

硕 士 研 究 生 读 书 报 告



题目 FDNeRF: Few-shot Dynamic Neural Radiance Fields for Face Reconstruction and Expression Editing读书报告

作者姓名 章腾

作者学号 22251169

指导教师 李启雷

学科专业 软件工程

所在学院 软件学院

提交日期 2022年12月

Reading Report of FDNeRF: Few-shot Dynamic Neural Radiance Fields for Face Reconstruction and Expression Editing

A Dissertation Submitted to

Zhejiang University

in partial fulfillment of the requirements for

the degree of

Master of Engineering

Major Subject: Software Engineering

Advisor: Qilei Li

By

Zhang Teng

Zhejiang University, P.R. China

2022

# 摘要

NeRF模型及其改进在机器人、城市测绘、自动导航、虚拟现实/增强现实等领域得到了广泛应用。现阶段不少研究人员将NeRF实现虚拟人作为重要的研究课题。在这篇论文中，作者提出了一种少镜头动态神经辐射场，能够从单目视频中提取的少量动态帧重建和编辑3D人脸的方法。并在其中引入了一种CFW模块，即在二维特征空间中将不同帧表情扭曲至一致，该模块具有标识自适应和三维约束能力。同时该架构支持面部表情的自由编辑，并支持视频驱动的3D重演。

**关键词：**三维人脸重建，表情编辑，NeRF，少样本和动态建模

Abstract

NeRF models and improvements are used in robotics, urban mapping, automated navigation, virtual/augmented reality, and more. At present, many researchers regard the realization of virtual human by NeRF as an important research topic. In this paper, the authors propose a method named Few-shot Dynamic Neural Radiance Fields which can reconstruct and edit 3D faces from a small number of dynamic frames extracted from monocular video. In addition, a CFW module is introduced to perform different expression conditioned warping in 2D feature space to deal with the inconsistency between dynamic frames. This module also has identity adaptive and 3D constrained. The architecture also supports free edits of facial expressions, and enables video-driven 3D reenactment.

**Keywords：**3D face reconstruction, expression editing, NeRF, few-shot and

dynamic modeling

1 引言

NeRF即神经辐射场，是一种用于三维场景建模的深度学习方法，它能够使用少量的监督信息，在不同的光照条件下生成图像。具体来说就是将全连接神经网络引入到物体的3维场景表示中。只需要同一物体不同角度的若干张图片作为监督，神经网络可以隐式地对该物体进行三维场景建模（获取三维几何和表面纹理信息），然后在新视角下通过体渲染（volume rendering）的方法渲染生成新的角度的二维图像。这种方法是与传统图形学渲染方法中通过物理学原理进行渲染完全不同的。因此Nerf作为一种新颖的视图合成和三维重建方法，NeRF模型及其改进在机器人、城市测绘、自动导航、虚拟现实/增强现实等领域得到了广泛应用。成为了研究的巨大热点。诸如Nerf的复杂度优化、动态场景人物建模、稀疏视角建模、质量优化等等问题，已经诞生了许多文章。

这篇文章提出一种少镜头动态神经辐射场（Few-shot Dynamic Neural Radiance Fields），即一个基于nerf的从单目视频中提取的少量动态帧重建和编辑3D人脸的方法。与现有的需要密集图像作为输入并只能对单一人物进行建模的动态神经辐射场不同，该篇所提出方法能够通过少量的输入来重建不同人的脸部。其能够接受视图不一致的动态输入，并支持任意的面部表情编辑，产生具有超出输入的新面部表情。

同时在本篇文章中，FDNeRF在处理中采用了一种新颖的带有3D约束的条件特征扭曲(CFW)模块，通过在2D特征空间中将源表情扭曲为目标表情来处理动态帧之间的不一致性。然后再采用重构模块，根据弯曲的特征空间预测亮度场空间点的颜色和密度。最终通过体渲染方法得到表情结果。

# 2方法

## 2.1方法概述

FDNeRF的整体结构如下所示。

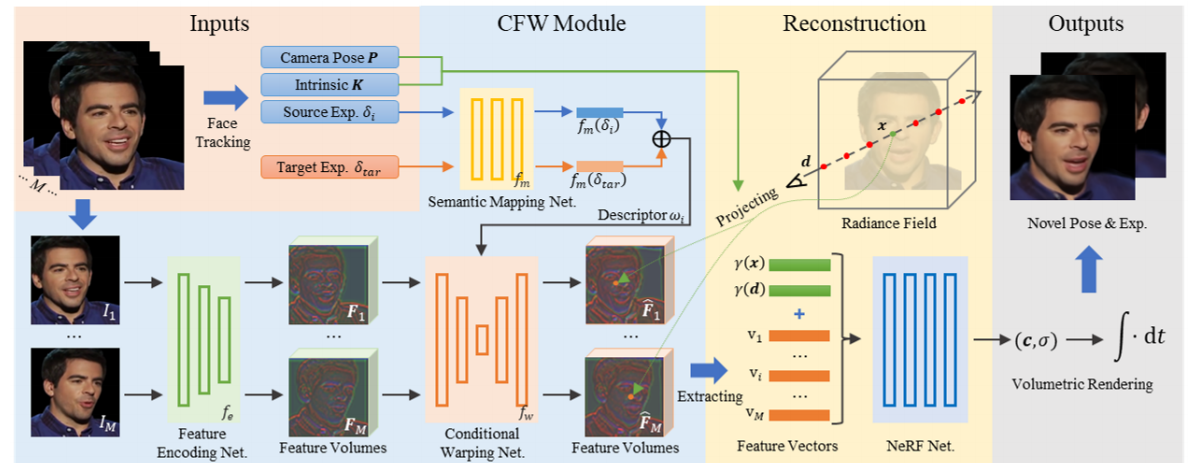


图1. FDNeRF架构图

从整体上来看， FDNeRF为了能达成最终的表情效果，其分成了四个阶段进行处理。（对应以上不同颜色部分的处理方式）

1. 在预处理阶段，给定少量动态图像，在该阶段实现人脸跟踪，估计相关表情参数𝛿𝑖、相机姿态P和内在矩阵K。
2. 在CFW模块中，使用特征编码网络𝑓𝑒为每一张图像𝐼𝑖提取深度特征体积𝐹𝑖，使用语义映射网络𝑓𝑚根据源和目标表达式参数生成运动描述符𝜔𝑖。然后使用描述符指导条件扭曲网络𝑓𝑤生成扭曲特征体积𝐹𝑖。
3. 在辐射场重建过程中，将查询点𝑥投影到每个图像平面，并提取对齐的特征向量𝑣𝑖。将这些向量(在CFW中抽取得到的Fi)，连同点的位置和方向，一起输入到NeRF网络来推断颜色和密度。
4. 最后通过体渲染来合成新的视图图像。

## 2.2 CFW（特征扭曲）模块和预处理

CFW模块部分主要由三个子网络组成，分别是类resnet特征编码网络𝑓𝑒、语义映射网络𝑓𝑚和条件翘曲网络𝑓𝑤。在该部分使用编码网络𝑓𝑒为每个输入帧𝐼𝑖获得一个深度特征卷𝐹𝑖，它对身份和表达式信息进行编码。其表达式如下（其中𝐹𝑖由从𝑓𝑒的前四层提取的特征图组成。）：

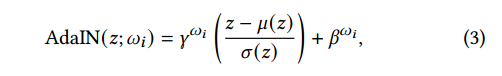


语义映射网络𝑓𝑚则提取引导特征体积𝐹𝑖扭曲的语义条件。该部分利用了现成的Face2face面部跟踪方法估计每个输入帧的表达式参数𝛿，面部姿态P和内在矩阵K。随后，𝑓𝑚将原始参数转换为潜在代码𝑓𝑚(𝛿𝑖)和𝑓𝑚(𝛿𝑡𝑎𝑟)在扭曲网络中提取更多的判别表示，实现细粒度引导。

同时该部分对于每一帧未对齐的帧，将其潜在代码𝑓𝑚(𝛿𝑖)与目标表达式代码𝑓𝑚(𝛿𝑡𝑎𝑟)连接起来，形成高维运动描述符𝜔𝑖来引导扭曲网络：



而对于条件扭曲网络𝑓𝑤，则使用类似编码器和解码器的结构。同时为了更充分地引导扭曲网络，本篇文章通过自适应实例规范化(AdaIN)操作符将运动描述符𝜔𝑖注入到𝑓𝑤的所有卷积层中。具体如下式所示，就是通过一个轻量级映射网络将把运动描述符𝜔𝑖分别转换为仿射参数𝛾𝜔𝑖和𝛽𝜔𝑖。（其中𝜇(·)和𝜎(·)计算𝑧的平均值和方差统计。）



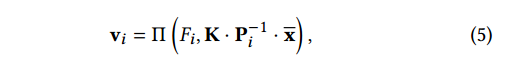
基于特征体积𝐹𝑖和描述符𝜔𝑖，扭曲网络𝑓𝑤将估计出一个变形流场，该变形流场表示输入特征体积𝐹𝑖与期望目标特征体积𝐹𝑖之间的坐标偏移量。最终通过双线性插值采样得到对齐的特征体（𝑓𝑤(𝐹𝑖，𝜔𝑖)表示扭曲网络估计的变形流场，Sample(𝑎，𝑏)表示根据流场𝑏在𝑎上进行插值采样操作。）：



## 2.4 辐射场重建

本篇文章的任务是从目标特征体𝐹重建具有所需表情的3D人脸，因此采用了类似MoFaNeRF的框架作为的重建模块，推导出每个空间点的颜色和密度，然后使用体绘制生成最终的几何形状和外观。

具体来说，其首先通过目标视图的每个像素投射相机光线，并沿着每个光线采样𝑁点进行体绘制（原始NeRF采用的方式），然后，我们利用上一阶段得到的内在矩阵K和对应的位姿P𝑖将光线上的每个采样点𝑝投影到每个帧坐标上，并通过双线性插值从目标特征体𝐹𝑖中提取相关的对齐特征向量v𝑖。（其中Π表示提取过程，K·p\_i^(-1)·x表示𝑖-th框架平面上的坐标，x是点𝑝的齐次坐标）



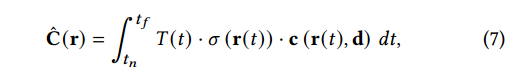
接着再将特征向量以及查询点𝑝的位置x和方向d输入到重构模块中，估计颜色c和密度𝜎值:（𝛾(·)是在NeRF当中经常被使用的将输入映射到高维傅里叶空间的函数，𝐺(·)是由神经网络形成的平均函数，以收集所有可用的信息。𝑀表示输入帧数）



同时在这一阶段中，为了消除视图之间的几何差异，其并未在神经网络开始开始处输入方向分量𝛾(d)来影响密度相关参数。相反，而仅将它放置输入到最后几个层中，只用以调整与颜色相关的参数。

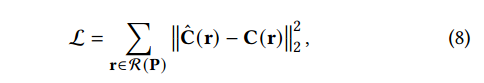
## 2.5 体渲染

这篇文章中并没有对体渲染方法做出改进，体渲染也是NeRF进行渲染的一般方法，渲染图像中每个像素的期望颜色C可以通过累积沿摄像机射线r的所有采样点的估计颜色C和密度𝜎来计算：（其中r(𝑡)= o +𝑡d表示从摄像机中心o出发的射线的点位置，𝑡𝑛和𝑡𝑓分别是射线的近边界和远边界）



## 2.6 神经网络优化方法

这篇文章中使用光度重建损失对CFW模块和重构模块部分的神经网络进行联合优化。其可以用下式描述（其中R (P)表示姿态P中的摄像机光线集合，C(R)表示目标图像中的像素颜色。）



在具体优化过程中，每次训练时都是从训练视频中随机选择𝑀（1-12中随机选择）帧作为输入帧，从其余帧中随机选择一帧作为目标帧。然后，利用输入帧和目标帧的表达式参数来指导CFW模块中的特征扭曲过程。

# 3 实验与结果

这篇文章中使用VoxCeleb数据集中的213个谈话视频作为实验数据集。通过对这些视频进行单目面部跟踪方法的预处理来估计每一帧的表情语义、面部姿势和内在矩阵，并将面部姿态作为图像的相机姿态。同时为了验证其少样本特性，在实验中只使用了三帧作为输入帧。

定量分析的结果如下图所示。在三种主流的NeRF评价指标上（峰值信噪比(PSNR)、结构相似指数度量(SSIM)]和学习感知图像块相似度性(LPIPS)），FDNeRF生成的结果在该数据集上的表现都达到了最高性能。

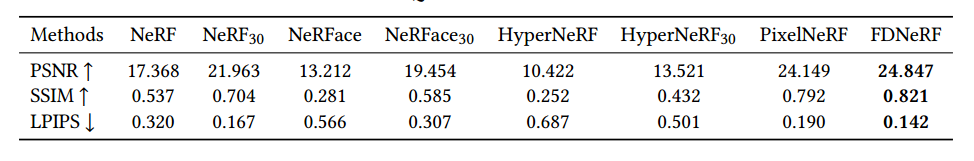


图2 评价指标比较

其中30下标表示该种NeRF是使用30帧输入图像进行训练，因为NeRF、NeRFFace、HyperNeRF都需要密集的3D建模输入视图，因此作者从数据集视频中提取额外的27帧以提高其性能。遗憾的是关于这个结果，作者没有给出实际图像效果的对比，仅给出了图像差异描述，因此我们就并不再深入探究。

作者还对FDNeRF的表情编辑能力进行了实验。并与NeRFace和MoFaNeRF两种同样能进行表情编辑的神经网络进行了横向对比。从下图中可以直观的看到FDNeRF具有最好的图像真实感，并且具有最少的图像噪音、在面部表情控制上也做得很好。



图3 表情编辑结果对比

在实验部分的最后，作者对视频驱动的表情重演进行了探索。在该部分，直接使用将视频序列的表情参数来应用于视频驱动的3D重演。但作者发现如果直接将帧信息（姿态、目标表情等）输入神经网络，相邻帧的估计参数之间的轻微不一致可能会导致再现结果中的不连续伪影。因此作者将CFW模块中的语义映射网络修改为接收一组参数，使其能够以一个连续帧的窗口的参数作为目标表达式语义，其中参数窗口设置为向前和向后的L长度帧（作者实验时使用的长度为13）。最终得到了一个良好的视频驱动的重演结果。



图4 基于目标帧驱动生成表情帧

# 3 总结

该篇论文从提出了一个使用从说话的人物头部单目视频中提取的少量动态帧的FDNeRF用于三维人脸重建和表情编辑。同时为了消除动态帧之间的不一致性，作者设计了基于表达式的特征扭曲模块，并设计了亮度场重建模块，以实现特征对齐的精确三维重建。并且用基于窗口的策略对该模型进行了扩展，使其可以进行时间连贯的视频驱动的重演。

但作者在最后也写到，FDNeRF在非面部区域的不一致性上存在局限。非面部区域(例如，头发和躯干)的不一致，不以表情为条件，将导致结果中产生一些模糊。这可能需要对其余部位也建立扭曲场来解决该问题。

但基本上来说，这篇文章是三维人脸重建和表情编辑领域的一个很大进步，一方面其极大的减少了训练所需的样本数量（仅需要三个单目样本），使得更多的数据能够被利用到该领域中，同时少样本也就带来了训练成本的下降，这也更符合工业界的要求，让其有了落地的可能。此外其3维约束的2维特征扭曲的思路，不仅仅只适应于面部表情变化，对于其他非刚性变化，同样能够采用该思路构建这个流程进行处理，是非常具有启发意义的。

最后，我认为该篇论文的一个最大缺点就是并未给出内部使用的神经网络更严格的定义，仅用算符来描述，仅说明在某处使用了某个神经网络，而没给出内部架构（也就不知道和现有神经网络架构到底有什么区别，作者进行了何种修改）。同时目前该篇论文的源码仍然是coming soon阶段，让人读完之后非常苦恼。此外我认为该篇论文实验部分相对单薄，一方面可能是面部表情的数据集本身较少，另一方面可能也是版面限制，导致其在实验部分对于结果的放出也不够多，实验结果并不是那么的有说服力。

# 4 参考文献

[1] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. 2022. FDNeRF: Few-shot Dynamic Neural Radiance Fields for Face Reconstruction and Expression Editing. In SIGGRAPH Asia 2022 Conference Papers (SA '22). Association for Computing Machinery, New York, NY, USA, Article 12, 1–9. https://doi.org/10.1145/3550469.3555404