

《GaussianTalker: Real-Time High-Fidelity Talking Head Synthesis with Audio-Driven 3D Gaussian Splatting》论文复现报告

摘要

GaussianTalker 是一种用于实时生成可控制姿态的说话头部的新颖框架。它利用了 3D 高斯泼溅（3D Gaussian Splatting, 3DGS）的快速渲染能力，同时解决了直接用语音音频控制 3DGS 的挑战。GaussianTalker 构建了头部的规范 3DGS 表示，并使其与音频同步变形。其中一个关键创新在于构建了头部的规范 3DGS 表示，并将 3D 高斯属性编码为共享的隐式特征表示，与音频特征合并，以操纵每个高斯属性。这种设计利用了空间感知特征，使相邻点之间相互作用。然后，特征嵌入被输入到空间-音频注意力模块，该模块预测每个高斯属性的逐帧偏移量。与之前用于操纵众多高斯及其复杂参数的拼接或乘法方法相比，GaussianTalker 更加稳定。本文关注 3DGS 模型的 few-shot 性能表现，探索 GaussianTalker 在较短的个人讲话视频（10-20s）做训练数据下的实际效果。通过复现 GaussianTalker 代码，结合定量与定性实验对其 few-shot 效果进行详细分析。

关键词：Talking Head Synthesis; 3D Gaussian Splatting; few shot

1 引言

利用任意语音音频驱动并生成说话的头部视频是一项热门的研究任务，它具有多种用途，包括数字人、虚拟形象、电影制作和视频会议等。近两年，众多的研究工作将神经辐射场 (NeRF) [12] 应用于创建可控制姿态的说话肖像。通过直接在 NeRF 的多层感知器 (MLP) 中对音频特征进行条件设定，这些方法可以合成包含与输入音频同步的嘴唇动作的 3D 头部结构。尽管这些基于 NeRF 的技术实现了高质量和一致的视觉输出，但其缓慢的推理速度限制了它们的实用性。

最近，3DGS [7] 被认为是 NeRF 的可行替代方案，它提供了与 NeRF 可比的渲染质量，同时显著提高了推理速度。GaussianTalker [3] 首次利用 3D 高斯表示来发挥其快速场景建模能力，以实现音频驱动的动态面部动画。通过构建规范头部形状的静态 3DGS 表示，并使其与音频同步变形，成功地模拟了音频特征与每个高斯椭球运动之间的相关性。原论文中的定性和定量实验证明了 GaussianTalker 在面部保真度、唇同步准确性和渲染速度方面优于先前方法。

本文的复现工作针对 few shot 条件下 GaussianTalker 的实际效果展开。常规的 Talking Head Synthesis 方法的训练数据为 3-5 分钟的单人讲话视频，较长的时常限制了其实际应用

场景。本文尝试在 10-20s 的个人讲话视频下训练 GaussianTalker 模型，并测试其实际效果，以探索 few shot 条件下基于 3DGS 的 Talking Head Synthesis 方法的可行性。

2 相关工作

2.1 音频驱动的说话肖像合成

音频驱动的说话肖像合成旨在基于音频输入创建逼真并拥有准确嘴唇动作的面部动画。早期基于 2D GAN 的方法 [14, 21] 实现了照片级真实感，但由于缺乏 3D 几何形状，无法控制头部姿态。为了实现头部姿态控制，一些工作 [11, 18] 利用基于中间表示的方法，其中 2D landmark 和 3D 可变形模型增强了模型唇同步效果，使其能够调整头部方向。然而，这些方法导致了新的问题，如中间表示带来额外误差，个人身份信息丢失和真实性的降低。

最近，由于神经辐射场 (NeRF) [12] 能够捕获复杂的场景，它已被用于探索说话肖像的合成。AD-NeRF [6] 率先使用 NeRF 的隐式表示进行音频条件输入，但头部和躯干的单独网络限制了其灵活性。后续基于 NeRF 的方法 [10, 16] 实现了高质量的结果，但渲染速度较慢。虽然 RAD-NeRF [17] 和 ER-NeRF [8] 提高了合成效率和质量，但可控制姿态的 3D 说话头部的实时渲染仍然具有挑战性。

2.2 基于 3D 高斯泼溅的面部重建

3DGS [7] 是点云渲染中的一项开创性技术，它利用众多各向异性的椭球来精确表示场景。每个点代表一个 3D 高斯分布，其均值、协方差、不透明度和球谐函数参数通过优化，以准确捕捉场景的形状和外观。这种方法有效地解决了点渲染中的常见问题，如输出结果的间隙。此外，通过与基于 tile 的光栅化算法相结合，3DGS 实现了快速训练和实时渲染能力。

先前的面部重建方法主要依赖于 3D 可变形模型 (3DMM) [5] 或利用神经隐式表示 [1]。最近的方法 [2, 4] 已转向采用 3DGS 表示，旨在利用快速训练和渲染的优势，同时实现有竞争力的照片级真实感。GaussianAvatars [15] 通过在 FLAME [9] 网格上绑定 3D 高斯来重建头部化身。MonoGaussianAvatar [2] 通过使用线性混合蒙皮 (LBS) 将 3D 高斯的平均位置从规范空间转移到变形空间来学习显式头部化身，并同时通过变形场调整其他高斯参数。GaussianHead [19] 采用运动变形场来适应面部运动，同时保留头部几何形状，并单独利用三平面来保留单个 3D 高斯的外观信息。然而，上述方法往往依赖于参数模型进行面部动画。与先前的工作相比，原论文工作的音频驱动方法不仅不需要语音之外的数据，而且并且能够应用于新的音频。

3 本文方法

3.1 本文方法概述

GaussianTalker 旨在实时合成由音频输入驱动的高保真、可控制姿态的说话头部图像。模型在一个说话肖像视频 $V = \{I_n\}$ 上进行训练，该视频由某个身份的 N 个图像帧组成。目标是重建一组规范的 3D 高斯，它们作为说话头部的平均表示。接着，模型学习一个变形模块，该模块根据相应的输入音频使 3D 高斯形变。在推理过程中，对于输入音频 a_n ，变形模块预

测每个高斯属性的偏移量，并且变形后的高斯在视点 π_n 处进行光栅化，以输出新的图像 \hat{I}_n 。整体框架如图 1 所示。

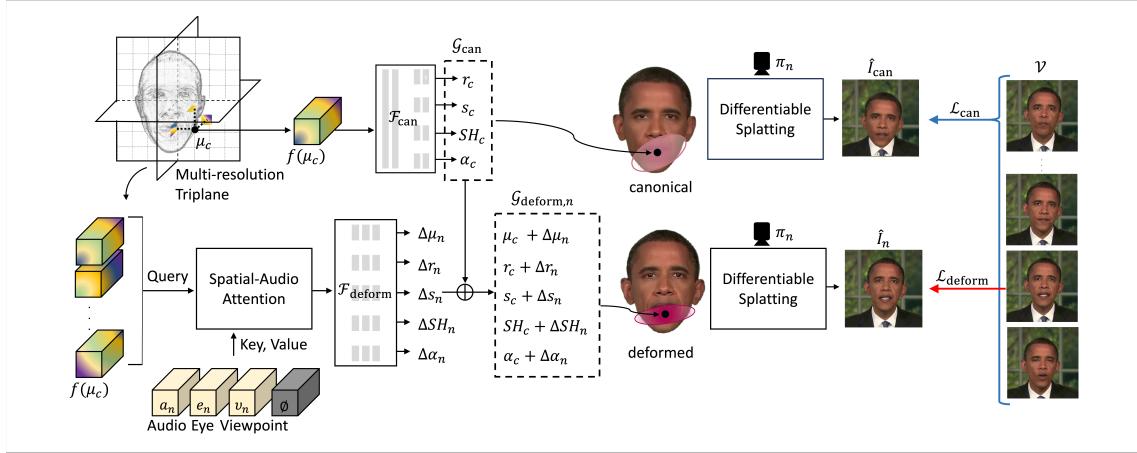


图 1. 方法示意图

3.2 使用三平面表示学习规范 3D 高斯

本节介绍使用 3D 高斯表示学习说话头部规范形状的细节。GaussianTalker 通过学习低维特征表示来修改 3D 高斯的表示，该低维特征表示稍后可与音频特征合并，以用于每个高斯的变形。具体来说，为了对规范 3D 头部的空间信息进行编码，GaussianTalker 利用隐式神经辐射场编码的空间信息的优势，对于每个规范的 3D 位置 μ_c ，从多分辨率三平面表示中提取特征嵌入 $f(\mu_c)$ 。这些特征嵌入用于计算每个点的缩放 S_c 、旋转 r_c 、球谐函数 SH_c 和不透明度 α_c 。这些计算得到的属性构成了说话头部的规范 3D 高斯，表示为：

$$G_{can} = \{\mu_c, r_c, s_c, SH_c, \alpha_c\} \quad (1)$$

为了对规范 3D 头部的空间信息进行编码，GaussianTalker 采用多分辨率三平面表示，它由三个正交的 2D 特征网格 $P = \{P^{xy}, P^{yz}, P^{zx}\}$ 构成。对于位置为 μ 的单个 3D 高斯，其每个坐标值在 $[0, R]$ 之间进行归一化，并且通过将点插值到每个平面的规则间隔的 2D 网格中来计算其相应特征。对于每个平面，这些特征使用哈达玛积进行组合，然后沿着不同维度进行拼接，以为每个规范高斯位置 μ_c 生成长度为 H 的最终特征向量 $f(\mu_c)$ ，如下所示：

$$f(\mu) = \bigcup \prod_{p \in P} \text{interp}(p, \zeta_p(\mu_c)), \quad (2)$$

其中 $\zeta_p(\mu)$ 表示 μ 在第 P 个平面上的投影，“interp”表示将点双线性插值到规则间隔的 2D 网格中。

规范 3D 高斯的属性预测是从相应的特征表示 $f(\mu_c)$ 中获取的。具体来说，GaussianTalker 使用一组多层感知器 (MLP) 层，记为 $F_{can}(\cdot)$ ，以便将中的特征映射为平均缩放 s_c 、平均旋转 r_c 、平均球谐函数 SH_c ，以及平均不透明度值 α_c ，如下所示：

$$\{s_c, r_c, SH_c, \alpha_c\} = F_{can}(f(\mu)) \quad (3)$$

3.3 学习音频驱动的 3D 高斯变形

为了对动态特征与大量 3D 高斯之间的关系进行建模, GaussianTalker 在注意力机制中将输入语音音频与编码特征融合, 以便为第 n 个图像帧生成音频感知特征 h_n 。后续帧中每个高斯属性的变形偏移量直接取决于特征 h_n 。最后, 第 n 个图像帧的变形后的 3D 高斯集定义为:

$$G_{\text{deform}, n} = \{\mu_c + \Delta\mu_n, r_c + \Delta r_n, s_c + \Delta s_n, SH_c + \Delta SH_n, \alpha_c + \Delta \alpha_n\}, \quad (4)$$

其中 $\Delta\mu_n, \Delta r_n, \Delta s_n, \Delta SH_n, \Delta \alpha_n$ 分别是第 n 帧中 3D 位置、缩放、旋转、球谐函数参数和不透明度的变形偏移量。

4 复现细节

4.1 与已有开源代码对比

由于原论文的代码已经开源, 本文在原论文的开源代码基础上修改数据集以及相应的划分代码, 以测试 GaussianTalker 方法在 few shot 条件下的性能表现。并与正常训练的模型和 few shot 条件下的其他方法分别进行对比, 以分析和评估 GaussianTalker 的 few shot 可行性。

4.2 实验环境

实验环境与原论文中保持一致, python 版本为 3.7, 神经网络的实现框架采用 tensorflow v2.8.0, 模型的训练和推理均在单张 NVIDIA RTX 3090 上进行。

4.3 创新点

原论文采用的数据集来自 AD-NeRF [6], GeneFace [20] 和 HDTF [22], 其中每个视频均为 3-5 分钟的个人讲话视频。为了测试模型在 few shot 条件下的性能表现, 本次复现从原训练数据中进行挑选并裁剪。最终选择了 Obama, May 以及 Macron 三人的个人讲话视频, 并分别裁剪成 10s, 15s 和 20s 三种不同时长作为训练数据, 剩下部分的视频作为推理时的 ground truth。

表 1. 使用不同长度的训练视频时的方法比较

Methods	ER-NeRF			SyncTalk			GaussianTalker		
	10s	15s	20s	10s	15s	20s	10s	15s	20s
PSNR \uparrow	30.012	31.432	31.290	30.068	32.199	34.687	31.068	32.147	32.476
SSIM \uparrow	0.918	0.917	0.919	0.917	0.929	0.943	0.927	0.928	0.931
LPIPS \downarrow	0.021	0.022	0.014	0.026	0.024	0.023	0.018	0.018	0.016
LMD \downarrow	3.392	3.017	2.803	3.237	2.982	2.574	3.166	3.037	2.928

5 实验结果分析

本次复现实验采用 10s, 15s, 20s 三种时长的视频作为训练数据，并将 GaussianTalker 与 ER-NeRF [8] 和 SyncTalk [13] 两种方法进行比较。实验结果如表 1 所示。指标 PSNR, SSIM 和 LPIPS 用以评价生成图片的质量，而 LMD 用以评价合成唇部与音频的同步精度。可以看出三种方法随训练视频时长的降低，图片质量和同步精度都有所下降。其中，SyncTalker 拥有最好的生成图片质量，而 ER-NeRF 则有更好的同步精度。GaussianTalker 相对于这两种方法表现较差。实验表明 GaussianTalker 在训练视频仅有 10-20s 的 few shot 条件下表现不佳，无法合成高质量的讲话视频。图 2 对 GaussianTalker 在 Obama 和 May 两个视频数据上的推理结果进行了可视化比较，同样说明了 GaussianTalker 所存在的问题。

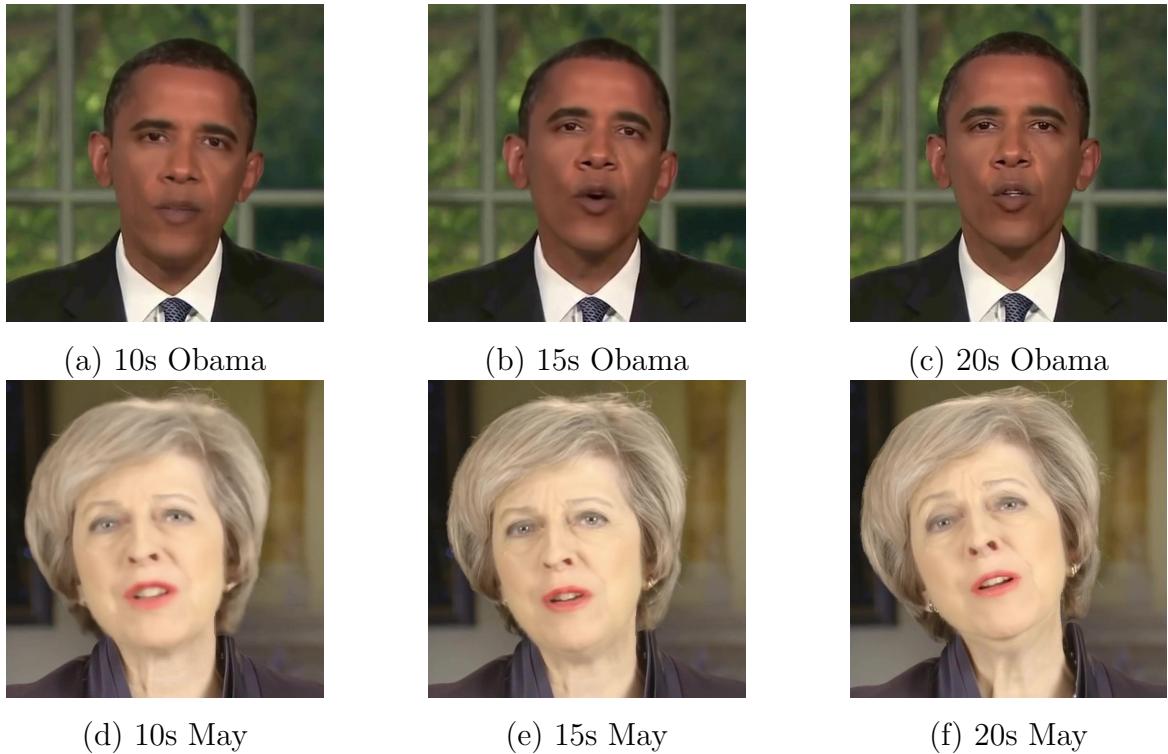


图 2. 采用不同时长视频训练的可视化比较

6 总结与展望

本次研究报告探索了一种利用音频驱动并生成说话头部视频的新颖框架 GaussianTalker，它创新性地利用了 3DGS 来对头部进行建模并驱动。在方法复现上，本文采用了原论文的开源代码并在其基础上对 GaussianTalker 的 few shot 效果展开了研究与分析。结果表明 GaussianTalker 本身在较短的个人数据训练下，难以进行全新音频的推理。因此未来，我会在 GaussianTalker 的基础上，进一步尝试将其改进为 few shot 方法，即能够在少量数据下实现较好的推理结果。

参考文献

- [1] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 20364–20373, 2022.
- [2] Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024.
- [3] Kyusun Cho, Joungbin Lee, Heeji Yoon, Yeobin Hong, Jaehoon Ko, Sangjun Ahn, and Seungryong Kim. Gaussiantalker: Real-time high-fidelity talking head synthesis with audio-driven 3d gaussian splatting, 2024.
- [4] Helisa Dhamo, Yinyu Nie, Arthur Moreau, Jifei Song, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Headgas: Real-time animatable head avatars via 3d gaussian splatting. In *European Conference on Computer Vision*, pages 459–476. Springer, 2025.
- [5] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022.
- [6] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Adnerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5784–5794, 2021.
- [7] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023.
- [8] Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin Gu. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7568–7578, 2023.
- [9] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.
- [10] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. In *European conference on computer vision*, pages 106–125. Springer, 2022.
- [11] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (ToG)*, 40(6):1–17, 2021.

- [12] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.
- [13] Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. SyncTalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 666–676, 2024.
- [14] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020.
- [15] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024.
- [16] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *European conference on computer vision*, pages 666–682. Springer, 2022.
- [17] Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *arXiv preprint arXiv:2211.12368*, 2022.
- [18] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 716–731. Springer, 2020.
- [19] Jie Wang, Jiu-Cheng Xie, Xianyan Li, Feng Xu, Chi-Man Pun, and Hao Gao. Gaussianhead: High-fidelity head avatars with learnable gaussian derivation. *arXiv preprint arXiv:2312.01632*, 2023.
- [20] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. Gene-face: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*, 2023.
- [21] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision*, pages 85–101. Springer, 2022.

- [22] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021.