arXiv:2305.03027v1 [cs.CV] 4 May 2023

NeRSemble：人头多视角辐射场重建

TOBIAS KIRSCHSTEIN，慕尼黑工业大学，德国；SHENHAN QIAN，慕尼黑工业大学，德国；SIMON GIEBENHAIN，慕尼黑工业大学，德国；TIM WALTER，慕尼黑工业大学，德国；MATTHIAS NIESSNER，慕尼黑工业大学，德国

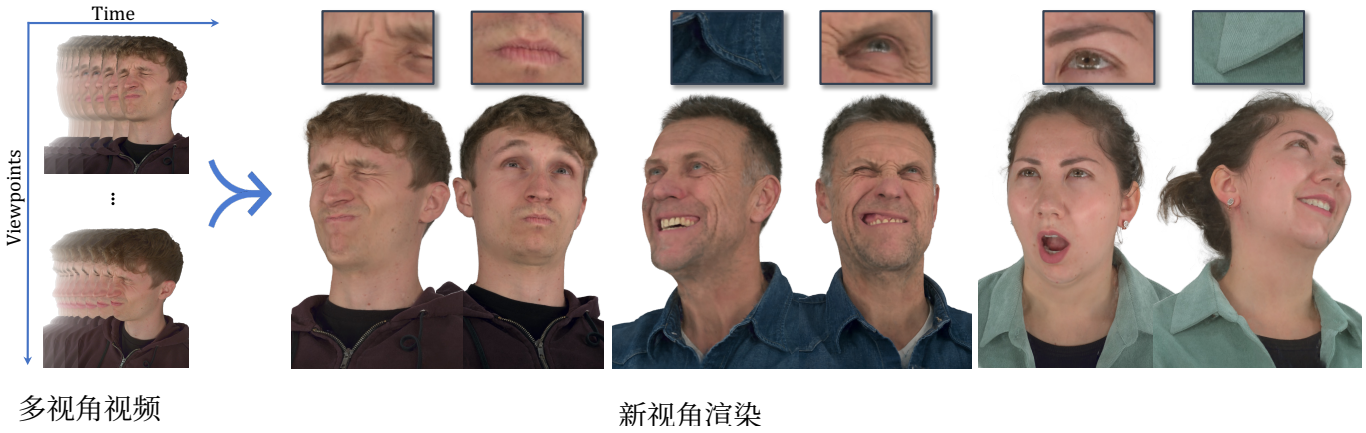


图 1. **NeRSemble：** 给定来自十二台摄像机的多视角视频录制数据（左），我们的方法能够合成处于复杂运动中的人头的高度逼真的新视图。我们从未见视角渲染出的结果（右）忠实呈现了静态场景部分以及发生高度非刚性变形的区域。除方法之外，我们还公开了高质量多视角视频采集数据，其中包含来自总计 222 名受试者的 3170 万帧。

我们专注于重建高保真人头辐射场，捕捉其随时间变化的动画，并合成任意时间步的新视点重新渲染结果。为此，我们提出了一种新型多视角采集装置，由 16 台经标定的机器视觉摄像机组成，可按每秒 73 帧的速度记录分辨率为 7.1 MP 的时间同步图像。借助该装置，我们采集了一个新数据集，其中包含 220 多个人头的 4700 多段高分辨率、高帧率序列，并据此推出一项新人头重建基准[1]。录制的序列涵盖广泛的面部动态，包括头部运动、自然表情、情绪和口语。为重建高保真人头，我们提出了基于哈希集成的动态神经辐射场（NeRSemble）。我们通过结合变形场与三维多分辨率哈希编码集成来表示场景动态。变形场可精确建模简单的场景运动，而哈希编码集成则有助于表示复杂动态。因此，我们获得了能够捕捉人头随时间运动的辐射场表示，并支持对任意新视点进行重新渲染。在一系列实验中，我们探究了本方法的设计选择，并证明我们的方法显著优于最先进的动态辐射场方法。

CCS 概念：• **计算方法** → 渲染；三维成像；体积模型；重建。

其他关键词和短语：神经辐射场，动态场景表示，新视角合成，多视角视频数据集，人头

1 引言

近年来，无论在社会还是产业领域，依赖于从捕获的场景表示中进行图像照片级真实感渲染的数字应用，其重要性都大幅提升。尤其是，动态人脸和头部的新视图合成，已成为众多图形应用关注的焦点，应用范围涵盖电子游戏、电影制作，以及虚拟现实或增强现实等场景。这里的关键任务是：给定一段记录人类演员展示面部表情或说话的视频，重建具有时间一致性的三维表示。该表示应能从任意视点和时间步，对人脸进行照片级真实感重新渲染。

然而，对于动态物体而言，重建能够实现逼真新视角渲染的三维表示尤其具有挑战性。在这里，我们不仅要重建人物的静态外观，还必须同时捕捉随时间变化的运动，并将其编码为紧凑的场景表示。对于人脸，这项任务更具挑战性，因为下游应用需要精细且高保真的细节，而对视觉伪影的容忍度较低。

[1]我们将发布所有采集的数据，包括全部 4734 段录制内容和基线代码，并提供一个新的公开基准，以支持该领域的后续研究。网站：https://tobias-kirschstein.github.io/nersemble

作者地址：Tobias Kirschstein，慕尼黑工业大学，德国，tobias.kirschstein@tum.de；Shenhan Qian，慕尼黑工业大学，德国，shenhan.qian@tum.de；Simon Giebenhain，慕尼黑工业大学，德国，simon.giebenhain@tum.de；Tim Walter，慕尼黑工业大学，德国，tim.michelbach@hotmail.com；Matthias Nießner，慕尼黑工业大学，德国，niessner@tum.de。

通常非常低。具体而言，人头具有多种特性，使得新视角合成（新视角合成）极具挑战性，例如头发的复杂性、反射特性的差异，以及人体皮肤的弹性会导致严重的非刚性变形和细微皱纹。

在静态场景中，我们已经看到神经辐射场表示（NeRF）[Mildenhall 等人 2020] 能够取得令人瞩目的新视角合成效果。这项开创性工作的核心思想是，将体渲染公式用作重建损失，并将由此得到的辐射场编码为基于神经场的表示。近年来，研究者对扩展 NeRF 以表示动态场景产生了浓厚兴趣。一些方法依赖变形场来建模动态变化的场景内容[Park 等人，2021a、2021b], ，另一些方法则提出用时间条件潜码取代变形场 [Li 等人，2022b]。这些方法在运动有限的短序列上取得了令人信服的结果；然而，要忠实重建具有复杂运动的人头，仍然颇具挑战。

在这项工作中，我们聚焦于在新设计的多视角采集装置背景下解决这些挑战，并提出 NeRSemble——一种结合变形场与灵活潜变量条件化优势的新方法，用于表示动态人头的外观。我们方法的核心思想是将潜特征存储在一组多分辨率哈希网格中，类似于 Instant NGP [Müller 等人 2022], ，并通过对这些网格进行混合来描述给定的时间步。重要的是，我们会在从哈希网格查询特征之前使用变形场。因此，变形场表示场景中的所有粗略动态，并对齐哈希网格的坐标系；哈希网格则负责建模精细细节和复杂运动。为了训练和评估我们的方法，我们设计了一套新的多视角采集装置，使用 16 台机器视觉相机，以 73 帧每秒的速度录制 7.1 MP 视频。借助这套装置，我们采集了一个新数据集，其中包含 222 个人头的 4734 个序列，总计 3170 万个独立帧。我们在这个新推出的数据集上评估了我们的方法，并证明其显著优于现有的动态神经辐射场重建方法。就分辨率和帧每秒数量而言，我们的数据集大幅超越所有可比数据集，并将向公众开放。此外，我们还将推出一个人头动态新视角合成的公开基准，这将有助于推动该领域的发展，并提高不同方法之间的可比性。

总而言之，我们的贡献如下：

- 一种基于神经辐射场表示的动态头部重建方法，将变形场与多分辨率哈希编码集合相结合。该方法能够利用稀疏相机阵列实现高保真的新视角合成，并支持对具有复杂运动的场景进行细致表示。
- 一个高帧率、高分辨率的人头多视角视频数据集，包含 220 多名受试者的 4700 多个序列。该数据集将公开发布，并包含一项人头动态新视角合成的新基准。

表 1. 现有的人脸多视角视频数据集。请注意，对于每个数据集，我们仅统计公开可访问的录制内容。

数据集	#受试者	#相机	分辨率	Fps
D3D面部动作编码系统 [2011]	10	6	1280 x 1024	60
BP4D-Spontaneous [2014]	41	3	1392 x 1040	25
数字间光场 [2017]	5	16	2048 x 1088	30
4DFAB [2018]	180	7	1600 x 1200	60
VOCASET [2019]	12	12	1600 x 1200	60
MEAD [2020]	48	7	1920 x 1080	30
MultiFace [2022]	13	150	2048 x 1334	30
Ours	222	16	3208 x 2200	73

2 相关工作

人脸建模与渲染是计算机图形学的核心课题，在许多应用中都发挥着关键作用，例如电子游戏、社交媒体、电信和虚拟现实。

2.1 三维可变形模型

三维可变形模型（3DMMs）是过去二十年来的主流方法。统一网格拓扑使得人们能够使用简单的统计工具表示身份特征和表情 [Blanz 和 Vetter 1999年；Li 等人 2017]。结合纹理后，便能生成令人信服的渲染结果 [Blanz 和 Vetter 1999年；Paysan 等人 2009], ，但基于网格的三维可变形模型在头发或精细的身份特定细节建模方面存在固有限制。近年来，神经场的应用 [Xie 等人 2022] 缓解了对拓扑一致网格的依赖。这些模型能够对完整的人头进行建模，包括头发[Yenamandra 等人 2021]和精细细节 [Giebenhain 等人 2022]。另一类研究中，Zheng 等人 [2022] 结合神经场与经典三维可变形模型的思路，对单目视频进行拟合。

2.2 神经辐射场

我们的工作旨在实现高度逼真的视频渲染结果，包括细致的发型和复杂的变形。因此，我们不采用三维可变形模型中的常见假设，而是专注于将单个多视角视频序列拟合到尽可能高的细节程度。神经辐射场（NeRF）[Mildenhall 等人 2020] 近期已成为新视角合成领域的最先进方法。早期的神经辐射场通常需要在单个场景上训练数小时甚至数天，而近期的研究进展已将训练时间缩短至数分钟。例如，可以通过基于网格的优化 [Fridovich-Keil 和 Yu 等人，2022年；Karnewar 等人，2022年；Sun 等人 2022], 张量分解 [Chen 等人 2022], 或 Instant NGP 的 [Müller 等人 2022] 多分辨率体素哈希来实现。

2.3 动态神经辐射场

将神经辐射场扩展到时变非刚性内容是另一个核心研究主题，近年来进展迅速。Pumarola 等人 [2020] 和 Park 等人 [2021a；2021b] 在规范空间中对单个神经辐射场建模，并显式建模从观测帧到规范空间的反向形变，以解释场景中的非刚性内容。旧：关于 the On the



图 2。左：我们定制搭建的多视角视频采集装置。右：16 个视点以及从录制视频中获取的面部细节 s.

另一方面，Li 等人 [2022b] 避免使用显式变形，而是将场景状态编码到一个潜向量中，并直接以此作为神经辐射场的条件。Wang 等人 [2022b]利用基于傅里叶变换的压缩来表示网格特征，从而表示 4D 辐射场。Lombardi 等人 [2019] 则将图像到体积生成器与变形场结合使用。

与我们的工作同期，Song 等人 [2023] 将快速神经辐射场骨干网络（即 TensoRF 或 Instant NGP）与滑动窗口方法相结合，以处理时间变化。Attal 等人 [2023] 将 4D 张量分解与学习式采样方法相结合，实现快速动态新视角合成。与这些工作不同，我们提出将基于哈希的分解与变形场结合。

2.4 视频视图合成

除了神经辐射场之外，也有其他不依赖辐射场骨干网络的视频视图合成方法。在早期的一项工作中，Zitnick 等人 [2004]利用几何辅助图像渲染来渲染动态场景的新视角。最近，Broxton 等人 [2020]通过构建多球面图像来生成自由视点视频，再将其转换为分层网格表示，以实现快速渲染和流媒体传输。Collet 等人 [2015], 则采用另一种方法，利用多视图立体系统获取动态表演的跟踪网格。尽管这些基于网格的方法能为较大场景带来出色的视频视图合成效果，神经辐射场的优势在于能够对头发等精细复杂的细节进行照片级真实感重建。

2.5 面部神经辐射场

有多项研究提出了专门面向人头领域的方法。值得一提的是，Gafni 等人 [2021] 使用拟合得到的 3DMM 参数来为神经辐射场提供条件输入，而 Athar 等人 [2022] 则将这种方法扩展到对基于 3DMM 几何结构推导出的显式形变进行建模。最近，Zielonka 等人 [2023] 提出了一种类似的方法，着重提升重建速度和实时渲染性能，方法是利用结合 Instant NGP 的跟踪 3DMM。Wang 等人 [2022a] 提出了一种可控制身份参数的生成式 NeRF。Hong 等人 [2022] 则采用类似方法，并加入了额外的表情参数。Lombardi 等人 [2021] 提出了一种高度优化的神经渲染方法，将颜色发射值显式存储在与 3DMM 表面松散绑定的体素网格中。在这项工作中，我们提出了一种无模板方法，因为我们认为，使用 FLAME [Li 等人 2017]这样的粗略几何代理，很难实现像素级精准新视图合成。

与我们的方法类似，Gao 等人 [2022] 最近提出融合来自多个哈希网格的特征。虽然他们的方法使用跟踪的 3DMM 参数，但 NeRSemble 会联合优化混合权重和其余模型参数。此外，我们还展示了在融合哈希网格之前加入变形场可带来显著提升。

3 人类 F 的多视角视频数据集 ACES

我们介绍了一个新颖的数据集，其中包含使用 16 台机器视觉相机拍摄的 222 名受试者的 4734 段多视角视频录制。我们的前向拍摄设备覆盖从左到右 93°、从上到下 32° 的视场。由于人类面部运动十分复杂，感知情绪也会受到细微差异的显著影响，我们采用 710 万像素的高分辨率，能够捕捉整个面部，细致到单根发丝和皱纹的程度，如图 2 所示。我们还以每秒 73 帧的速度录制，以确保不会遗漏任何细微动作。综合来看，我们的数据集独特地融合了高分辨率、高帧率录像和众多受试者，目前其他数据集均无法与之媲美（见表 1）。

3.1 采集

每次录制会话包含 25 个短序列，因此每位参与者约有 3 分钟的多视角视频素材。我们要求参与者做出多种不同的面部表情，以

表 2. **我们的多视角视频人脸数据集的统计信息。**

#参与者	#序列	#帧	总时长	磁盘空间
222（157名男性 / 65名女性）	4734	3170万	7 小时 30 分钟	203 TB

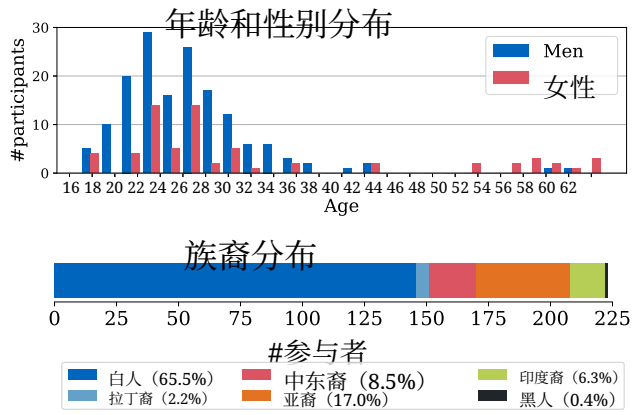


图 3。**我们数据集中的参与者统计信息**。我们记录的序列涵盖了不同性别参与者的广泛年龄和族裔。

最大限度地丰富运动类型。具体而言，我们的采集脚本包含 9 个涵盖不同面部肌肉群的表情序列、1 个包含快速运动的头发序列、4 个情绪序列、10 段带音频的句子，以及 1 个较长序列，在此序列中，受试者可以自由做出任意的面部变形和头部运动。

为获得高质量的视频录制效果，我们采用 3ms 的快门速度，从而能够捕捉快速运动，同时避免运动模糊。此外，我们使用较小的镜头光圈，以确保面部区域的各处图像均清晰锐利。这种组合能够获得高质量的采集效果，但会减少摄像机传感器接收到的入射光量，因此需要使用 8 块高亮度 LED 灯板对场景进行照明。我们还在灯具上使用漫射板，以减少皮肤上的镜面反射。此外，我们的摄像机采用精确时间协议（PTP），以实现精确的时间同步。同步时钟具有亚微秒级精度，因此各视频帧实际上是同时采集的。最后，我们使用色卡校准每台摄像机的白平衡系数和伽马参数。由此获得的视频录制结果在不同视点之间具有一致的色彩，并能够捕捉精细细节，如图 2 所示。

3.2 数据处理

我们通过使用精细棋盘格并结合束调整优化流程，估计每台相机各自的外参矩阵和共享的内参矩阵。这使得估计出的相机姿态十分准确；我们在合成环境中验证，其误差处于亚毫米级。此外，我们的录制背景是一面白墙，并在录制前进行拍摄。之后，可通过图像抠像方法从这些空背景中为每一帧生成高质量的前景分割图，例如使用 BackgroundMatting v2 [Lin 等人 2021]。



图 4. **我们数据集的结构**。我们要求每位参与者执行相同的表情序列。

3.3 基准测试

我们的数据集使我们能够研究从多视角视频中进行逼真人头重建，这正是本工作的目标。此外，采集的数据还支持远超新视角合成的多种用途，例如人头泛化、沉浸式视频会议、适用于 VR 的虚拟形象渲染、微表情研究、重新演绎、动画制作等。因此，我们计划向学术界发布完整的数据集。此外，我们将从数据集中选取具有代表性的录制内容，编制一个针对人脸新视角合成的基准。我们希望这一举措能够促进不同方法之间的可比性，并最终推动高保真人头重建研究的发展。

3.4 数据隐私

鉴于采集数据的敏感性，我们数据集中的所有参与者均签署了符合 GDPR 要求的协议。请注意，遵守 GDPR 意味着每位参与者都有权要求及时删除其数据。我们将在分发数据集时保障这些权利。

4 使用哈希集成的动态神经辐射场

我们的目标是找到一种时空表示，使经历复杂非刚性变形的人头能够实现高度逼真的新视角合成。为此，我们提出将变形场与 4D 场景体积分解相结合，沿时间维度将其分解为一组 3D 特征体积，以重建场景的动态（见图 5）。

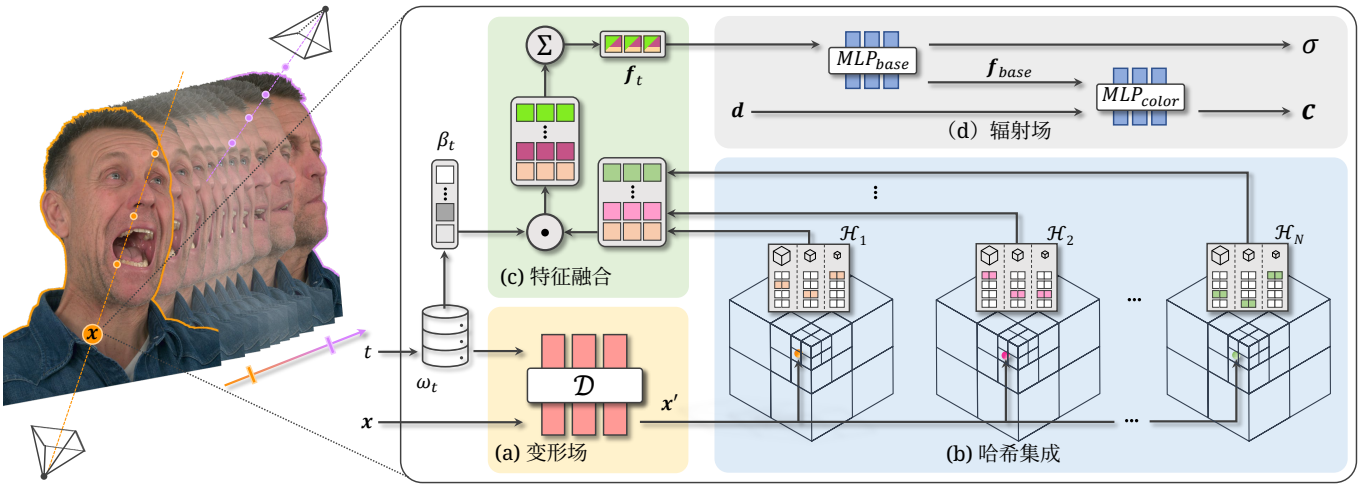


图 5. **方法概述**。 NeRSemble 使用体渲染表示动态新视角合成的时空辐射场（左侧）。右侧展示 NeRSemble 如何获取射线在时间 t 上一点 x 的密度 σ(x) 和颜色值 c(x，d)。(a) 给定变形代码 ω_t，点 x 会被映射至规范空间中的 $x' = D(x\ \omega_t)$。(b) 使用所得点从我们集成中的第 i 个哈希网格查询特征 $H_i(x')$。(c) 使用权重 β_t 对所得特征进行混合。请注意， ω_t和 β_t都对解释时间变化有所贡献。(d) 我们使用由两个小型多层感知机组成的高效渲染头，根据混合后的特征预测密度 σ(x) 和视角相关颜色 c(x，d)。

4.1 预备知识：神经辐射场

我们的工作建立在近期神经辐射场（神经辐射场） [Mildenhall 等人 2020], 的成功基础之上，后者利用密度场 σ(x)和视角相关颜色场 c(x，d)进行体渲染。给定一条射线 $r(\tau) = o + \tau d$，以及一个颜色值

$$C(r) = \int_{\tau_n}^{\tau_f} \underbrace{T(\tau)\sigma(r(\tau))}_{w(\tau)} c(r(\tau), d)\, d\tau, \tag{1}$$

通过沿射线从近裁剪面 τ_n 到远裁剪面 τ_f进行积分获得，其中 $T(\tau) = \exp\left(-\int_{\tau_n}^{\tau} \sigma(r(s))\, ds\right)$表示截至 τ的累积透射率。

优化的目标是求解用于编码所得辐射场的多层感知机（MLP）的最优参数。近期研究表明，纯体素网格 [Fridovich-Keil 和 Yu 等人 2022]，以及显式网格与多层感知机的组合 [Müller 等人 2022]，都是辐射场表示的有效替代方案。

Instant NGP。 我们的方法依赖于 Instant NGP [Müller 等人 2022],提出的体素哈希方案，该方案结合多分辨率特征 f(x) 与两个小型多层感知机来表示神经辐射场的三维场：

$$[\sigma(x), f_{base}(x)] = MLP_{base}(f(x)) \tag{2}$$

$$c(x, d) = MLP_{color}(f_{base}(x), d). \tag{3}$$

值得注意的是，特征存储在多分辨率哈希网格 H 中，使得 $f(x) = H(x)$。哈希网格 H 提供了一种节省内存的方式，可将 3D 场景体积编码到足以让小型 MLP 表示最复杂场景的程度。

4.2 多分辨率哈希集成

我们对动态场景的表示受到经典混合形状 [Blanz 等人 2003] 的启发。我们假设，在时间 t，场景的任意状态都可以表示为从一组多分辨率哈希网格 $\{H_i\}_{i=1}^N$中选取的特征向量的组合，我们将其称为哈希网格集成。为了获得时间 t的混合辐射场，我们将其表示为其特征的线性组合

$$f_t(x) = \sum_{i=1}^{N} \beta_{t,i} \mathcal{H}_i(x), \tag{4}$$

使用混合权重 β，这些权重与哈希集成、共享的 MLPbase 和 MLPcolor 一同进行优化。

由于混合是在特征空间中进行的，这种混合操作使模型能够表示复杂的运动。随后，混合后的特征由 MLPbase 和 MLPcolor 进行解码。

4.3 特征的空间对齐

如果集成中的所有独立元素都在共享坐标系中运行，则网格特征的混合效果最佳。例如，传统的混合形状依赖于由网格拓扑的顶点排序所确定的完美对应关系。由于我们在没有这种结构的情况下混合特征，因此按照 Park 等人 [2021a] 的方法，使用一个 $SE(3)$场对变形进行显式建模，该变形场由基于坐标的 MLP 表示。更具体地说，我们的变形场

$$\mathcal{D} : \mathbb{R}^3 \times \mathbb{R}^{d_{def}} \to \mathbb{R}^3, (x, \omega_t) \mapsto x' \tag{5}$$

给定描述当前表情的条件编码 ω_t，将观测空间中的点 x 映射到规范空间中的对应点 x'。随后，变形场会在不同时间步之间寻找对应点，并将它们映射到同一个规范空间。

利用这些学习到的对应关系，我们修改公式 4，使其在规范空间中运行：

$$\mathrm{f}^{(t)} = \sum_{i=1}^{N} \beta_{t,i} \mathcal{H}_i \left(\mathcal{D} \left(\mathrm{x}, \boldsymbol{\omega}_t \right) \right). \tag{6}$$

这样一来，融合在两个不同时间步观察到的同一运动点的特征会更加容易。

4.4 预热阶段

通过这种组合，哈希集成和变形相互竞争，以解释人脸动态。因此，优化过程很可能陷入局部极小值，此时 D 无法提供有意义的变形。因此，我们在优化过程中提出了一个预热阶段，以鼓励 D 学习观测空间与规范空间之间有意义的对应关系。

在优化的前 E个训练轮次中，我们禁用除一个哈希网格之外的所有哈希网格，使模型本质上模拟可变形 NeRF。在此阶段，变形场 D 及其变形编码 $\boldsymbol{\omega}_t$ 是解释动态行为的唯一手段。因此，D 能够不受干扰地学习有意义的变形，这对于之后有效融合哈希表特征至关重要。

预热之后，第一个哈希表与我们的变形场便能够解释场景的低频动态。在接下来的 Etrans 个训练轮次中，我们继续将其余所有哈希表加入优化。依次插入这些表使我们能够表示精细尺度的运动和细节，而这些原本无法由 D 表示。

为了确保平滑过渡，我们调整混合权重

$$\beta_{t,i}(s)^* = \alpha_i(s) \beta_{t,i} \quad (\forall i \in \{1, ..., N\}), \tag{7}$$

其中，i表示哈希集成的索引，$\alpha_i(s)$ 是 Park 等人引入的窗函数。[2021a] 和 s 计划在 Einit 至 Etrans+Einit 训练轮次之间从 1 线性增加到 N。关键的是，$\alpha_1(s)$ 在整个优化过程中保持不变，从而确保第一个哈希表始终处于激活状态。

4.5 深度监督

由于我们的多视图数据集能够通过传统方法计算深度图，因此本文也研究了额外深度监督的作用。给定射线的深度 $z^{\mathrm{gt}}(\boldsymbol{r})$，我们将深度损失计算为

$$\mathcal{L}_{\mathrm{depth}} = \mathbb{E}_{\boldsymbol{r} \sim \mathcal{R}_{\mathrm{d}}} \left[\left(z(\boldsymbol{r}) - z^{\mathrm{gt}}(\boldsymbol{r}) \right)^2 \right], \tag{8}$$

其中，射线 $\boldsymbol{r}$的预期深度为 $z(\boldsymbol{r}) = \int_{\tau_n}^{\tau_f} w(\tau) \cdot \tau \, \mathrm{d}\tau$。

由于实际中的深度观测并不完整，因此只对已知深度的射线 $\boldsymbol{r} \in \mathrm{R}_{\mathrm{d}}$ 计算深度损失。

此外，我们采用 Urban Radiance Fields (URF) [Rematas 等人 2022] 提出的两种视线先验，以进一步利用深度约束。首先，我们使用

$$\mathcal{L}_{\mathrm{empty}} = \mathbb{E}_{\boldsymbol{r} \sim \mathcal{R}_{\mathrm{d}}} \left[\int_{\tau_n}^{z(\boldsymbol{r})-\epsilon} w(\tau)^2 d\tau \right] \tag{9}$$

在表面前方雕刻出空空间，其中 ϵ 在训练过程中会像 URF 中那样呈指数衰减。其次，

$$\mathcal{L}_{\mathrm{near}} = \mathbb{E}_{\boldsymbol{r} \sim \mathcal{R}_{\mathrm{d}}} \left[\int_{z(\boldsymbol{r})-\epsilon}^{z(\boldsymbol{r})+\epsilon} \left(w(\tau) - \mathcal{N}\left(\tau \mid z(\boldsymbol{r}), \left(\frac{\epsilon}{3} \right)^2 \right) \right)^2 d\tau \right] \tag{10}$$

促使深度观测附近的体密度遵循逐渐收窄的高斯分布。结合 $\mathrm{L}^{\mathrm{depth}}$，这三种先验构成了深度监督，旨在提高重建的几何保真度。

4.6 背景去除

我们采用连续值阿尔法图 $M(\boldsymbol{r})$，以阻止模型重建背景部分。我们使用一种施加稀疏性约束的 L1 损失，对穿过背景像素的射线上的密度进行惩罚：

$$\mathcal{L}_{\mathrm{mask}} = \mathbb{E}_{\boldsymbol{r} \sim \mathcal{R}_{\mathrm{bg}}} \left[\left\| (1 - T(\tau_f)) - M(\boldsymbol{r}) \right\|_1 \right], \tag{11}$$

其中，$T(\tau f)$ 是射线 $\boldsymbol{r}$的总透射率，$M(\boldsymbol{r})$ 是其在预计算阿尔法图中对应的阿尔法值。

4.7 优化目标

最终损失由以下各项组成：

$$\mathcal{L} = \mathcal{L}_{\mathrm{rgb}} + \mathcal{L}_{\mathrm{mask}} + \underbrace{\mathcal{L}_{\mathrm{depth}} + \mathcal{L}_{\mathrm{near}} + \mathcal{L}_{\mathrm{empty}}}_{\text{depth supervision}} + \mathcal{L}_{\mathrm{dist}} \tag{12}$$

其中，$\mathrm{L}^{\mathrm{rgb}}$ 是标准的均方误差颜色损失，我们仅在前景射线上计算该损失。我们还加入了失真损失 $\mathrm{L}_{\mathrm{dist}}$，用于惩罚孤立的低密度区域 [Barron 等人 2022]。由于我们的场景仅包含人头，且人头大致呈凸形，因此我们还会在指向中心的随机射线上计算 $\mathrm{L}_{\mathrm{dist}}$。这样可将该项的正则化效应扩展到头部后方空间，该区域在我们的场景中通常会被遮挡。

最后，我们为每个损失项设置相应的权重：
λ_{depth}，λ_{dist}，λ_{near}，$\lambda_{\mathrm{empty}} = 1\mathrm{e}{-4}$ 和 $\lambda_{\mathrm{mask}} = 1\mathrm{e}{-2}$。

4.8 动态场景表示讨论

与张量分解的关系。公式 4 可解释为四维张量分解的一种特殊情况，类似于 Chen 等人 2022] 在 TensoRF [中提出的向量-矩阵分解。表示动态场景的时空张量 $\mathrm{T} \in \mathbb{R}^{D_T \times D_X \times D_Y \times D_Z}$ 可以分解为四个向量-张量外积之和：

$$\mathcal{T} \approx \sum_{a \in A} \sum_{i=1}^{N} v_i^a \circ M_i^{A \setminus \{a\}}, \tag{13}$$

其中 $\circ$ 表示向量-张量外积，$A = \{X、Y、Z、T\}$ 是轴索引集合，$v_i^a \in \mathbb{R}^{D_a}$ 是一个向量，而 $M^{A \setminus \{a\} i}$ 是一个三维张量，例如 $M^{A \setminus \{X\}_i} \in \mathbb{R}^{D_T \times D_Y \times D_Z}$。

我们方法的公式 4 可以看作公式 13 的一个特例，其中仅使用了 $a = T$ 对应的项。我们没有将特征存储在稠密网格 $M^{A \setminus \{a\} i}$ 中，而是采用一种内存高效的哈希表表示 $M^{A \setminus \{T\} i} = \mathrm{H}_i$，向量 v_i^a则对应我们的混合权重$v_{i,t}^a = \beta_{t,i}$。

我们最终的方法在查询 H_i 之前引入了一个变形场 D（见公式 6），因此不同于这种张量分解视角。通过利用变形场解释部分运动，该方法有效地对齐了哈希表中跨时间步的空间特征。

Attal 等人 [2023]、Cao 和 Johnson [2023]、Fridovich-Keil 等人 [2023], 在同期研究中提出了另一种实现四维张量分解的方法：他们结合 6 个可能的二维特征平面中的特征，而不是像公式 13 那样对一维和三维张量进行 4 次外积。

与 *HyperNeRF* 的关系。 HyperNeRF [Park 等人 2021b] 在其规范空间 NeRF 中加入所谓的环境维度，以解决变形场无法建模的拓扑问题。我们的方法并未加入连续的环境维度，而是使用多个哈希网格对规范空间进行建模，实质上引入了一种具有类似作用的离散 环境维度。

5 实验结果

我们在新视角合成（NVS）任务上评估了本方法，任务是基于数据集中 10 个不同序列的多视角视频录制进行新视角合成，这些序列展现了不同类型的面部和头部运动。验证序列包含头部旋转、大笑、眨眼、说话、头发摆动、各种嘴部运动，以及一个自由表情序列。所有视频均以 73 fps 拍摄，每段包含 300–500 帧。我们从 16 个可用视点中选取 12 个作为输入，并在剩余 4 个视点上评估新视角合成任务。所选评估视角在相机布局中均匀分布；由于存在极端观察角度，评估方案颇具挑战性（见图 7）。

5.1 数据准备

在开展实验之前，我们利用数据集受控这一特点，便于从图像输入中重建动态三维场景。具体而言，我们执行以下预处理步骤：

深度图生成。 我们采用标准的 COLMAP 流程，为 12 个训练视角中的每个视角生成深度图 [Schönberger 和 Frahm，2016年；Schönberger 等人 2016]。为去除噪声深度测量值，我们舍弃被少于 3 台摄像机观测到的深度值。

背景抠像。 我们使用 Background Matting v2 [Lin 等人 2021]，根据拍摄帧及其对应的背景图像获取阿尔法图。为确保最佳质量，我们使用其 ResNet101 [He 等人 2016] 版本，并在细化阶段将误差阈值设为 0.01。

图像下采样。 在所有实验中，我们将图像下采样为原来的二分之一，得到 1604×1100 像素，这对所有方法来说都足够。在时间维度上，我们不进行下采样，而是以完整的 73 fps 开展所有实验。

颜色校正。尽管我们对摄像机进行了颜色校准，不同视角之间的亮度仍可能存在细微差异。为解决这一问题，我们首先使用面部分割掩码 [Yu 等人 2018] 从面部、躯干和头发区域采样像素值。我们随后通过最优传输 [Flamary 等人 2021]求解仿射颜色变换矩阵，以对齐不同视角间获得的颜色分布。

5.2 浮点伪影去除

基于网格的场景表示通常缺少纯多层感知机架构所具有的平滑先验。因此，它们往往会生成细小的浮点伪影，影响重新渲染结果的视觉质量。由于我们的哈希集成基于 Instant NGP，因此也继承了这一倾向。为解决这一问题，我们为每个序列指定紧密贴合的轴对齐包围盒，并且仅在其中重建辐射场。除了向所有基线方法开放的紧密场景盒拟合外，NeRSemble 还采用了以下两种抑制浮点伪影的技术，并在第 5.7 节中进行了消融研究。

视锥体剔除。我们排除那些被少于 2 个训练相机看到的空间区域，因为这些区域尤其容易产生浮点伪影。这些区域在训练或推理期间都不会被查询。

占据网格过滤。 在推理之前，我们对 Instant NGP 骨干网络在训练过程中跟踪的密度网格应用低通滤波器，并且仅在最大的连通分量内进行渲染，从而有效地丢弃密度中较小的孤立区域。

5.3 实现细节

我们使用 PyTorch [Paszke 等人 2019] 在 Nerfstudio [Tancik* 等人 2022] 框架中实现了我们的方法，该框架采用 NerfAcc [Li 等人, 2022c] 实现的 Instant NGP。

我们使用预热调度器 $E\text{init} = E\text{trans} = 40\text{k}$ 训练所有模型 300k 次迭代，在单张 Nvidia RTX A6000 上约需一天。以 1604×1100 像素的分辨率推理单帧，大约需要 25 秒。

所有模型组件的学习率均设为 $1e^{-3}$，并且每 20k 次迭代衰减为原来的 0.8。对于变形场 D，我们改用 0.5 的衰减因子，以确保预热阶段结束后学习率足够低。

此外，我们使用 $N = 32$ 个哈希表，每个表均采用 Müller 等人 [2022] 提供的默认超参数配置。对于变形场 D，我们采用 [Park 等人, 2021a] 提出的 $SE(3)$ 场的默认配置，并将可学习形变编码 $\boldsymbol{\omega}_t$ 的维度设为 128。我们的混合权重 $\beta \in \mathbb{R}^N$ 为每个哈希表各设置一个权重。

5.4 基线方法

我们将本方法与多种用于动态场景新视角合成的最先进方法进行比较。具体而言，我们与以下方法进行比较：

5.4.1 动态 NeRF。

Nerfies [Park 等人 2021a] 是可变形 NeRF 的代表方法。我们使用与 HyperNeRF 相同的实现，但不包含环境维度。

HyperNeRF [Park 等人 2021b] 通过扩展 Nerfies 来解决拓扑问题。由于其官方实现存在内存问题，我们将其代码移植到 Nerfstudio 框架，并谨慎选择

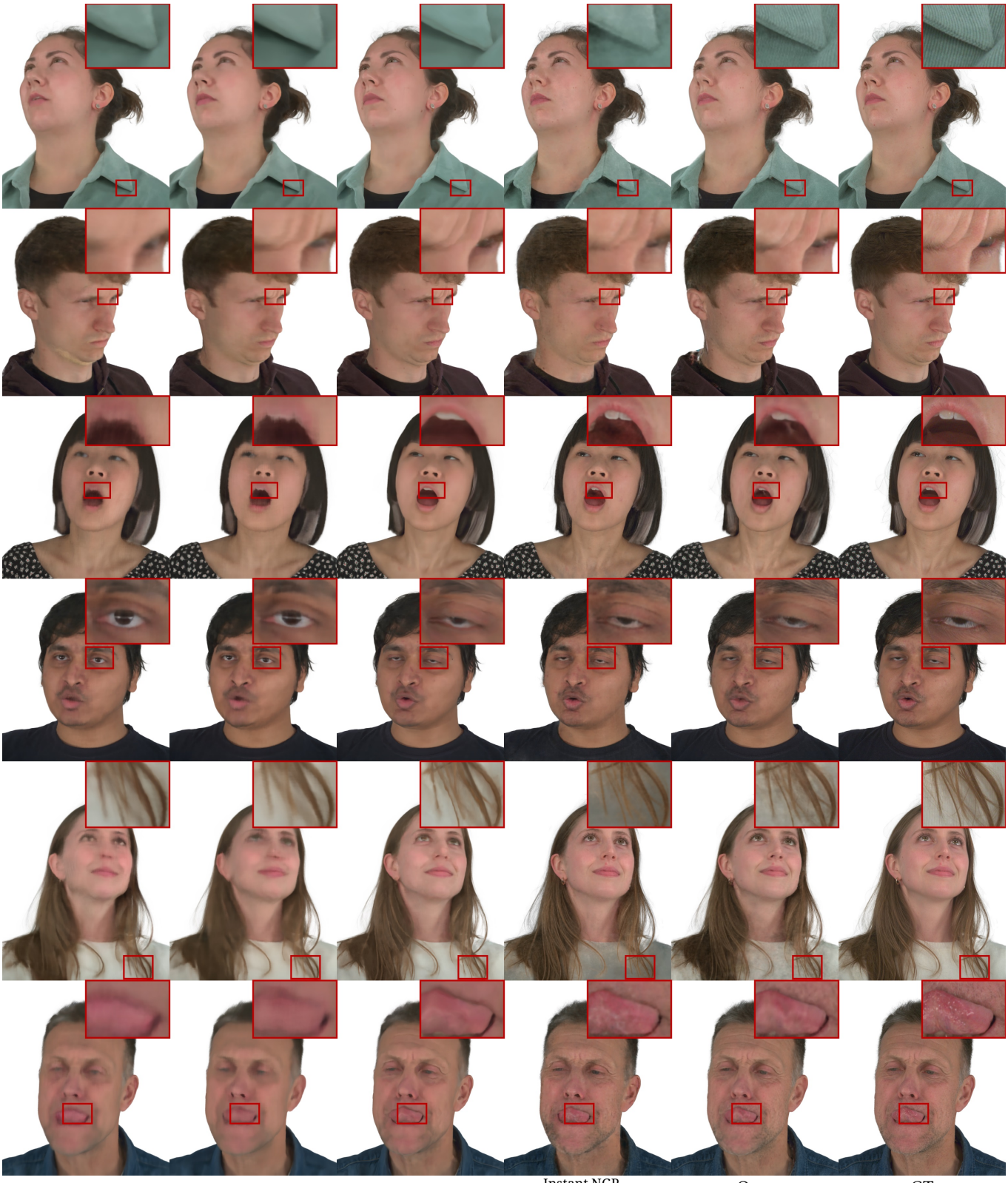


图6. **定性结果**。即使面对复杂表情，我们的方法也能重建出高质量细节。



图 7. **相机布局的空间分布**。红色标记的是用于评估的 4 个视角。点云重建结果通过 COLMAP 获得，并用于额外的深度监督。

超参数，以匹配官方实现的性能。

DyNeRF [Li 等人，2022b], 与此不同，它不要求使用变形场来表示动态内容，而是直接通过依赖时间的潜在编码来调节神经辐射场。由于没有可用的公开代码，我们在 Nerfstudio 中实现了 DyNeRF，并针对我们的数据分布对其进行微调。

5.4.2 与时间无关的方法。此外，我们数据集的多视角特性支持逐帧进行 3D 重建。因此，我们还考虑了两种不考虑时间的基线方法。首先，我们对 COLMAP 点云应用泊松曲面重建（PSR）[Kazhdan 等人 2006]。其次，我们对每一帧分别运行官方的 Instant NGP [Müller 等人 2022]。

5.4.3 面部专用方法。此外，我们还将神经头部化身（NHA）[Grassal 等人 2022] 和 NeRFace [Gafni 等人 2021] 作为面部专用动态重建方法的代表进行比较。这两种方法都依赖由跟踪的统计网格模型提供的几何先验。NHA 是一种基于网格的方法，它在 FLAME 基础上优化顶点偏移，并预测视角和表情相关的纹理。NeRFace 直接利用三维形变模型参数作为神经辐射场的条件，因此与 DyNeRF 类似。由于这两种方法最初都是为单目应用设计的，我们通过采用定制的多视角 FLAME 跟踪器，并在训练期间提供全部 12 个视角，将它们扩展到我们的多视角场景。请注意，依赖三维形变模型使这两种方法都具备一定的重动画能力，但在观测足够密集时，可能会损害渲染质量。

5.4.4 背景建模。为进行公平比较，我们要求所有基于神经辐射场的基线方法在不使用密度的情况下表示背景，即将所有剩余的透射率着色为白色。为此，我们还使用阿尔法掩码将真实图像中的所有背景像素设置为白色。根据我们的经验，这项简单技术能够让所有基线方法都学到前景人物的良好重建结果。

表3。**定量评估**。我们将方法与两种非时序基线方法以及三种动态重建方法进行了比较。我们在数据集中的10个具有多样性的序列上，对未见过的验证视角进行评估。我们的方法在三项指标上均优于基线方法。最后两行展示了针对我们方法的核心架构组件和训练过程所进行的消融实验。

	方法	峰值信噪比 ↑	结构相似性 ↑	学习感知图像块相似度 ↓
Static	PSR	12.5	0.774	0.341
	Instant NGP	28.8	0.864	0.254
Dynamic	Nerfies	29.5	0.849	0.299
	HyperNeRF	29.6	0.848	0.304
	DyNeRF	30.6	0.860	0.254
	Ours	**31.8**	**0.875**	**0.212**
Parts	NGP + 定义。	30.8	0.864	0.231
	哈希集成	30.5	0.857	0.257
Ablation	无深度	31.5	0.873	0.217
	无预热	31.0	0.866	0.234
	仅 16 张表格	31.5	0.871	0.218

5.5 评估方案

我们在 4 个留出相机视角上评估所有方法。图 7 展示了评估相机的空间排列。此外，为了节省计算时间，我们仅评估每台评估相机中均匀分布的 15 个时间步上的预测结果。我们在多个序列上验证过，与评估完整序列相比，仅评估 15 个时间步时，所有采用的图像指标差异最多为 0.02 分。

指标。我们报告三种图像指标，用于评估单个重建帧的视觉质量：峰值信噪比（PSNR）、结构相似性（SSIM）[Wang 等人 2004], 以及学习感知图像块相似度（LPIPS）[Zhang 等人 2018]。为聚焦面部区域，所有指标均在将预测结果与阿尔法掩码混合后，按帧进行评估。此外，我们还计算了 aJOD 指标 [Mantiuk 等人 2021]，该指标被 [Li 等人，2022b], 采用，用于衡量与参考视频之间的感知差异。

5.6 与最先进方法的比较

表3显示，NeRSemble 在所有图像指标上均以定量结果优于所有基线方法。尤其是，在对高频细节敏感的结构相似性和学习感知图像块相似度指标上，我们的方法取得了显著提升。这一观察结果与图 6 中的定性比较相符，我们的方法重建出了更好的面部细节。我们建议读者观看补充视频，以便更深入地直观分析我们的方法。

时间一致性评估。PSNR、SSIM 和学习感知图像块相似度（LPIPS）等逐帧指标无法考虑闪烁等时间伪影。因此，我们采用感知视频指标 JOD [Man- tiuk 等人 2021] 来衡量渲染视频与其对应真实值之间的视觉相似度。对于所有主要基线方法，我们以训练帧率的三分之一（即 24.3 fps）渲染视频，并对 JOD 值取平均。

表 4. 使用感知质量指标 Just-Objectionable-Difference（JOD）[Mantiuk 等人 2021]对时间一致性**进行评估**。数值越高，表示时间闪烁越少，且越接近真实视频。

方法	Instant-NGP	Nerfies	HyperNeRF	DyNeRF	Ours
JOD ↑	6.75	7.23	7.27	7.69	**7.86**

表 5. **与面部专用方法的评估比较**。 NeRSemble 的表现优于神经头部化身 [Grassal 等人 2022] 和 NeR-Face [Gafni 等人 2021]。

方法	峰值信噪比 ↑	结构相似性 ↑	学习感知图像块相似度 ↓
神经头部化身	31.0	0.927	0.041
NeRFace	35.2	0.956	0.047
Ours	**37.5**	**0.968**	**0.023**

表 6. **在 Neural 3D Video 数据集上的定量比较**。尽管 NeRSemble 的功能设计受到面部混合形状的启发，但它也能合理地建模通用动态场景。

方法	NeRFPlayer（Instant NGP）	NeRFPlayer (TensoRF)	HyperReel	Ours
峰值信噪比 ↑	30.3	30.7	**31.1**	29.9

所有验证视角和全部 10 个验证序列上的得分。时间评估结果见表 4。请注意，Instant NGP 基线方法完全不考虑时间，因此会在视频渲染中产生明显的闪烁伪影。图 8 展示了此类伪影的一个示例。相比之下，NeRSemble 提供了流畅的时间体验。

与 *Instant NGP* 的比较。 如图 6 所示，Instant NGP 基线方法逐帧生成的图像效果令人满意。然而，由于我们的相机布局较为稀疏，它很容易产生浮点伪影和散乱表面。相比之下，NeRSemble 会在多个时间步之间约束空间，这极大地有助于消除浮点伪影。即使 NeRSemble 在训练时未采用任何反漂浮伪影策略或额外损失项，这一点仍然成立。即便是我们模型的这种精简版本，其性能仍优于 Instant NGP（见表7顶行）。这表明，在我们的稀疏设置下，通过独立建模每一帧来提高表达能力（例如，Instant NGP 基线方法的参数量是我们模型的 10–15 倍）并不能带来更好的重建结果。

与面部专用方法的比较。 为了与 NHA 和 NeRFace 进行比较，我们在数据集中的 7 个序列上进行评估，排除其中 3 个运动较复杂的序列，因为 NHA 的预处理流程未能预测面部标志点、分割掩码和法线。此外，NHA 只合成头部，不包括躯干。因此，为了公平比较，我们只评估面部区域。表 5 显示，尽管这两种基线方法都是专为面部设计的，NeRSemble 仍优于它们。

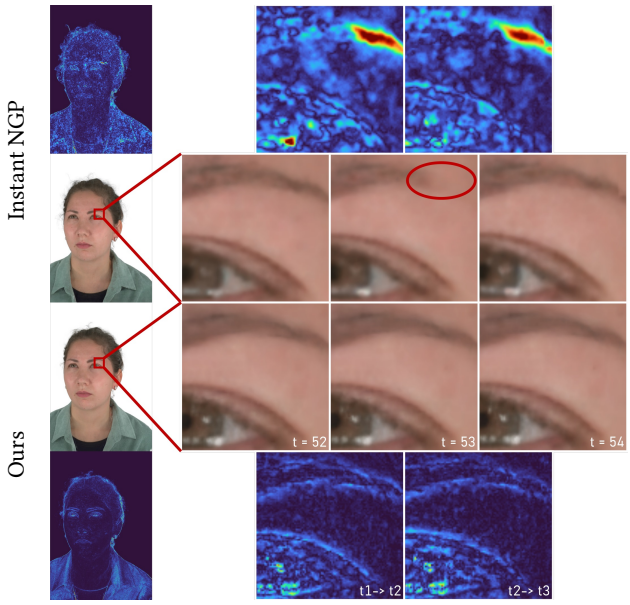


图 8. **时间一致性**。 我们展示了一个新视角的重新渲染结果及其时间差分图像（左侧）。右侧展示了 Instant NGP 基线方法在三个相邻帧之间产生的闪烁伪影：眉毛在帧与帧之间时而缩小、时而增大。相比之下，NeRSemble 具有更好的时间一致性。

在 *Neural 3D Video* 数据集上的比较 *[Li 等人，2022b]*。NeRSemble 对动态场景的内容没有强假设，因此适用于更一般的场景。为展示 NeRSemble 的通用性，我们在 Neural 3D Video 数据集 [Li 等人，2022b]中公开提供的 6 个序列上进行评估。我们遵循 NeRFPlayer [Song 等人 2023] 和 HyperReel [Attal 等人 2023] 的评估方案：将录制内容下采样至 1352 x 1014 像素分辨率，留出顶部中央视角用于评估，并报告所有 6 个序列的平均指标。我们还使用 COLMAP [Schönberger 等人2016] 重新计算了相机位姿，因为数据集提供的位姿略有偏差。表 6 展示了定量结果。我们未将原始 DyNeRF [Li 等人，2022b] 和 StreamRF [Li 等人，2022a] 纳入评估，因为它们的数值仅在 6 个可用序列中的 1 个上计算，因此无法与 NeRFPlayer 和 HyperReel 的结果进行比较。评估结果表明，尽管 NeRSemble 的功能设计受到面部混合形状的启发，它仍能合理地建模通用动态场景。不过，由于我们的方法依赖 Instant NGP，因此也继承了它的一些弱点。具体而言，它无法像 HyperReel 那样建模光折射。因此，NeRSemble 无法完美捕捉 Neural 3D Video 数据集中常见的窗玻璃和玻璃瓶所产生的效果。

5.7 消融实验

除了与基线方法进行比较外，我们还开展了多项实验，以验证我们的设计选择并了解 NeRSemble 的内部工作机制。



(a) NGP + 定义。 (b) 哈希集成 (c) 无预热 (d) 16 个哈希表 (e) 我们的方法 (f) 真实值

图 9. **模型组件消融实验。** 将 Instant NGP 与变形场结合（a），可以在场景中刚性运动的区域（例如躯干）生成清晰的细节，但难以处理更具挑战性的运动，例如嘴部运动。另一方面，采用哈希编码集成（b）能更好地处理复杂运动，但通常会生成更模糊的重建结果。结合这两个组件（c）可以发挥两种架构各自的优势，但在刚性运动区域生成的细节仍不及采用变形场的 Instant NGP。通过采用预热阶段，即使使用 16 个表格，也能恢复清晰的细节（d）；增加哈希编码的数量还可以进一步改善效果（e）。

架构组件的贡献。 我们通过消融实验评估了使用哈希集成和变形场的效果。表3显示，无论是采用 Instant NGP 骨干网络的变形场（NGP + Def.），还是单纯的哈希集成，其性能都无法达到最终模型的水平。然而，这两种架构各自都是强大的基线方法。在图9中，我们展示了定性结果。结果表明，基于变形的方法通常能够生成更锐利的重建结果，但面对更具挑战性、难以通过变形进行建模的运动时会表现不佳。另一方面，哈希集成具有通过特征融合对任意动态场景进行建模的表达能力，但对于简单运动通常会产生更加模糊的结果，因为它缺少变形场所提供的先验。表3中的定量结果证实了这些发现：哈希集成取得了较高的峰值信噪比，但学习感知图像块相似度值较差。

哈希表数量。 与使用 32 个哈希表相比，NeRSemble 使用 16 个哈希表时仅有轻微的性能损失。这证实了帧数与哈希表数量之间的比例具有良好的可扩展性，并且信息能在各个表格之间得到有效共享。

预热阶段的影响。 不进行预热的训练效果始终较差。我们认为，这是因为一开始就让模型使用所有哈希网格，会妨碍它学习与变形场对应的关系。因此，学到的哈希编码对齐效果较差，无法有效融合。在视觉上，这表现为渲染结果略显模糊。这一发现与 HyperNeRF 提出的在训练初期禁用环境维度的做法一致。

表7。 **浮点伪影去除技术的消融实验。** 视锥体剔除（VFC）和占据网格过滤（OGF）对各项指标的影响都微乎其微，因为它们去除的浮点伪影大多位于评估方案中未涵盖的区域。请注意，不使用任何额外损失项（L）或浮点伪影去除技术的 NeRSemble 基础版本，其表现已足以与表3中的所有基线方法相竞争。

$\mathcal{L}$	VFC	OGF	峰值信噪比 ↑	结构相似性 ↑	学习感知图像块相似度 ↓
☐	☐	☐	30.4	0.868	0.230
☒	☐	☐	31.8	**0.875**	0.213
☒	☒	☐	31.8	**0.875**	0.213
☒	☐	☒	**31.9**	**0.875**	**0.212**
☒	☒	☒	31.8	**0.875**	**0.212**

深度监督的影响。 由于移除深度监督仅会轻微影响性能，我们推测，12 个输入视图中的 RGB 信息已足以监督几何结构。不过，利用来自其他渠道的深度监督（例如拟合的三维形变模型或训练好的深度预测网络）仍可能有所裨益，因为这能融入来自 RGB 视频帧以外来源的数据先验。

浮点伪影去除技术。我们通过消融实验研究了三种抑制浮点伪影的策略。首先，我们单独考察所有额外损失项的效果，即掩码损失、深度监督和失真损失，并注意到它们对性能有显著影响，如表7所示。另一方面，视锥体剔除和占据网格过滤不会影响报告的指标，但在渲染新的相机轨迹时仍能提升视觉质量。



图 10. **混合权重**。 通过设置 $\beta_{t_1,\ i} = 0\ (i > 1)$ 来探究第一个哈希网格的内容，结果表明第一个哈希网格存储了某种平均表示（左图）。在右图中，我们依次为 $i \in \{2,\ 3,\ 4\}$ 设置 $\beta_{t_1,\ i} = 0.75$。每个表都存储了超出变形网络表示能力的额外细节。请注意，所有示例均使用 ω_{t_1}，而 t_1 表示第一帧。

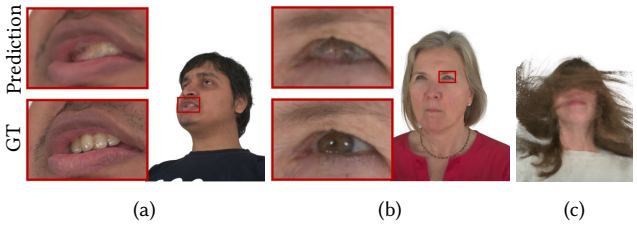


图 11. **失败案例**。 口腔内部的严重遮挡有时会造成空洞脸错觉，导致牙齿被错误地重建在口腔后部（a）。眼睛中光源的镜面反射可能导致罕见的眼部伪影（b）。变形场可能无法对极快的头发运动进行建模，因而妨碍规范哈希网格为某些帧合成清晰的结果（c）。

各个哈希网格的内容。 我们在图 10 中分析了各个哈希网格 H_i 的内容。为此，我们修改序列 t_1第一帧 t_1的学习得到的混合权重 $\beta_{t_1,\ i}\ (i > 1)$，同时保持 $\beta_{t_1,\ 1}$ 和变形编码 ω_{t_1} 不变。该实验表明，变形场 D 负责场景的刚体运动，因为修改 β_{t_1} 会产生良好对齐的外观变化，而头部保持静止。此外，H_1 似乎存储了与该人物的平均脸相当的表示。其余哈希网格的行为则类似于动态体积纹理，进一步为场景添加了其他方式无法解释的细节，例如拓扑复杂的变形、表情相关皱纹或光照变化。我们认为，第一个哈希网格 H_1 具有特殊地位，是因为它在整个训练过程中始终处于激活状态，而所有其他哈希网格则是在预热阶段逐步引入的。

5.8 局限性

在我们的实验中，我们展示了仅凭稀疏的多视角记录也能取得令人信服的结果；然而，仍存在各种局限性。由于 NeRSemble 通过变形场对不同时间步之间的显式对应关系进行建模，因此无法完美捕捉快速的头发运动（见图 11c）。为此，通过光流或可微物理引入运动先验，或许是一个值得未来探索的方向。

此外，我们的方法目前侧重于通过优化动态辐射场表示，恢复特定序列的外观和运动。因此，我们的方法无法学习能够泛化到不同序列的先验。我们认为，未来在可跨不同身份和面部表情泛化的动态神经辐射场方面有很大的研究潜力。对逼真 4D 虚拟形象分布的学习先验有助于进一步约束优化过程。这对于单目输入，或采集面部区域（例如口腔内部）尤为重要，因为这些区域在大部分录制时间内通常会被遮挡，因此重建质量可能较差（见图 11）。

6 结论

在这项工作中，我们提出了一种新方法和数据集，专注于从多视角视频输入中对动画人头进行辐射场重建。为此，我们引入了一个全新的多视角视频基准，其中包含多种不同的人头，涵盖超过 220 个身份和 4700 个序列。我们进一步提出了一种新方法，用于生成任意视点和时间步下的照片级真实感重渲染结果，并希望我们的数据集及其配套基准能够为相关领域作出重要贡献，推动数字人领域的未来研究。

我们提出的用于时空NeRF的新型表示利用变形场分离粗略运动，并使用哈希网格编码的集成来建模细微变形并增加模型的时间容量。我们的实验表明，NeRSemble能够根据多视角视频输入生成时间上连贯且细节丰富的体积重建结果，其性能显著优于现有基线方法，尤其是在序列包含复杂运动时。

致谢

本研究得到 ERC Starting Grant Scan2CAD (804724)、德国研究基金会（DFG）项目“Making Machine Learning on Static and Dynamic 3D Data Practical”，以及德国研究基金会（DFG）研究组“Learning and Simulation in Visual Computing”的支持。我们还要感谢 Maximilian Knörl 协助采集数据，并感谢 Angela Dai 为视频配音。

参考文献

ShahRukh Athar、Zexiang Xu、Kalyan Sunkavalli、Eli Shechtman 和 Zhixin Shu。2022年。RigNeRF：完全可控的神经三维人像。发表于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集。20364–20373。Benjamin Attal、Jia-Bin Huang、Christian Richardt、Michael Zollhoefer、Johannes Kopf、Matthew O'Toole 和 Changil Kim。2023年。HyperReel：基于射线条件采样的高保真六自由度视频。计算机视觉与模式识别会议（2023年）。Jonathan T. Barron、Ben Mildenhall、Dor Verbin、Pratul P. Srinivasan 和 Peter Hedman。2022年。Mip-NeRF 360：无界抗锯齿神经辐射场。计算机视觉与模式识别会议（2022年）。V. Blanz、C. Basso、T. Poggio 和 T. Vetter。2003年。图像与视频中的面部重动画。（2003年），641–650。Volker Blanz 和 Thomas Vetter。1999年。用于三维面部合成的可变形模型。发表于第*26*届计算机图形学与交互技术年度会议论文集（*SIGGRAPH '99*）。ACM Press/Addison-Wesley Publishing Co.，美国，187–194。https://doi.org/10.1145/311535.311556

Michael Broxton、John Flynn、Ryan Overbeck、Daniel Erickson、Peter Hedman、Matthew Duvall、Jason Dourgarian、Jay Busch、Matt Whalen 和 Paul Debevec。2020年。采用分层网格表示的沉浸式光场视频。*ACM* 图形学汇刊（*TOG*）39, 4（2020年），86–1。Ang Cao 和 Justin Johnson。2023年。HexPlane：一种用于动态场景的快速表示方法。计算机视觉与模式识别会议（2023年）。Anpei Chen、Zexiang Xu、Andreas Geiger、Jingyi Yu 和 Hao Su。2022年。TensoRF：张量辐射场。载于欧洲计算机视觉会议（*ECCV*）。Shiyang Cheng、Irene Kotsia、Maja Pantic 和 Stefanos Zafeiriou。2018年。4dfab：一个用于面部表情分析和生物特征识别应用的大规模四维数据库。载于*IEEE*计算机视觉与模式识别会议论文集。5117–5126。Alvaro Collet、Ming Chuang、Pat Sweeney、Don Gillett、Dennis Evseev、David Calabrese、Hugues Hoppe、Adam Kirk 和 Steve Sullivan。2015年。高质量、可流式传输的自由视点视频。*ACM* 图形学汇刊（*ToG*）34, 4（2015年），1–13。Darren Cosker、Eva Krumhuber 和 Adrian Hilton。2011年。一种符合面部动作编码系统（FACS）标准的三维动态动作单元数据库及其在三维动态可变形人脸建模中的应用。载于*2011*年国际计算机视觉会议。IEEE，2296–2303。Daniel Cudeiro、Timo Bolkart、Cassidy Laidlaw、Anurag Ranjan 和 Michael Black。2019年。三维说话风格的捕获、学习与合成。载于*IEEE*计算机视觉与模式识别会议论文集（*CVPR*）。10101–10111。Rémi Flamary、Nicolas Courty、Alexandre Gramfort、Mokhtar Z. Alaya、Aurélie Bois- bunon、Stanislas Chambon、Laetitia Chapel、Adrien Corenflos、Kilian Fatras、Nemo Fournier、Léo Gautheron、Nathalie T.H. Gayraud、Hicham Janati、Alain Rakotoma- monjy、Ievgen Redko、Antoine Rolet、Antony Schutz、Vivien Seguy、Danica J. Sutherland、Romain Tavenard、Alexander Tong 和 Titouan Vayer。2021年。POT：Python最优传输。机器学习研究期刊22, 78（2021年），1–8。http://jmlr.org/papers/v22/20-451.html Sara Fridovich-Keil、Giacomo Meanti、Frederik Rahbæk Warburg、Benjamin Recht 和 Angjoo Kanazawa。2023年。K-Planes：空间、时间和外观中的显式辐射场。载于计算机视觉与模式识别会议。Fridovich-Keil 和 Yu、Matthew Tancik、Qinhong Chen、Benjamin Recht 和 Angjoo Kanazawa。2022年。Plenoxels：无神经网络辐射场。载于计算机视觉与模式识别会议。Guy Gafni、Justus Thies、Michael Zollhöfer 和 Matthias Nießner。2021年。用于单目四维人脸数字化身重建的动态神经辐射场。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集。8649–8658。Xuan Gao、Chenglai Zhong、Jun Xiang、Yang Hong、Yudong Guo 和 Juyong Zhang。2022年。从单目视频重建个性化语义人脸NeRF模型。*ACM* 图形学汇刊（亚洲计算机图形学国际会议论文集）41, 6（2022年）。https://doi.org/10.1145/3550454.3555501 Simon Giebenhain、Tobias Kirschstein、Markos Georgopoulos、Martin Rünz、Lourdes Agapito 和 Matthias Nießner。2022年。神经参数化头部模型的学习。Philip-William Grassal、Malte Prinzler、Titus Leistner、Carsten Rother、Matthias Nießner 和 Justus Thies。2022年。基于单目RGB视频的神经头部化身。载于*IEEE/CVF*计算机视觉与模式识别会议论文集。18653–18664。Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。2016年。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集。770–778。Yang Hong、Bo Peng、Haiyao Xiao、Ligang Liu 和 Juyong Zhang。2022年。HeadNeRF：一种实时NeRF参数化头部模型。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）。Animesh Karnewar、Tobias Ritschel、Oliver Wang 和 Niloy Mitra。2022年。Relu场：微小却强大的非线性。载于*ACM SIGGRAPH 2022*年会议论文集。1–9。Michael M. Kazhdan、Matthew Bolitho 和 Hugues Hoppe。2006年。泊松曲面重建。载于第四届欧洲图形学几何处理研讨会论文集（意大利撒丁岛卡利亚里）（*SGP '06*，第*256*卷），Alla Sheffer 和 Konrad Polthier（编）。欧洲图形学协会，瑞士 Aire-la-Ville，瑞士，61–70。http://dl.acm.org/citation.cfm?id=1281957.1281965 Lingzhi Li、Zhen Shen、Li Shen、Ping Tan 等。2022a。用于三维视频合成的流式辐射场。载于神经信息处理系统进展。Ruilong Li、Matthew Tancik 和 Angjoo Kanazawa。2022c。NerfAcc：一种通用的NeRF加速工具箱。*arXiv*预印本 *arXiv:2210.04847*（2022年）。Tianye Li、Timo Bolkart、Michael. J. Black、Hao Li 和 Javier Romero。2017年。从四维扫描中学习面部形状和表情模型。*ACM* 图形学汇刊（亚洲计算机图形学国际会议论文集）36, 6（2017年），194:1–194:17。https://doi.org/10.1145/3130800. 3130813 Tianye Li、Mira Slavcheva、Michael Zollhoefer、Simon Green、Christoph Lassner、Changil Kim、Tanner Schmidt、Steven Lovegrove、Michael Goesele、Richard Newcombe 等。2022b。从多视角视频合成神经三维视频。载于*IEEE/CVF*计算机视觉与模式识别会议论文集。5521–5531。Shanchuan Lin、Andrey Ryabtsev、Soumyadip Sengupta、Brian L Curless、Steven M Seitz 和 Ira Kemelmacher-Shlizerman。2021年。实时高分辨率背景抠图。载于*IEEE/CVF*计算机视觉与模式识别会议论文集。8762–8771。

Stephen Lombardi、Tomas Simon、Jason Saragih、Gabriel Schwartz、Andreas Lehrmann 和 Yaser Sheikh。2019年。神经体积：从图像中学习动态可渲染体积。*ACM* 图形学汇刊 38，4，文章 65（2019年7月），14 页。Stephen Lombardi、Tomas Simon、Gabriel Schwartz、Michael Zollhoefer、Yaser Sheikh 和 Jason Saragih。2021年。用于高效神经渲染的体积基元混合。*ACM* 图形学汇刊（*TOG*）40, 4（2021年），1–13。Rafał K. Mantiuk、Gyorgy Denes、Alexandre Chapiro、Anton Kaplanyan、Gizem Rufo、Romain Bachy、Trisha Lian 和 Anjul Patney。2021年。FovVideoVDP：宽视场视频的可见差异预测器。*ACM* 图形学汇刊 40，4，文章 49（2021年7月），19 页。https://doi.org/10.1145/3450626.3459831 Ben Mildenhall、Pratul P. Srinivasan、Matthew Tancik、Jonathan T. Barron、Ravi Ra- mamoorthi 和 Ren Ng。2020年。神经辐射场：将场景表示为用于视图合成的神经辐射场。载于 欧洲计算机视觉会议。Thomas Müller、Alex Evans、Christoph Schied 和 Alexander Keller。2022年。采用多分辨率哈希编码的即时神经图形基元。*ACM* 图形学汇刊 41，4，文章 102（2022年7月），15 页。https://doi.org/10.1145/3528223. 3530127 Keunhong Park、Utkarsh Sinha、Jonathan T. Barron、Sofien Bouaziz、Dan B Goldman、Steven M. Seitz 和 Ricardo Martin-Brualla。2021a。Nerfies：可变形神经辐射场。国际计算机视觉大会（2021年）。Keunhong Park、Utkarsh Sinha、Peter Hedman、Jonathan T. Barron、Sofien Bouaziz、Dan B Goldman、Ricardo Martin-Brualla 和 Steven M. Seitz。2021b。HyperNeRF：用于拓扑变化神经辐射场的高维表示。*ACM* 图形学汇刊 40，6，文章 238（2021年12月）。Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga、Alban Des- maison、Andreas Kopf、Edward Yang、Zachary DeVito、Martin Raison、Alykhan Tejani、Sasank Chilamkurthy、Benoit Steiner、Lu Fang、Junjie Bai 和 Soumith Chin- tala。2019年。PyTorch：一种命令式风格的高性能深度学习库。载于 神经信息处理系统进展 *32*。Curran Associates, Inc.，8024–8035。http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf Pascal Paysan、Reinhard Knothe、Brian Amberg、Sami Romdhani 和 Thomas Vetter。2009年。用于姿态和光照不变人脸识别的三维人脸模型。载于 *2009*年第六届 *IEEE* 先进视频与信号监控国际会议。Ieee，296–301。Albert Pumarola、Enric Corona、Gerard Pons-Moll 和 Francesc Moreno-Noguer。2020年。D-NeRF：动态场景的神经辐射场。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集。Konstantinos Rematas、Andrew Liu、Pratul P. Srinivasan、Jonathan T. Barron、Andrea Tagliasacchi、Tom Funkhouser 和 Vittorio Ferrari。2022年。Urban Radiance Fields。计算机视觉与模式识别会议 （2022年）。Neus Sabater、Guillaume Boisson、Benoit Vandame、Paul Kerbiriou、Frederic Babon、Matthieu Hog、Tristan Langlois、Remy Gendrot、Olivier Bureller、Arno Schubert 和 Valerie Allie。2017年。多视图光场视频的数据集与流程。载于 *CVPR*研讨会。Johannes Lutz Schönberger 和 Jan-Michael Frahm。2016年。重新审视运动恢复结构。载于 计算机视觉与模式识别会议（*CVPR*）。Johannes Lutz Schönberger、Enliang Zheng、Marc Pollefeys 和 Jan-Michael Frahm。2016年。非结构化多视图立体视觉的逐像素视图选择。载于 欧洲计算机视觉会议（*ECCV*）。Liangchen Song、Anpei Chen、Zhong Li、Zhang Chen、Lele Chen、Junsong Yuan、Yi Xu 和 Andreas Geiger。2023年。NeRFPlayer：采用分解式神经辐射场的可流式传输的动态场景表示。*IEEE*可视化与计算机图形学汇刊 29，5（2023年），2732–2742。Cheng Sun、Min Sun 和 Hwann-Tzong Chen。2022年。直接体素网格优化：辐射场重建的超快速收敛。载于 计算机视觉与模式识别会议。Matthew Tancik*、Ethan Weber*、Evonne Ng*、Ruilong Li、Brent Yi、Terrance Wang、Alexander Kristoffersen、Jake Austin、Kamyar Salahi、Abhik Ahuja、David McAllister 和 Angjoo Kanazawa。2022年。Nerfstudio：神经辐射场开发框架。https://github.com/nerfstudio-project/nerfstudio Daoye Wang、Prashanth Chandran、Gaspard Zoss、Derek Bradley 和 Paulo Gotardo。2022a。MoRF：用于多视图神经头部建模的可变形辐射场。载于*ACM SIGGRAPH 2022*年会议论文集 （温哥华，不列颠哥伦比亚省，加拿大）（*SIGGRAPH '22*）。计算机协会，纽约，纽约州，美国，文章 55，9 页。https://doi.org/10.1145/3528233.3530753 Kaisiyuan Wang、Qianyi Wu、Linsen Song、Zhuoqian Yang、Wayne Wu、Chen Qian、Ran He、Yu Qiao 和 Chen Change Loy。2020年。MEAD：用于情感说话人脸生成的大规模视听数据集。载于 欧洲计算机视觉会议。Liao Wang、Jiakai Zhang、Xinhang Liu、Fuqiang Zhao、Yanshun Zhang、Yingliang Zhang、Minye Wu、Jingyi Yu 和 Lan Xu。2022b。用于实时动态辐射场渲染的傅里叶 PlenOctrees。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集（*CVPR*）。13524–13534。Zhou Wang、Alan C Bovik、Hamid R Sheikh 和 Eero P Simoncelli。2004年。图像质量评估：从误差可见性到结构相似性。*IEEE*图像处理汇刊 13，4（2004年），600–612。

Cheng-hsin Wuu、Ningyuan Zheng、Scott Ardisson、Rohan Bali、Danielle Belko、Eric Brockmeyer、Lucas Evans、Timothy Godisart、Hyowon Ha、Alexander Hypes、Tay- lor Koska、Steven Krenn、Stephen Lombardi、Xiaomin Luo、Kevyn McPhail、Laura Millerschoen、Michal Perdoch、Mark Pitts、Alexander Richard、Jason Saragih、Junko Saragih、Takaaki Shiratori、Tomas Simon、Matt Stewart、Autumn Trimble、Xinshuo Weng、David Whitewolf、Chenglei Wu、Shoou-I Yu 和 Yaser Sheikh。2022年。Multiface：神经人脸渲染数据集。载于 *arXiv*。https://doi.org/10.48550/ ARXIV.2207.11243 Yiheng Xie、Towaki Takikawa、Shunsuke Saito、Or Litany、Shiqin Yan、Numair Khan、Federico Tombari、James Tompkin、Vincent Sitzmann 和 Srinath Sridhar。2022年。视觉计算及其他领域中的神经场。《计算机图形论坛》（2022年）。https://doi.org/10.1111/cgf.14505 Tarun Yenamandra、Ayush Tewari、Florian Bernard、Hans-Peter Seidel、Mohamed El- gharib、Daniel Cremers 和 Christian Theobalt。2021年。i3DMM：人头深度隐式三维可变形模型。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集。12803–12813。Changqian Yu、Jingbo Wang、Chao Peng、Changxin Gao、Gang Yu 和 Nong Sang。2018年。BiseNet：用于实时语义分割的双边分割网络。载于

欧洲计算机视觉会议（*ECCV*）论文集. 325–341. Richard Zhang、Phillip Isola、Alexei A Efros、Eli Shechtman 和 Oliver Wang。2018年。深度特征作为感知度量的非凡效能。载于*IEEE*计算机视觉与模式识别会议论文集. 586–595. Xing Zhang、Lijun Yin、Jeffrey F Cohn、Shaun Canavan、Michael Reale、Andy Horowitz、Peng Liu 和 Jeffrey M Girard。2014年。BP4D-Spontaneous：高分辨率自发三维动态面部表情数据库。《图像与视觉计算》32, 10 (2014), 692–706. Yufeng Zheng、Victoria Fernández Abrevaya、Marcel C. Bühler、Xu Chen、Michael J. Black 和 Otmar Hilliges。2022年。I M Avatar：基于视频的隐式可变形头部虚拟形象。载于计算机视觉与模式识别（*CVPR*）。Wojciech Zielonka、Timo Bolkart 和 Justus Thies。2023年。即时体积式头部虚拟形象。计算机视觉与模式识别会议（2023年）。C Lawrence Zitnick、Sing Bing Kang、Matthew Uyttendaele、Simon Winder 和 Richard Szeliski。2004年。使用分层表示进行高质量视频视图插值。*ACM* 图形学汇刊（*TOG*）23, 3 (2004), 600–608.