Learning Dynamic Facial Radiance Fields for Few-Shot Talking Head Synthesis 阅读报告

张芷苒 PB21081601

2024 年 6 月 26 日

摘要

本次阅读报告选择 ECCV2022 paper "Learning Dynamic Facial Radiance Fields for Few-Shot Talking Head Synthesis"，该文章实现了一种用于少样本 talking head 合成的动态面部辐射场（DFRF），该方法能在少量训练数据下迅速将模型泛化到未知身份，生成高质量音频驱动的 talking head 视频。在我的大创项目中，我复现了这个项目，定量结果附在了报告后。

1 研究背景

谈话头像（talking head）合成是一种新兴技术，在电影配音、虚拟化身和在线教育中有广泛应用。

1.1 相关工作-Neural Radiance Fields

神经辐射场 (NeRF) 使用全连接网络以体素网格的形式存储 3D 几何和外观信息。NeRF 的核心思想是使用神经网络来表示场景中每个点的辐射亮度（radiance）。与传统的图像生成模型不同，NeRF 不仅仅生成颜色值，还生成了与场景中每个点的方向相关的辐射亮度。这允许生成高质量、高分辨率的三维重建，尤其在虚拟现实（VR）和计算机图形学领域具有潜在应用。

1.2 研究原因

现有的 talking head 技术存在一些局限性，阻碍了它们的实际应用。基于二维的方法难以生成较为自然的说话风格；基于 3D 的经典方法由于使用 3DMM 中间表示法，会造成信息损失；基于 NeRF 的方法可以合成比较自然的说话头像视频，但因为需要针对每个身份训练特定的模型，其计算成本相对较高，而且训练需要大量的数据集。因此，更具挑战性的对话头合成任务设置要求对于一个仅有简短训练视频片段的任意人物，只需经过少量次数的迭代，就能构建出高质量的个性化音频驱动人像视频。DFRF（动态面部辐射场）模型很好地完成了这个任务。

不同于现有的直接将三维几何和外观编码到网络中的 NeRF 方法，DFRF 利用少量的静态头像图像数据，通过学习动态面部辐射场，生成逼真的动态头像。这一过程涉及到从静态图像中提取面部特征，并通过辐射场模型将这些特征转换为动态图像序列。因此，面部辐射场只需少量参考图像就可以灵活地调整到新的身份。此外，为了更好地建模面部变形，DFRF 还引入了一个基于音频信号的可微面部变形模块，将参考图像变形到查询空间。

仅使用数十秒的训练片段，DFRF 可以在 4 万次迭代中为新的身份合成自然且高质量的音频驱动 talking head 视频。

2 模型框架

下图展示了 DFRF 模型的框架。[SLZ+22]

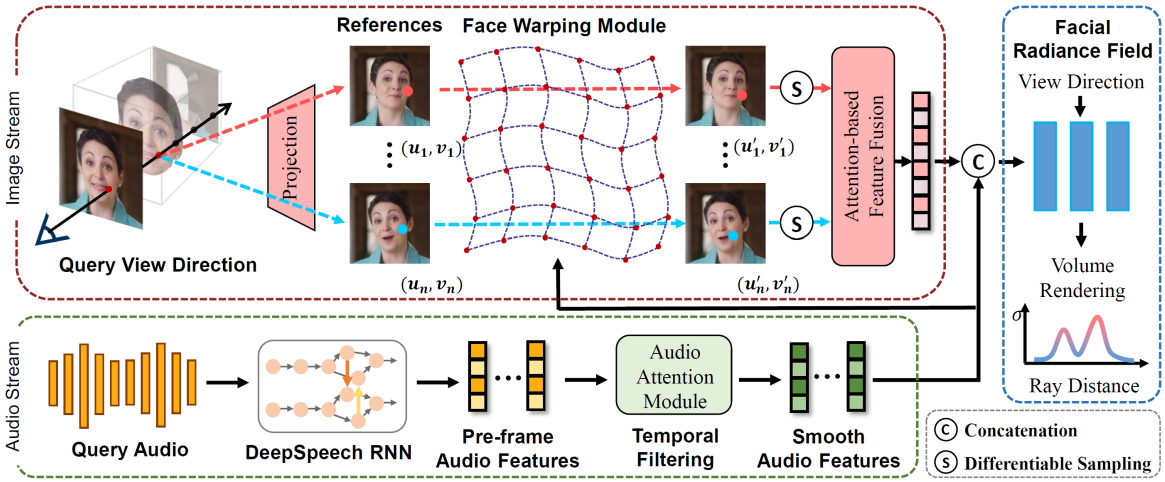

图 1: DFRF 模型示意图

DFRF 模型的结构可以分为以下几个主要模块：

2.1 图像流部分（Image Stream）

- 输入图像：输入的是查询视角方向的图像。
- 参考图像（References）：多张参考图像，用于辅助生成动态头像。
- 面部变形模块（Face Warping Module）：该模块将参考图像变形到查询视角。通过变形，将参考图像中的特征点（如 (u_1, v_1), (u_n, v_n)）映射到新的特征点（如 (u'_1, v'_1), (u'_n, v'_n)）。
- 特征融合模块（Attention-based Feature Fusion）：该模块基于注意力机制（Attention Mechanism）将变形后的图像特征进行融合。

2.2 音频流部分（Audio Stream）

- 查询音频（Query Audio）：输入的是音频信号。
- DeepSpeech RNN：使用递归神经网络（RNN）从音频信号中提取特征。
- 音频特征提取（Pre-frame Audio Features）：提取每帧音频的特征。
- 时间过滤（Temporal Filtering）：对提取的音频特征进行时间上的过滤和平滑。
- 音频注意力模块（Audio Attention Module）：进一步处理音频特征，生成平滑的音频特征（Smooth Audio Features）。

2.3 面部辐射场（Facial Radiance Field）

- 输入视角方向：结合图像流和音频流的特征，输入到面部辐射场模块。
- 体积渲染（Volume Rendering）：通过体积渲染技术，生成图像序列。这一步包括从射线距离中进行体积密度（σ）的计算。

2.4 特征融合与采样（Feature Fusion and Sampling）

- 特征拼接（Concatenation）：将图像流和音频流的特征拼接在一起。
- 可微采样（Differentiable Sampling）：对拼接后的特征进行可微分的采样，生成最终的动态图像。

模型的工作流程如下：

1. 图像和音频输入：模型接收查询视角方向的图像和音频信号作为输入。
2. 特征提取和变形：从输入图像中提取特征点，并通过面部变形模块将参考图像的特征点映射到新的视角。
3. 音频特征处理：通过 DeepSpeech RNN 从音频信号中提取特征，并通过时间过滤和注意力模块生成平滑的音频特征。
4. 特征融合：将图像流和音频流的特征通过注意力机制进行融合，并通过特征拼接将其整合在一起。
5. 面部辐射场和渲染：将融合后的特征输入到面部辐射场模块，通过体积渲染生成动态头像。

通过这种方式，DFRF 模型能够在少量的训练数据下，生成高质量的 talking head，并能够快速泛化到新的身份。

3 方法解析

3.1 问题陈述

现有谈话头像技术的一些局限性阻碍了其在实际中的应用。二维方法在生成自然的谈话风格方面存在困难。经典的三维方法由于使用中间 3DMM 表示，存在信息丧失。基于 NeRF 的方法合成了优质的谈话头像视频，但计算成本较高，因为每个身份需要特定的模型训练，并且需要大量数据集。因此，原论文关注更具挑战性的谈话头像合成任务。对于任意身份，只有一个简短的训练视频片段可用，一个高质量的个性化音频驱动肖像动画模型应在少量迭代内构建。

因此，原论文提出了一种动态面部辐射场（DFRF）用于少样本谈话头像合成。图像特征被引入作为条件，从参考图像到对应面部辐射场建立快速映射。为了更好地建模面部变形，原论文设计了一个可微面部扭曲模块，将参考图像变形到查询空间。具体来说，为了快速收敛，原论文首先在不同身份上训练一个基础模型，捕捉头部的结构信息，并建立从音频到唇形运动的通用映射。在此基础上，进行高效微调，快速泛化到新的目标身份。下面将详细介绍这些设计。

3.2 动态面部辐射场

新兴的 NeRF 提供了一种强大且优雅的三维场景表示框架，它通过一个 $\mathcal{F}_\theta$ 将场景编码到三维体积空间中，可以通过沿着摄像机射线集成颜色和密度将三维体积渲染成图像。具体来说，使用 P 作为体素空间中所有三维点的集合，三维查询点 $p = (x, y, z)$ P 和二维视图方向 $d = (\theta, \Phi)$ 作为输入，MLP 推断相应的 RGB 颜色 c 和密度 σ，公式为 $(c, \sigma) = \mathcal{F}(p, d)$。

在本工作中，原论文采用 NeRF 作为三维感知谈话头像建模的骨干。谈话头像任务关注音频驱动的面部动画。然而，原始 NeRF 仅设计用于静态场景。因此，原论文通过引入音频条件提供缺失的变形通道，如图1的音频流所示。原论文首先使用预训练的基于 RNN 的 DeepSpeech 模块提取每帧音频特征。为了实现帧间一致性，进一步引入一个时间过滤模块计算平滑的音频特征 A，这可以表示为其邻近音频特征的自注意力融合。使用这些音频特征序列 A 作为条件，原论文可以学习音频-唇形映射。这个音频驱动的面部辐射场公式为 $(c, \sigma) = \mathcal{F}(p, d, A)$。

由于身份信息隐含地编码到面部辐射场中，并且在渲染时没有提供显式身份特征，因此这个面部辐射场是针对个人的。对于实现谈话头像的其他方法而言，每个新身份都需要从头开始在大数据集上进行优化，这导致计算成本昂贵且需要长时间的视频训练。为了摆脱这些限制，原论文设计了一个参考机制，使经过良好训练的基础模型能够在仅有短视频片段的新身份上快速泛化。如图1所示，这种基于参考的架构概述。具体来说，以 N 个参考图像 $M = \{M_n \in \mathbb{R}^{H \times W} \mid 1 \leq n \leq N\}$ 及其相应的摄像机位置 T_n 作为输入，使用两层卷积网络计算其像素对齐的图像特征 $F = \{F_n \, \mathbb{R}^{H \times W \times D} | 1 \; n \; N\}$，不进行下采样。本工作中，特征维度 D 设为 128，H 和 W 分别表示图像的高度和宽度。使用多个参考图像提供了更好的多视角信息。对于三维查询点 $p = (x, y, z)$ P，原论文使用内参 $\{Kn\}$ 和摄像机姿态 $\{R_n, T_n\}$ 将其投影回这些参考图像的二维空间，并得到相应的二维坐标。使用 $p_n^{ref} = (u_n, v_n)$ 表示第 n 个参考图像中的二维坐标，这个投影公式为：

$$p_n^{ref} = \mathcal{M}(p, K_n, R_n, T_n), \tag{1}$$

其中 M 是从世界空间到图像空间的传统映射。这些对应的像素级特征 $\{F_n(u_n, v_n)\}$ $\mathbb{R}^{N \times D}$ 从 N 个参考图像中采样并通过注意力模块融合，得到最终特征 $\tilde{F} = Aggregation(\{F_n(u_n, v_n)\})$ $\mathbb{R}^D$。这些特征网格包含丰富的身份和外观信息。使用它们作为面部辐射场的附加条件，使模型能够从少数观测帧快速泛化到新面部外观。最终公式为：

$$(c, \sigma) = \mathcal{F}_\theta(p, d, A, \tilde{F}). \tag{2}$$

3.3 可微面部扭曲

在第 3.2 节中，原论文将查询三维点投影回这些参考图像的二维空间，如公式1，以获得条件像素特征。这种操作基于 NeRF 的先验知识，即从不同视点投射的相交射线应对应相同的物理位置，从而产生相同的颜色。这种严格的空间映射关系在刚性场景中成立，但谈话面部是动态的。当说话时，唇部和其他面部肌肉会根据发音移动。直接应用公式1 到可变形的谈话面部可能导致关键点不匹配，例如标准体积空间中接近嘴角的三维点被映射回参考图像的像素空间，如果参考面部显示不同的嘴形，则映射点可能偏离实际的嘴角。这样的不准确映射会导致从参考图像获取的条件像素特征错误，进一步影响说话嘴部变形的预测。

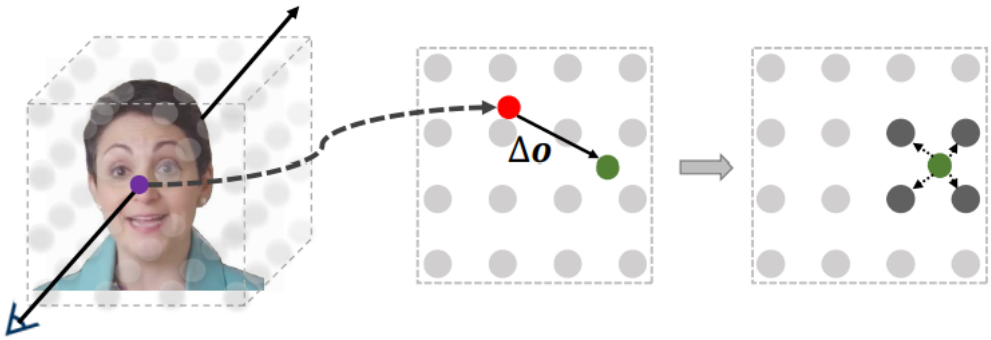

图 2: 可微分人脸变形的可视化示意图。一查询的 3D 点（紫色）被投影到参考图像空间（红色）。然后学习一个偏移 Δo 来将其变形到查询空间（绿色），在该空间中通过双线性插值计算其特征。

为了解决这一限制，原论文提出了一个基于音频条件和三维点对点的面部扭曲模块 D_η。它为每个投影点 p^{ref} 回归偏移量 $\Delta o = (\Delta u, \Delta v)$，如图1的图像流所示。具体来说，$D_\eta$ 被实现为一个三层 MLP 的变形场，其中 是可学习参数。为了回归偏移量 Δo，需要有效利用查询图像和这些参考图像之间的动态差异。音频信息 A 反映了查询图像的动态，而参考图像的变形可以通过图像特征 $\{Fn\}$ 隐含地看出。因此，原论文将这两个部分与查询三维点坐标 p 一起作为 D_η 的输入。预测偏移量的过程公式为：

$$\Delta o_n = \mathcal{D}_\eta(p, A, F_n(u_n, v_n)). \tag{3}$$

然后将预测的偏移量 Δo_n 加到 p_n^{ref} 上，如图 3 所示，得到查询点 p 的确切对应坐标 $p_n^{ref'}$：

$$p_n^{ref'} = p_n^{ref} + \Delta o_n = (u_{\ n}, v_{\ n}), \tag{4}$$

其中 $u_{\ n} = u_n + \ u_n$, $v_{\ n} = v_n + \ v_n$。

由于硬索引操作 Fn(un , vn) 不可微分，梯度无法反向传播到这个扭曲模块。因此，原论文引入了一个软索引函数实现可微扭曲，其中每个像素的特征通过其周围点的特征双线性插值获得。通过这种方式，变形场 D_η 和面部辐射场 F_θ 可以端到端地联合优化。图2显示了这种软索引操作的可视化。对于绿色点，其像素特征通过其四个最近邻点的特征双线性插值得到。为了更好地约束这个扭曲模块的训练过程，原论文引入了一个正则项 L_r，以限制预测偏移量的值在合理范围内，防止失真，

$$L_r = \frac{1}{N \cdot |\mathcal{P}|} \sum_{p \in \mathcal{P}} \sum_{n=1}^{N} \sqrt{\Delta u_n^2 + \Delta v_n^2}, \tag{5}$$

其中 P 是体素空间中所有三维点的集合，N 是参考图像的数量。此外，原论文认为低密度的点更可能是背景区域，应该具有较低的变形偏移量，在这些区域，应该施加更强的正则化约束。为了更合理地约束，原论文将上述 L_r 更改为：

$$L_r' = (1-\sigma) \cdot L_r, \tag{6}$$

其中 σ 表示这些点的密度。最终的动态面部辐射场公式为：

$$(c, \sigma) = \mathcal{F}_\theta(p, d, A, \tilde{F}'), \tag{7}$$

(c,) = F (p, d,A, F')

其中 $\tilde{F}' = Aggregation(Fn(u_n, v_n))$。

通过这个面部扭曲模块，所有参考图像都可以变形到查询空间，从而更好地建模谈话面部变形。后续的消融研究证明了这个组件在生成更准确且音频同步的嘴部运动方面的有效性。

3.4 体积渲染

体积渲染用于将公式7中的颜色 c 和密度 σ 集成到面部图像中。原论文将背景、躯干和脖子部分一起视为渲染“背景”，并从原始视频中逐帧恢复。原论文设置每条射线最后一点的颜色为对应背景像素，以渲染自然的背景，包括躯干部分。这里原论文遵循原始 NeRF 的设置，摄像机射线 r 的累计颜色 C 在音频信号 A 和图像特征 $\tilde{F}'$ 的条件下公式为：

$$C\left(r;\theta,\eta,R,T,A,\tilde{F}'\right) = \int_{z_{near}}^{z_{far}} \sigma(t)\cdot c(t)\cdot T(t)\,dt, \tag{8}$$

其中 θ 和 η 是面部辐射场 F 和面部扭曲模块 D_η 的可学习参数。R 是旋转矩阵，T 是平移向量。$T(t) = \exp\left(-\int_{z_{\text{near}}}^{t} \sigma\left(r(s)\right) ds\right)$ 是沿摄像机射线的积分透过率，z_{near} 和 z_{far} 是摄像机射线的近远界限。原论文遵循 NeRF 设计了一个 MSE 损失 $L_{MSE} = ||C-I||^2$，其中 I 是真实颜色。结合公式6中的正则项，整体损失函数公式为：

$$L = L + \lambda \cdot L'_r. \tag{9}$$

3.5 实现细节

原论文在不同身份上从粗到细训练一个基础辐射场。在粗略训练阶段，通过公式2在 LMSE 监督下训练面部辐射场 F_θ，捕捉头部结构并建立从音频到唇形运动的通用映射。然后原论文将面部扭曲模块加入训练，通过公式7联合优化偏移回归网络 D_η 和 F_θ，使用公式9中的损失函数。

对于任意未见过的身份，只需基于训练好的基础模型短时间的训练剪辑进行微调。通过短时间微调可以学习个性化的嘴部发音模式，使得渲染的图像质量大幅提高。这样，这个微调模型就可以用于推理 (inference) 了。

4 实验结果

生成伪造人脸的过程包括预处理数据，训练面部形变模块，渲染视频三个步骤。我们使用信院显卡集群的单卡 3090 进行训练，训练视频均为 512×512 分辨率，25fps。具体的结果详见压缩包内的视频。

4.1 原视频相同驱动音频

4.1.1 生成效果

使用与原视频相同的音频作为驱动，生成效果如附件 1.mp4、2.mp4 所示，其中视频左边为 ground truth，右边为用 15s 视频训练 50k 个 epoch 的生成结果。

4.1.2 定量分析

将 dfrf 生成的视频 1、视频 2 分别放入 sync net[CZ17] 检测唇音同步，并用 ad-nerf 用 15s 视频训练同样迭代次数生成的视频作为参考：

	ad-nerf	视频 1	视频 2
offset（偏移量）	-14	-2	2
confidence（置信度）↑	0.654	4.676	2.832

表 1: 可见在唇音同步性上 dfrf 明显优于 ad nerf

我们还对用 15s 的视频分别训练 20k，30k，40k，50k 个 epoch 的结果评估其 psnr（峰值信噪比），ssim（结构相似度）和 lpips（感知损失），并以 15s 视频训练 40k 次迭代的 ad nerf 的生成结果作为参考：

	ad-nerf	20k	30k	40k	50k
PSNR↑	29.45	32.02	31.98	31.85	31.88
SSIM↑	0.936	0.980	0.980	0.979	0.979
LPIPS↓	0.039	0.045	0.042	0.040	0.039

表 2: 可见 dfrf 随着迭代次数增加，在 psnr 和 ssim 基本保持不变的情况下，lpips 指标逐渐变好并超越了 ad nerf，可以认为在图像质量方面 dfrf 能达到甚至超越 ad nerf

根据上述定量分析实验结果，可见 dfrf 生成的视频在唇音同步和图像质量方面均取得不错的效果。

4.2 原视频不同驱动音频

由于 dfrf 是一个音频驱动的生成模型，因此可以用不同的音频来驱动同一个人脸的视频生成。

4.2.1 生成效果

用不同语言的音频分别驱动同一个视频，生成效果如附件 french.mp4，english.mp4，chinese.mp4 所示。

4.2.2 定量分析

将上面三个视频分别放入 sync net 检测唇音同步结果如下：

	french	english	chinese
offset（偏移量）	-3	-1	-1
confidence（置信度）↑	2.232	3.145	2.564

表 3: 可见使用不同音频驱动，dfrf 生成的视频仍能保持唇音同步

参考文献

[CZ17] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer, 2017.

[SLZ+22] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. 2022.