不能明确建模的分布的潜空间编码的挑战。通过这样做, 鉴别器鼓励编码器推断出位于  $\mathcal{W}$  而不是  $\mathcal{W}_*$  的潜空间编码。本研究使用分别作用于每个潜空间编码条目鉴别器, 表示为  $D_{\mathcal{W}}$ 。在每次迭代中, 计算每个  $E(x)_i$  的 GAN 损失, 并对所有  $i$  进行平均。本研究使用非饱和的 GAN 损失<sup>[25]</sup>, 并加上  $R_1$  正则化<sup>[64]</sup>,

$$\mathcal{L}_{\text{adv}}^D = -\mathbb{E}_{w \sim \mathcal{W}} [\log D_{\mathcal{W}}(w)] - \mathbb{E}_{x \sim p_X} [\log(1 - D_{\mathcal{W}}(E(x)_i))] + \frac{\gamma}{2} \mathbb{E}_{w \sim \mathcal{W}} [\|\nabla_w D_{\mathcal{W}}(w)\|_2^2], \quad (4.10)$$

$$\mathcal{L}_{\text{adv}}^E = -\mathbb{E}_{x \sim p_X} [\log D_{\mathcal{W}}(E(x)_i)] \quad (4.11)$$

下文将介绍实现反演编码器和训练方案。反演编码器基于 Pixel2Style2Pixel (pSp) 编码器<sup>[65]</sup>建立, 如图 4.5 所示。与原始的 pSp 编码器并行生成  $N$  个风格

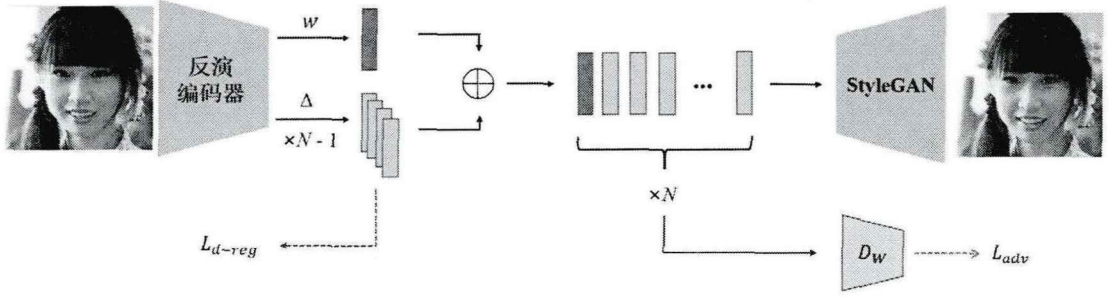


图 4.5 反演编码器网络架构

编码不同，本文遵循上文中提到的原则，生成一个单一的基本潜空间编码，用  $w$  表示，并生成一系列  $N - 1$  个偏移向量（在图 4.5 中用绿色表示）。然后，这些偏移量与基本潜空间编码  $w$  相加，得到最终的  $N$  个潜空间编码，然后将其输入到一个固定的、预先训练过的 StyleGAN2 生成器中，以获取反演后生成的图像。

为了训练反演编码器，本文采用确保低失真度的损失函数，这是在训练编码器时常见的做法，并且采用明确鼓励生成的潜空间编码保持接近  $w$  的损失函数，从而提高生成图像的感知质量和可编辑性。

在 pSp 中提出的关键思想之一是身份损失，它专门设计用于在面部领域准确反演真实图像。

$$\mathcal{L}_{\text{sim}}(x) = 1 - \langle C(x), C(G(e4e(x))) \rangle \quad (4.12)$$

其中  $C$  是一个使用 MOCOv2<sup>[66]</sup> 训练的 ResNet-50<sup>[43]</sup> 网络， $G$  是预训练的 StyleGAN2 生成器。在这里， $\mathcal{L}_{\text{sim}}$  明确鼓励编码器最小化重建图像的特征向量和其源图像之间的余弦相似度。由于提取特征的通用性质， $\mathcal{L}_{\text{sim}}$  可以应用于任意领域。需要注意的是，在面部领域，采用了 pSp 中使用的身份损失，使用预训练的 ArcFace<sup>[40]</sup> 面部识别网络来提取特征向量。此外，采用常用的  $\mathcal{L}_2$  和  $\mathcal{L}_{\text{LPIPS}}$ <sup>[67]</sup> 损失来学习像素级和感知级相似性。最终失真损失定义为：

$$\mathcal{L}_{\text{dist}}(x) = \lambda_{l2}\mathcal{L}_2(x) + \lambda_{lrips}\mathcal{L}_{\text{LPIPS}}(x) + \lambda_{\text{sim}}\mathcal{L}_{\text{sim}}(x) \quad (4.13)$$

为了提高感知质量和可编辑性，首先，应用了一个式 (4.9) 的  $\mathcal{L}_{\text{d-reg}}$  损失，以确保在学习偏移量  $\Delta_i$  时接近  $\mathcal{W}_*$ 。其次，使用了式 (4.10) 和式 (4.11) 的对抗损失以鼓励每个学习到的潜空间编码位于分布  $\mathcal{W}$  内。可编辑性损失具体由下式给出：

$$\mathcal{L}_{\text{edit}}(x) = \lambda_{\text{d-reg}}\mathcal{L}_{\text{d-reg}}(x) + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}(x) \quad (4.14)$$

整体损失目标被定义为失真度损失和可编辑性损失的加权组合：

$$\mathcal{L}(x) = \mathcal{L}_{\text{dist}}(x) + \lambda_{\text{edit}}\mathcal{L}_{\text{edit}}(x) \quad (4.15)$$

## 4.4 实验与分析

### 4.4.1 数据集与实现细节

关于掩膜网络,本文使用 StyleGAN2 生成器在 FFHQ 数据集<sup>[38]</sup>上训练。样式空间  $S$  的维度为 9088,其中 6048 个维度对应于卷积层,3040 个维度对应于生成器  $G$  的 ToRGB 块。在实验中,使用生成器的前 12 层(即 5632 个维度),因为其余的只会影响图像的细粒度细节,使用所有层导致在重演性能的微不足道的提升。设置身份损失权重为 10.0,形状损失权重为 1.0,训练模型进行 70K 次迭代,批大小为 6,使用 Adam 优化器<sup>[50]</sup>,学习率为  $10^{-4}$ 。

关于反演编码器,使用 FFHQ<sup>[38]</sup>数据集进行训练,使用 CelebA-HQ<sup>[68]</sup>测试数据集进行评估。设置损失权重如下: $\lambda_{l2} = 1$  和  $\lambda_{l_{lips}} = 0.8$ 。当使用潜空间鉴别器  $D_w$  进行训练时,使用  $\lambda_{adv} = 0.1$ 。当应用渐进式增长的增量训练时,使用  $\lambda_{d-reg} = 2e^{-4}$ 。对于人脸领域,在预训练的 ArcFace<sup>[40]</sup> 人脸识别网络上设置  $\lambda_{sim} = 0.1$ 。最后,设置  $\lambda_{edit} = 1$ 。在前 20,000 个训练步骤中,只有第一个  $w$  风格向量被训练。在前 20,000 步之后,每 2,000 个训练步骤逐渐增加下一个潜空间编码条目的增量。最后,批大小为 8,使用 Adam 优化器<sup>[50]</sup>,学习率为  $10^{-4}$ 。

### 4.4.2 实验效果

图 4.6展示了本文的人脸重演结果。实验结果表明本文的重演方法在保持源人脸身份信息方面表现出色。具体而言,它成功保留了诸如发型、发色以及面部基本形状等关键身份信息,确保了源人脸特征的连续性和一致性。同时,该方法在传递目标人脸的头部姿态和面部表情方面亦取得了显著成效。这体现在目标人脸的偏航、俯仰和滚动等头部姿态以及细微的面部表情变化被精准地映射到了合成图像上。这一观察结果清晰地表明了本文所采用的掩膜网络在面部姿态与身份特征分离方面的有效性。它不仅能够有效地将面部姿态信息从身份特征中剥离出来,而且在转换过程中避免了任何不必要的风格细节的转移。这一特性使得本方法在处理复杂多变的人脸图像时更具鲁棒性和通用性。

图 4.7展示了本文在纹理重建上的定性实验效果。最左边一列代表输入图片,涵盖了不同的身份、表情以及姿态信息。完成纹理重建后,将得到的具有纹理的 3D 人脸旋转以充分显示重建的纹理效果。实验结果表明本文的纹理重建方法在纹理质量及保真度方面表现出色,在处理各种不同情况的人脸的输入图片时能重建出稳定的高质量高保真纹理贴图。





图 4.6 人脸重演实验效果

4.4.3 对比实验

为了定量对比评估人脸重演的效果，本文采用了三个评估指标。首先，报告 ArcFace<sup>[40]</sup> 身份相似度评分，以衡量源图像和重新生成的图像之间的身份保存情况。此外，为了测量源和目标之间的头部姿态转移误差，本文计算了三个欧拉角 (即偏航，俯仰和滚转) 差异的 L1 距离，将其表示为“姿态误差”。同样，为了测量源和目标之间的面部表情传递误差，本研究计算表情系数差异的 L1 距离，将其记为“表情误差”。

在表 4.1 中，显示了在一组 5K 个随机生成的图像对上，将本文方法与最先进的 Fast Bi-layer<sup>[69]</sup>、ID-disentanglement (ID-d)<sup>[70]</sup>、PIR<sup>[71]</sup> 和 StyleFusion<sup>[72]</sup> 进行定量的比较评估。本研究的方法在所有指标上都优于 IDdisentanglement (ID-d)<sup>[70]</sup> 和 StyleFusion<sup>[72]</sup>。本研究法获得了与 PIR<sup>[71]</sup> 相近的身份相似度与表情误





图 4.7 纹理重建实验效果

差,但是在姿态误差方面表现要更好,与 Fast Bi-layer<sup>[69]</sup>在表情误差上取得了相近的效果,但是在身份相似度和姿态误差方面表现更好。

在图 4.8 中,将 PR3D 与两种最先进的方法进行了定性比较: DECA<sup>[15]</sup> 和 Deep3D<sup>[14]</sup>。值得注意的是,DECA<sup>[15]</sup>使用从原图提取的方法进行纹理重建,在侧视图作为输入时因自遮挡而导致纹理存在显著缺陷,并且由于从原图提取纹理的重建方法依赖于高精度的形状重建,而 DECA 的形状重建精度不足,这导致在眼睛和嘴巴等部位的纹理效果较差,在图 4.9 中针对具体部位的重建细节进行了对比。Deep3D<sup>[15]</sup>使用卷积神经网络来回归纹理系数,在构造的线性纹理空间中找到对应的纹理,但是由于构造的纹理空间与真实人脸纹理间存在域间隙,所以 Deep3D 在保真度方面存在不足,例如输入图像中存在的皱纹。相比之下,本研究的方法 PR3D 由于高精度的形状重建框架,在重建的形状和输入图片之间



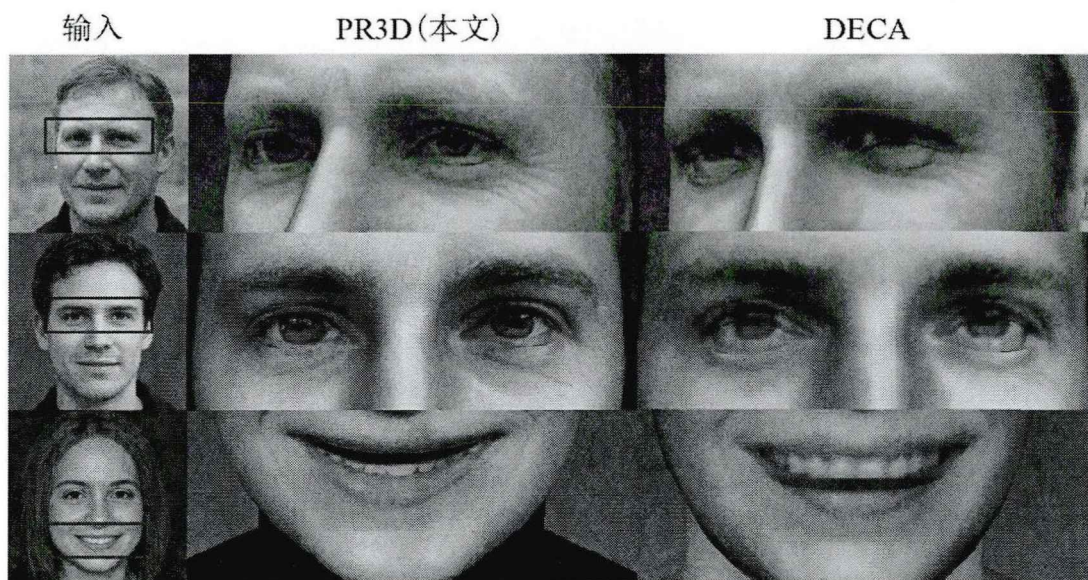
表 4.1 本文人脸重演与其他重演方法定量比较

| 方法                            | 身份相似度 | 姿态误差 | 表情误差 |
|-------------------------------|-------|------|------|
| Fast Bi-layer <sup>[69]</sup> | 0.43  | 1.3  | 0.10 |
| ID-d <sup>[70]</sup>          | 0.57  | 2.1  | 0.12 |
| PIR <sup>[71]</sup>           | 0.69  | 1.6  | 0.11 |
| StyleFusion <sup>[72]</sup>   | 0.63  | 3.75 | 0.12 |
| 本研究                           | 0.69  | 1.1  | 0.11 |

形成了很好的对齐效果，并且由于 StyleGAN2 的逼真生成能力，补充了由于自遮挡等造成的纹理缺失，得到覆盖几乎整个面部区域的高分辨率纹理贴图。



图 4.8 本文纹理重建与其他方法的定性比较

图 4.9 与 DECA<sup>[15]</sup>在具体部位纹理重建细节的比较

## 4.5 本章小结

本章介绍了本文提出的基于 StyleGAN2 的高保真纹理重建算法。首先介绍了 StyleGAN2 的潜在空间和对于真实图片的反演权衡问题。然后介绍了基于 StyleGAN2 的纹理重建算法的提出动机与具体设计。接下来通过实验与分析来说明本文提出的纹理重建方法比现有的其他前沿方法达到了更高的保真度。



## 第5章 总结与展望

### 5.1 研究内容总结

本文提出了 PR3D(Precise and Realistic 3D face reconstruction), 它是一种能够从单张图像中精确且逼真地重建 3D 人脸的先进技术。PR3D 包括基于半监督学习的高精度形状重建算法, 以及基于 StyleGAN2 的高保真纹理重建算法, 显著提升了 3D 人脸重建的精确性与真实感。

在形状重建方面, 对于同一身份的图像, 姿势、配饰、光照和表情等因素不会影响它们的中性形状。以前的工作往往使用单个编码器来回归所有参数, 旨在将重建的 3D 人脸与输入图像(无论是在关键点上还是在渲染像素上)尽可能对齐。然而, 它们忽视了身份特征在确定 3D 形状中的关键作用, 往往导致同一身份的不同照片的重建结果差异显著。受此启发, 本文设计了一个新颖的半监督学习框架。该框架的核心包含两个关键组件: 身份编码器和辅助编码器。身份编码器负责从输入图像中提取与个体身份相关的特征, 并将其编码为 FLAME 形状参数; 而辅助编码器则提供额外的辅助信息, 如表情、姿态等。为了进一步优化 FLAME 参数, 引入了一个联合优化过程, 该过程基于可微分渲染技术, 通过最小化渲染结果与输入图像之间的差异来迭代调整参数。在半监督学习策略中, 本文提出了一种混合关键点损失函数。该函数结合了自然场景人脸图像的稀疏关键点损失和有标注人脸图像的密集关键点损失。通过同时利用这两种损失, 模型能够充分利用有标注数据集的精确标注信息, 同时从大量无标注的自然场景人脸图像中学习到丰富的形状变化模式。这种混合损失的设计使得 PR3D 方法在各种输入图像上表现出强大的鲁棒性, 特别是在面对具有严重遮挡或极端姿态变化的挑战性场景时, 仍能保持出色的重建性能。

为证明本文的基于半监督学习的形状重建方法的高精度, 与近期其他先进方法进行了定性和定量的对比实验。一是与其他人脸重建方法进行定性的视觉比较, 实验结果表明在不同身份、姿态、光照等条件下的人脸图片作为输入时, 本文的方法展现出更好的准确性和鲁棒性, 展现出最好的视觉效果。二是在 NoW 基准和 REALY 基准上的定量比较, 本文在两个基准上都达到了最先进的效果。

为了满足实时应用的需求, 本文进一步从 PR3D 中蒸馏出一个轻量级模型——fast-PR3D。该模型在保持较高重建精度的同时, 显著提升了重建速度, 从而在速度和准确性之间取得了良好的平衡。这使得 fast-PR3D 模型能够在实际应用中实现实时三维人脸重建, 为各种实时交互和动态场景提供了有力的支持。

在纹理重建方面。基于构造的方法致力于构建一个纹理空间, 并寻找与输入图像最为匹配的纹理, 然而与真实的纹理空间相比, 构建的纹理空间往往存在显



著的差异,这可能导致重建结果的纹理不够真实或自然。另一类方法是基于提取的方法,它们直接从输入图像中提取纹理信息,然而由于侧视图像中的自遮挡问题,所得到的纹理贴图往往存在缺陷,如部分区域纹理信息缺失或失真。本文提出了一种基于 StyleGAN2 风格空间的人脸重演纹理提取方法。该方法利用预训练的 StyleGAN2 生成模型的强大生成能力,通过训练反演编码器和隐空间的掩膜网络,实现人脸重演,然后从源图像和具有互补姿态的重演图像中提取纹理信息。通过结合这两种来源的纹理信息,成功地补充了原始图像中由于自遮挡引起的缺失纹理区域,并生成了高质量、高保真度的纹理贴图。这种方法显著提升了纹理重建的视觉效果和逼真度,使得重建结果更加接近真实人脸的外观。

为证明本文的纹理重建方法的高保真度,与其他先进方法进行对比。实验结果显示了本文所采用的人脸重演方法在面部姿态与身份特征分离方面的有效性,还证明了本文纹理重建方法在纹理质量及保真度方面表现出色,优于其他纹理重建工作,在处理各种不同情况的人脸的输入图片时能重建出稳定的高质量高保真纹理贴图。

综上所述,本文关注于基于单幅人脸图像的高质量三维人脸重建算法研究,兼顾了三维人脸重建领域的形状重建、纹理重建以及实时重建部分,实验结果表明了本文提出方法的优越性。

## 5.2 未来工作展望

本文针对基于单幅人脸图像的三维人脸重建问题提出了 PR3D 方法,以实现在形状和纹理方面高质量的三维人脸重建。但是还有以下方向可以进行改进和探索:

在形状重建方面,由于依赖于 FLAME 进行三维人脸重建,所以同时受到了其表达能力的限制,无法恢复人脸的高频细节,如皱纹、酒窝等,下一步工作可以试图通过引入细节图或非线性运算来恢复生动的高频细节。第二,本文在嘴部的形状重建精度和其他部位相比存在不足,未来可以针对嘴部,结合人脸分割模型设计相关损失函数,以提升在嘴部的重建效果。此外,探究数据集(比如不同种族、年龄)与实际效果之间的关系也是很有意义的研究方向,因为不同种族或年龄之间的五官特征可能会有一些差异,但需要注意的是,这些特征并不是绝对的,因为人种之间也存在很大的个体差异。

在纹理重建方面,由于依赖于 StyleGAN2 的隐空间,而对于一些风格差异较大的输入图片,其反演得到的潜在编码并不能很好地反应真实图片,下一步工作可以专注于提升 StyleGAN2 在人脸域的生成能力,使其对于各种不同情况的输入人脸图片都能实现高质量高保真的纹理重建效果。

## 参考文献

- [1] 何龙健, 钟子乐, 邹大辉, 等. 面向医疗整容的三维人脸重建与编辑系统[J]. 计算机系统应用, 2022, 31(12): 69-77.
- [2] HU L, SAITO S, WEI L, et al. Avatar digitization from a single image for real-time rendering [J]. ACM Transactions on Graphics (ToG), 2017, 36(6): 1-14.
- [3] BLANZ V, ROMDHANI S, VETTER T. Face identification across different poses and illuminations with a 3d morphable model[C]//Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition. 2002: 202-207.
- [4] CAO C, WENG Y, LIN S, et al. 3d shape regression for real-time facial animation[J]. ACM Transactions on Graphics (TOG), 2013, 32(4): 1-10.
- [5] BLANZ V, VETTER T. A morphable model for the synthesis of 3d faces[M]//Seminal Graphics Papers: Pushing the Boundaries, Volume 2. 2023: 157-164.
- [6] JACKSON A S, BULAT A, ARGYRIOU V, et al. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. 2017: 1031-1039.
- [7] DOU P, SHAH S K, KAKADIARIS I A. End-to-end 3d face reconstruction with deep neural networks[C]//Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2017: 5908-5917.
- [8] PAYSAN P, KNOTHE R, AMBERG B, et al. A 3d face model for pose and illumination invariant face recognition[C]//Proceedings of the Sixth IEEE International Conference on Advanced Video and Signal-Based Surveillance. 2009: 296-301.
- [9] LI T, BOLKART T, BLACK M J, et al. Learning a model of facial shape and expression from 4d scans.[J]. ACM Transactions on Graphics (ToG), 2017, 36(6): 194-1.
- [10] BAI H, KANG D, ZHANG H, et al. Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 362-371.
- [11] ALDRIAN O, SMITH W A. Inverse rendering of faces with a 3d morphable model[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(5): 1080-1093.
- [12] BAS A, SMITH W A, BOLKART T, et al. Fitting a 3d morphable model to edges: A comparison between hard and soft correspondences[C]//Proceedings of the 2017 Asian Conference on Computer Vision. 2017: 377-391.
- [13] GERIG T, MOREL-FORSTER A, BLUMER C, et al. Morphable face models-an open framework[C]//Proceedings of the Thirteenth IEEE International Conference on Automatic Face and

- Gesture Recognition. 2018: 75-82.
- [14] DENG Y, YANG J, XU S, et al. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1-11.
- [15] FENG Y, FENG H, BLACK M J, et al. Learning an animatable detailed 3d face model from in-the-wild images[J]. ACM Transactions on Graphics (ToG), 2021, 40(4): 1-13.
- [16] WOOD E, BALTRUŠAITIS T, HEWITT C, et al. 3d face reconstruction with dense landmarks [C]//Proceedings of the Seventeenth European Conference on Computer Vision. 2022: 160-177.
- [17] ZIELONKA W, BOLKART T, THIES J. Towards metrical reconstruction of human faces[C]// Proceedings of the Seventeenth European Conference on Computer Vision. 2022: 250-269.
- [18] GENOVA K, COLE F, MASCHINOT A, et al. Unsupervised training for 3d morphable model regression[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 8377-8386.
- [19] 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)[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. 2017: 1021-1030.
- [20] CHEN Z, WANG Y, GUAN T, et al. Transformer-based 3d face reconstruction with end-to-end shape-preserved domain transfer[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(12): 8383-8393.
- [21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the Thirty-First Annual Conference on Neural Information Processing Systems. 2017: 5998-6008.
- [22] GUO J, ZHU X, YANG Y, et al. Towards fast, accurate and stable 3d dense face alignment [C]//Proceedings of the Sixteenth European Conference on Computer Vision. 2020: 152-168.
- [23] ZHU X, LIU X, LEI Z, et al. Face alignment in full pose range: A 3d total solution[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 41(1): 78-92.
- [24] GECER B, PLOUMPIS S, KOTSIA I, et al. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1155-1164.
- [25] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks [J]. Communications of the ACM, 2020, 63(11): 139-144.
- [26] DENG J, CHENG S, XUE N, et al. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7093-7102.

- [27] LEE G H, LEE S W. Uncertainty-aware mesh decoder for high fidelity 3d face reconstruction [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6100-6109.
- [28] LATTAS A, MOSCHOGLOU S, GECER B, et al. Avatarme: Realistically renderable 3d facial reconstruction" in-the-wild"[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 760-769.
- [29] LUO H, NAGANO K, KUNG H W, et al. Normalized avatar synthesis using stylegan and perceptual refinement[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 11662-11672.
- [30] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of stylegan[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8110-8119.
- [31] SANYAL S, BOLKART T, FENG H, et al. Learning to regress 3d face shape and expression from an image without 3d supervision[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7763-7772.
- [32] AMBERG B, ROMDHANI S, VETTER T. Optimal step nonrigid icp algorithms for surface registration[C]//Proceedings of the 2007 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2007: 1-8.
- [33] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: A unified embedding for face recognition and clustering[C]//Proceedings of the 2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2015: 815-823.
- [34] KENDALL A, GAL Y. What uncertainties do we need in bayesian deep learning for computer vision?[C]//Proceedings of the Thirty-First Annual Conference on Neural Information Processing Systems. 2017: 5574-5584.
- [35] ISOLA P, ZHU J Y, ZHOU T, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2017: 1125-1134.
- [36] NIRKIN Y, MASI I, TUAN A T, et al. On face segmentation, face swapping, and face perception[C]//Proceedings of the Thirteenth IEEE International Conference on Automatic Face and Gesture Recognition. 2018: 98-105.
- [37] WANG Y, TAO X, QI X, et al. Image inpainting via generative multi-column convolutional neural networks[C]//Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems. 2018: 329-338.
- [38] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern

- Recognition. 2019: 4401-4410.
- [39] CHAI Z, ZHANG H, REN J, et al. Realy: Rethinking the evaluation of 3d face reconstruction [C]//Proceedings of the Seventeenth European Conference on Computer Vision. 2022: 74-92.
- [40] DENG J, GUO J, LIU T, et al. Sub-center arcface: Boosting face recognition by large-scale noisy web faces[C]//Proceedings of the Sixteenth European Conference on Computer Vision. 2020: 741-757.
- [41] RAMAMOORTHY R, HANRAHAN P. An efficient representation for irradiance environment maps[C]//Proceedings of the Twenty-Eighth Annual Conference on Computer Graphics and Interactive Techniques. 2001: 497-500.
- [42] facebookresearch. Pytorch3d[EB/OL]. 2020. <https://github.com/facebookresearch/pytorch3d>.
- [43] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
- [44] LUGARESI C, TANG J, NASH H, et al. Mediapipe: A framework for perceiving and processing reality[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1-4.
- [45] zllrunning. Face-parsing.pytorch[EB/OL]. 2018. <https://github.com/zllrunning/face-parsing.PyTorch>.
- [46] CAO Q, SHEN L, XIE W, et al. Vggface2: A dataset for recognising faces across pose and age[C]//Proceedings of the Thirteenth IEEE International Conference on Automatic Face and Gesture Recognition. 2018: 67-74.
- [47] AN X, ZHU X, GAO Y, et al. Partial fc: Training 10 million identities on a single machine [C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. 2021: 1445-1449.
- [48] LIU Z, LUO P, WANG X, et al. Deep learning face attributes in the wild[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision. 2015: 3730-3738.
- [49] PASZKE A, GROSS S, MASSA F, et al. Pytorch: An imperative style, high-performance deep learning library[C]//Proceedings of the Thirty-Third Annual Conference on Neural Information Processing Systems. 2019: 8024-8035.
- [50] KINGMA D P, BA J. Adam: A method for stochastic optimization[A]. 2014. arXiv preprint arXiv:1412.6980.
- [51] LOSHCHELOV I, HUTTER F. Fixing weight decay regularization in adam[A]. 2018. arXiv preprint arXiv:1711.05101.
- [52] RAI A, GUPTA H, PANDEY A, et al. Towards realistic generative 3d face models[A]. 2023. arXiv preprint arXiv:2304.12483.



- [53] FENG Y, WU F, SHAO X, et al. Joint 3d face reconstruction and dense alignment with position map regression network[C]//Proceedings of the Fifteenth European Conference on Computer Vision. 2018: 534-551.
- [54] DANĚČEK R, BLACK M J, BOLKART T. Emoca: Emotion driven monocular face capture and animation[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 20311-20322.
- [55] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 4510-4520.
- [56] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[A]. 2017. arXiv preprint arXiv:1704.04861.
- [57] TZELEPIS C, TZIMIROPOULOS G, PATRAS I. Warpedganspace: Finding non-linear rbf paths in gan latent space[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. 2021: 6393-6402.
- [58] WU Z, LISCHINSKI D, SHECHTMAN E. Stylespace analysis: Disentangled controls for stylegan image generation[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12863-12872.
- [59] ABDAL R, QIN Y, WONKA P. Image2stylegan: How to embed images into the stylegan latent space?[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. 2019: 4432-4441.
- [60] BOUNARELI S, TZELEPIS C, ARGYRIOU V, et al. Stylemask: Disentangling the style space of stylegan2 for neural face reenactment[C]//Proceedings of the Seventeenth IEEE International Conference on Automatic Face and Gesture Recognition. 2023: 1-8.
- [61] TOV O, ALALUF Y, NITZAN Y, et al. Designing an encoder for stylegan image manipulation [J]. ACM Transactions on Graphics (TOG), 2021, 40(4): 1-14.
- [62] SANCHEZ E, VALSTAR M. A recurrent cycle consistency loss for progressive face-to-face synthesis[C]//Proceedings of the Fifteenth IEEE International Conference on Automatic Face and Gesture Recognition. 2020: 53-60.
- [63] NITZAN Y, BERMANO A, LI Y, et al. Face identity disentanglement via latent space mapping [A]. 2020. arXiv preprint arXiv:2005.07728.
- [64] MESCHEDER L, GEIGER A, NOWOZIN S. Which training methods for gans do actually converge?[C]//Proceedings of the 2018 International Conference on Machine Learning. 2018: 3481-3490.
- [65] RICHARDSON E, ALALUF Y, PATASHNIK O, et al. Encoding in style: a stylegan encoder for image-to-image translation[C]//Proceedings of the 2021 IEEE/CVF Conference on

- Computer Vision and Pattern Recognition. 2021: 2287-2296.
- [66] CHEN X, FAN H, GIRSHICK R, et al. Improved baselines with momentum contrastive learning[A]. 2020. arXiv preprint arXiv:2003.04297.
- [67] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 586-595.
- [68] KARRAS T, AILA T, LAINE S, et al. Progressive growing of gans for improved quality, stability, and variation[A]. 2017. arXiv preprint arXiv:1710.10196.
- [69] ZAKHAROV E, IVAKHNENKO A, SHYSHEYA A, et al. Fast bi-layer neural synthesis of one-shot realistic head avatars[C]//Proceedings of the Sixteenth European Conference on Computer Vision. 2020: 524-540.
- [70] NITZAN Y, BERMANO A, LI Y, et al. Face identity disentanglement via latent space mapping [A]. 2020. arXiv preprint arXiv:2005.07728.
- [71] REN Y, LI G, CHEN Y, et al. Pirenderer: Controllable portrait image generation via semantic neural rendering[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. 2021: 13759-13768.
- [72] KAFRI O, PATASHNIK O, ALALUF Y, et al. Stylefusion: A generative model for disentangling spatial segments[A]. 2021. arXiv preprint arXiv:2107.07437.