中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-15

论文引用格式:Zhou Tianyang, Mao Yuxiang, Ye Yongjing, Xia Shihong. Animatable 3D Gaussian head avatars with texture prior[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI: 10.11834/jig.260036. (周天阳, 毛宇翔, 叶永竞, 夏时洪. 融合纹理先验的3D高斯人头化身重建与动画[J/OL]. 中国图象图形学报, XXXX:1-15. DOI: 10.11834/jig.260036.)[DOI:10.11834/jig.260036]

融合纹理先验的3D高斯人头化身重建与动画

周天阳[1,2],毛宇翔[1,2],叶永竞[1],夏时洪[1,2*]

1. 中国科学院计算技术研究所,北京　100190; 2. 中国科学院大学,北京　101408

摘　要: **目的**　提出了一种基于3DGS的3D人头化身建模方法TPAvatar(Avatar with Texture Prior),能够从多视角或单目视频序列数据中高效重建高保真可动画的3D人头化身,解决现有方法重建训练速度慢、难以重建精细皱纹细节的问题。**方法**　TPAvatar通过构建一个轻量化网络模型学习高斯属性的特征隐空间,并首次提出利用预训练的DINOv2模型从建模对象的纹理图中提取视角无关的身份外观先验,构建UV空间对齐的身份特征。在表情驱动方面,TPAvatar为每个高斯点构建一组隐式表情特征基,通过网格绑定和表情特征基的线性组合实现模型的高效动画。**结果**　在多视角数据集NeRSemble和单目数据集INSTA上的实验结果表明,TPAvatar可以实现高质量的多视角或单目人头化身重建。在多视角重建场景中,与基线方法GaussianAvatars/GEM相比,TPAvatar将重建时间从8/12小时缩短到了1.5小时,同时取得了更高的重建质量,在测试集上LPIPS分别降低了0.0037/0.0131;与基线方法RGBAvatar相比,TPAvatar在保持快速重建优点的同时显著提升了视角泛化性,在新视角合成任务中LPIPS降低了0.0155。在单目重建场景中,相对于最优基线方法RGBAvatar, LPIPS降低了0.0016。**结论**　TPAvatar是一种可实时动画的3D人头化身重建方法,适用于多视角或单目视频输入下的个性化3D人头化身重建任务,通过融合纹理特征和构建表情特征基提升了模型的动画质量和视角泛化性,实现了快速训练、高效推理以及高质量的重建与动画。代码链接:https://doi.org/10.57760/sciencedb.j00240.00128。

关键词: 计算机图形学;三维重建;3D人头化身;3D高斯泼溅;纹理先验;表情特征基;表情动画

Animatable 3D Gaussian head avatars with texture prior

Zhou Tianyang[1,2], Mao Yuxiang[1,2], Ye Yongjing[1], Xia Shihong[1,2]

1. The Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China; 2. University of Chinese Academy of Sciences, Beijing 101408, China

Abstract: **Objective** Constructing vivid personalized 3D head avatar models at the lowest possible cost from 2D images is an important research problem in the fields of computer graphics and virtual reality. Although recently the 2D video generation models have achieved significant breakthroughs, explicit 3D avatars still play an irreplaceable role in fields such as virtual reality and human-computer interaction. Using 3D Gaussian Splatting (3DGS) as the rendering and 3D representation method can effectively improve the rendering quality and efficiency of the avatar models. However, the existing avatar modeling methods based on 3DGS either require several hours or even days of training time or have difficulty reconstructing fine

收稿日期:2026-01-15;**修回日期**:2026-03-31

*通信作者:夏时洪,通信作者,男,研究员,主要研究方向为计算机图形学、虚拟现实、人工智能。E-mail:xsh@ict.ac.cn;夏时洪 xsh@ict.ac.cn

基金项目:中国科学院计算技术研究所创新课题(E461020003);国家重点研发计划项目(2022YFF0902302);

Supported by: the Knowledge Innovation Program of the Institute of Computing Technology of the Chinese Academy of Sciences under Grant E461020003; National Key Research and Development Program of China under Grant 2022YFF0902302

wrinkle details, making it hard to achieve both fast and high-quality reconstruction simultaneously. In the term of the implementation details of these methods, during the training of the avatar models, the existing methods all rely on the mesh that tracked from the original data in the preprocessing stage to provide geometric information for the avatar model, but at the same time, the texture map corresponding to the mesh model is discarded as an additional output of the preprocessing stage. This leads to the avatar model having to learn the appearance information from scratch from the original image. In addition, the existing methods achieves the expression animation of avatar models through the linear combination of Gaussian attribute blendshapes. However, taking the rotation attribute of Gaussians as an example, the addition of quaternions is mathematically meaningless. This will lead to unreasonable rotation interpolation during the expression animation process, affecting the stability of the animation and the rationality of the final avatar model′s geometry. To solve these problems, we propose TPAvatar (Head Avatar with Texture Prior), a novel method to create photorealistic head avatars from the multi-view video sequences or monocular video sequence of the subject based on 3DGS, achieving real-time and high-fidelity animation. **Method** By learning a latent feature space for Gaussian attributes, TPAvatar achieves a significant reduction in neural network parameters, resulting in a compact neural network model. Specifically, TPAvatar is the first to leverage a pre-trained DINOv2 model to extract view-independent identity appearance prior from the texture map of the specific subject, construct a UV-aligned identity feature map, and provide improved initialization for the Gaussian model. For expression driving, TPAvatar establishes a set of implicit expression feature blendshapes for each Gaussians within the local space of each triangle of the mesh. By combining mesh binding with linear combinations of these expression feature blendshapes, it enables efficient and expressive animation of the avatar. The identity feature map and the expression feature map are first summed pixel-by-pixel, and then decoded by a Gaussian decoder to obtain Gaussian attribute maps defined in the UV-local coordinate. These Gaussian attributes are subsequently transformed into the global coordinate, and the final images are rendered using 3DGS. **Result** Experimental results on the multi-views dataset NeRSemble and the monocular dataset INSTA demonstrate that TPAvatar can effectively handle multi-view reconstruction and monocular reconstruction tasks. Comparing with existing methods such as GaussianAvatars, GEM, and RGBAvatar, TPAvatar achieves shorter training time, faster inference speed, and higher-quality reconstruction and animation, effectively balancing high fidelity and real-time performance. Specifically, we evaluate these methods on subjects of different ages and genders under two tasks: novel view synthesis on the validation set and novel expression synthesis on the test set. In the multi-view reconstruction scenario, compared with the baseline method GaussianAvatars/GEM, TPAvatar has shortened the reconstruction time from 8/12 hours to 1.5 hours while achieving higher reconstruction quality: on the test set, PSNR increased by 1.5608/10.3556, and LPIPS decreased by 0.0037/0.0131; compared with the baseline method RGBAvatar, TPAvatar significantly improved the generalization of new view synthesis while maintaining the advantage of fast reconstruction, with PSNR increasing by 0.5139 and LPIPS decreasing by 0.0155. In the monocular reconstruction scenario, compared with the existing SOTA baseline method RGBAvatar, TPAvatar′s PSNR increased by 0.1176 and LPIPS decreased by 0.0016. TPAvatar achieves an animation speed of 164 FPS, enabling real-time animation performance. Ablation studies further verify the effectiveness of both the identity feature module and the expression driving module. **Conclusion** By integrating texture features and constructing expression feature bases, TPAvatar enhances the animation quality and perspective generalization of the 3DGS head avatar model, making it a practical and efficient 3D avatar reconstruction method that capable of real-time rendering and animation. It suitable for personalized 3D head avatar reconstruction tasks under multi-view or monocular video input. The decoupled design of identity and expression endows the proposed method with the potential to be extended to single-image avatar reconstruction. Code is available in:https://doi.org/10.57760/sciencedb.j00240.00128

Key words: computer graphics; 3D reconstruction; 3D head avatar; 3D Gaussian splatting; texture prior; expression feature blendshapes; expression animation

0 引 言

虚拟化身(avatar)是运用数字技术创造出来的“数字人”(Gao 等,2024)。近年来化身建模技术已经被广泛地应用于电影、游戏等领域,具有重要的社会价值和研究意义(Hao 等,2024)。传统的3D建模依赖艺术家手工处理,导致高昂的时间和人工成本。因此,从摄像机拍摄的2D图像序列中端到端重建可动画的化身是目前3D化身建模领域主要的研究目标(Zhao 等,2024)。

3D 高斯泼溅(3D Gaussian Splatting, 3DGS)(Kerbl 等, 2023)是一种高效的3D场景表示方法,可以从2D图像中重建3D静态场景。利用3DGS构建可动画化身是目前化身建模研究的热点问题(Pan 等,2025)。然而,现有研究要么依赖参数量较大的深度神经网络模型,需要数小时乃至数十小时的时间训练表情驱动信号到高斯模型的映射(Xu 等, 2024; Saito 等, 2024; Aneja 等, 2025);要么只是将表情信号到模型的映射建模为简单的线性关系,难以重建精细的皱纹细节(Shao 等, 2024; Qian 等, 2024)。这些方法很难同时做到快速重建和高质量重建。

为了实现轻量化的高保真人头化身建模,本文提出了一种3D人头化身重建方法TPAvatar(Head Avatar with Texture Prior),可以从多视角或单目视频数据中快速重建高保真可动画的人头化身。图1的红色虚线框标识了TPAvatar与现有方法在设计思路上的两处主要不同。具体地,TPAvatar的设计基于对现有方法的两个核心洞察:

1)现有3D高斯人头重建方法基本都需要在预处理阶段从输入图像中跟踪得到人头网格模型,为后续高斯模型提供几何先验和动画的驱动信号。现有基于光度损失的跟踪方法(Qian 等, 2024)输出的网格模型包含每张输入图像的FLAME(Li 等, 2017)几何参数和对应的纹理图。然而,现有方法均将纹理图视作预处理阶段的多余输出丢弃不用,仅利用网格模型的几何参数作为高斯模型的输入,高斯的外观属性需要在训练过程中从头学习。事实上,网格模型的纹理图是一种良好的视角无关、姿态无关的表示,可以为UV空间中的高斯模型提供建模对象身份相关的外观先验信息。然而,现有方法主要聚焦于规范高斯模型的初始化方式和表情驱动模块的设计,忽略了预处理阶段纹理图中的有效信息。

2)在表情驱动模块的设计上,RGBAvatar(Li 等, 2025)、GBS(Ma 等, 2024)等方法采用混合形状(Blendshapes)作为化身模型驱动方式,以表情系数作为混合系数,将若干高斯属性基线性相加实现表情驱动。SynShot(Zielonka 等, 2025c),GHA(Xu 等, 2024)等工作为了解耦身份和表情,为每个高斯点计算身份分量和表情分量,最后简单地将两部分分量直接相加作为最终高斯点的属性。然而对于3DGS的旋转属性而言,尽管相加后的归一化函数保证了最终的结果是一个四元数,但四元数的加法本身在几何上是没有意义的。这将导致表情动画过程中可能产生不合理的旋转插值,影响动画的稳定性与模型面部几何的合理性。

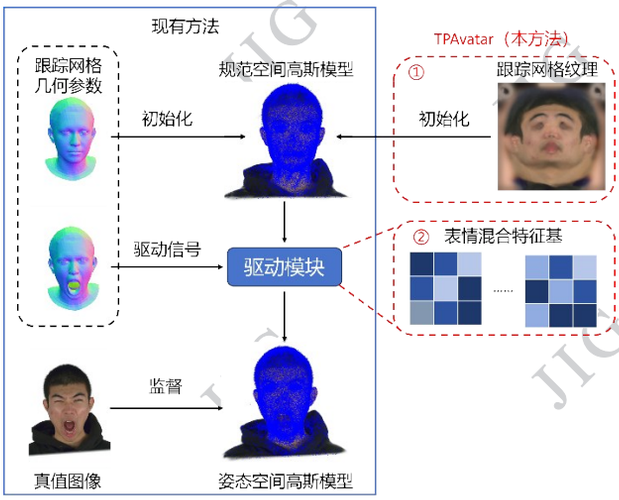


图1 TPAvatar与现有方法的思路对比

Fig. 1 Comparison between TPAvatar and existing methods.

针对第一个问题,TPAvatar首次提出将跟踪纹理图作为模型输入,使用预训练的DINOv2编码器(Oquab 等, 2024)将FLAME纹理图编码为身份相关的语义特征场,为高斯模型提供身份外观先验。实验表明,编码器可以有效感知纹理图中的身份与人脸部位语义信息。与从原始图像中编码身份外观特征相比,从纹理中提取特征可以将身份相关的外观特征从表情、相机视角、光照等因素中解耦。

针对第二个问题,TPAvatar提出使用隐式特征基的线性组合代替高斯属性基的线性组合,从训练数据中学习一个线性可插值的高维特征隐空间,利用特征基的线性组合实现了表情驱动,避免了对高

斯属性直接相加。对于每个高斯点,TPAvatar维护一个高斯特征向量。模型不是直接学习输入表情信号到高斯属性的映射,而是先学习输入信号到特征隐空间的映射。高维的特征隐空间为高斯属性的学习提供了良好的上下文。与GEM(Zielonka等,2025a)等方法相比,TPAvatar将高斯模型的几何和外观语义信息存储为一个显式的张量,神经网络模型无需隐式地根据低频位置信号从头编码学习高斯的语义特征,因此有效地减少了神经网络模型所需的参数量,提高了训练效率。

综上,本文的主要贡献如下:

1)TPAvatar是一种轻量化的3D高斯人头化身建模方法,用较短的训练时间取得了比基线方法更好的建模质量,并且支持超过150FPS的实时动画。

2)TPAvatar是第一个利用跟踪网络的纹理信息作为高斯化身模型输入的方法,从纹理中提取身份外观先验初始化高斯模型,有效提高了重建质量。

3)TPAvatar使用特征基的线性组合代替高斯属性的线性组合,避免了对高斯属性直接执行加法可能产生的问题,实现了有效、高效的表情驱动。

4)实验验证了方法在多视角和单目重建任务上的有效性。与基线方法GaussianAvatars,GEM相比,TPAvatar用更短的时间取得了更高的重建质量;与基线方法RGBAvatar相比,TPAvatar在保持快速重建优点的同时显著提升了多视角输入下的建模质量。

1 相关工作

1.1 可动画3D人头化身重建

现有研究从化身表示方法的角度可以分为基于参数化网格的方法、基于隐式神经辐射场(neural radiance fields,NeRF)(Mildenhall等,2022)的方法、基于3DGS的方法。

早期的化身建模研究(Blanz等,1999;Li等,2017;Feng等,2021)使用参数化3DMM网格表示3D化身。3DMM模型使用低维参数高效表示面部几何,与成熟的网格渲染管线一起确保了高效的动画和渲染。然而由于表示能力有限,网格模型难以直接从图像中重建皱纹等复杂的外观细节和精细的几何。但是网格模型可以使用低维参数高效编码人脸的形状和表情,因此后续研究多利用网格为模型提供几何先验,将网格建模作为预处理阶段的重要一步。

NeRF擅长建模复杂的外观细节和精细的几何,弥补了传统网格表示能力的不足,经典的工作包括AvatarMAV(Xu等,2023),INSTA(Zielonka等,2023b),NeRFBlendShape(Gao等,2022)等。但NeRF的缺点是训练和渲染的速度慢,对同一场景的训练常常需要数十小时甚至数天的时间。此外,与网格等显式3D表征方法相比,NeRF并不擅长建模动态场景:由于构建动画模型时需要进行病态的逆蒙皮映射,模型在动画过程中容易出现模糊或伪影。

与NeRF相比,3DGS更擅长建模动态场景,在实现高质量实时渲染的同时极大地提高了渲染速度。GaussianAvatars(Qian等,2024),GHA(Xu等,2024)等工作从单目或多视角同步的图像序列数据中重建3D高斯人头化身,实现了高保真的化身重建与驱动;与AvatarMAV等基于NeRF的方法相比,提高了模型的保真度和表情泛化性,降低了训练和推理成本。

1.2 基于3DGS的可动画人头化身重建方法

3D高斯人头化身建模问题的核心是学习一个中性表情对应的规范人头高斯模型和一个表情驱动模块。动画过程中,驱动模块需要根据表情信号预测每个高斯点相对规范模型的偏移量,从而对规范模型"变形"。现有研究的驱动模块建模方法可以分为四类:基于网格绑定的方法(Shao等,2024;Qian等,2024)、基于多层感知器(multilayer perceptron,MLP)的方法(Xu等,2024;Chen等,2024;Aneja等,2025)、基于卷积神经网络(convolutional neural networks,CNN)的方法(Teotia等,2024;Giebenhain等,2024;Saito等,2024)、基于Transformer的方法(Wu等,2025)。其中,基于网格绑定的方法无需神经网络模型,动画渲染速度快,但是对于动态表情皱纹的建模质量差,对于新表情新视角的泛化能力不足。GaussianHeads(Teotia等,2024),RGCA(Saito等,2024)等工作使用深度神经网络架构(如StyleU-net)预测动画过程中高斯点的属性,提升了模型的细节建模能力,但同时需要更长的训练和动画推理时间。

1.3 3D高斯人头化身的轻量化表示

构建一个轻量级的化身模型,同时做到高质量建模和高效训练推理,是现有方法需要解决的重要

问题。为此，一些工作探索了3D高斯人头化身模型的轻量化表示。GEM(Zielonka等，2025a)提出了一种高斯化身模型的蒸馏方法，利用主成分分析将预训练的高斯模型蒸馏为一个线性的小模型。然而，蒸馏的过程不可避免地造成了模型重建质量的损失。此外，GEM在蒸馏之前仍需要构建建模对象的高斯模型，因此只是降低了模型的部署和推理成本，并没有实际降低模型的训练成本。TGA(Wang等，2025)利用三平面隐式存储高斯的颜色，减少了模型的内存开销。但是方法在动画过程中仍主要依赖网格驱动，难以建模人脸精细的几何细节。RGBAvatar(Li等，2025)学习一组高斯属性基，通过属性基的线性组合实现表情动画，可以实现对流式视频输入的实时重建。然而由于属性基表征能力有限，RGBAvatar对于新视角新表情的泛化能力不足。

此外，GAGAvatar(Chu等，2024)等工作从大量野外图像数据中训练了一个跨身份的重建模型，实现了从单张图像中直接前馈重建3D高斯人头化身模型。但是由于数据的匮乏、先验的缺失，目前这类方法的重建质量仍难以与使用高质量多视角数据重建的方法相提并论。

2 相关技术基础

2.1 3D高斯泼溅(3DGS)

3DGS(Kerbl等，2023)可以实现3D场景的高质量建模与实时渲染。一个表示静态场景的高斯模型$\boldsymbol{G}$是一个若干高斯点组成的集合，每个高斯点$\boldsymbol{g}$有五个属性：位置$\boldsymbol{x}\in\mathbf{R}^3$，颜色$\boldsymbol{c}\in\mathbf{R}^3$，不透明度$o\in\mathbf{R}$，缩放$\boldsymbol{s}\in\mathbf{R}^3$和四元数形式的旋转向量$\boldsymbol{r}\in\mathbf{R}^4$。由于篇幅有限，具体的渲染算法请参阅Kerbl等人(2023)的原始论文。

2.2 3D人头化身重建问题的形式化建模

人头化身重建问题可以抽象为公式(1)：

$$\theta^*=\arg\max_{\theta}\mathbb{E}_{\left(\boldsymbol{G},\boldsymbol{f}_{\mathrm{id}},\boldsymbol{f}_{\mathrm{exp}}\right)\sim p_{\mathrm{data}}}\left[\log \mathrm{P}_{\theta}\left(\boldsymbol{G}|\boldsymbol{f}_{\mathrm{id}},\boldsymbol{f}_{\mathrm{exp}}\right)\right]\#(1)$$

式中θ^*是学习到的最优模型参数，p_{data}是训练数据的分布，模型参数θ将身份特征$\boldsymbol{f}_{\mathrm{id}}$和表情特征$\boldsymbol{f}_{\mathrm{exp}}$映射为高斯模型$\boldsymbol{G}$。对于个性化化身建模问题，身份特征$\boldsymbol{f}_{\mathrm{id}}$可以视作固定的先验，因此有：

$$\theta^*=\arg\max_{\theta}\mathbb{E}\left[\log \mathrm{P}_{\theta}\left(\boldsymbol{f}_{\mathrm{exp}}|\boldsymbol{G};\boldsymbol{f}_{\mathrm{id}}\right)+\log \mathrm{P}_{\theta}\left(\boldsymbol{G}|\boldsymbol{f}_{\mathrm{id}}\right)\right]\#(2)$$

所以化身重建问题本质上是一个在身份先验约束下的最大后验概率推断问题：模型θ需要寻找既能解释给定表情、又符合身份先验的高斯表示$\boldsymbol{G}$，实现稳定的表情动画。其中$\mathrm{P}_{\theta}\left(\boldsymbol{f}_{\mathrm{exp}}|\boldsymbol{G};\boldsymbol{f}_{\mathrm{id}}\right)$约束模型的表情一致性，$\mathrm{P}_{\theta}\left(\boldsymbol{G}|\boldsymbol{f}_{\mathrm{id}}\right)$约束模型的身份一致性。

3 方 法

3.1 方法概览

图2展示了TPAvatar的流程图。TPAvatar以建模人物的多视角或单目图像序列数据作为输入。TPAvatar使用预训练的编码器从网格纹理图中编码得到一个身份特征图。为了实现表情驱动，TPAvatar维护了一组可学习的表情混合特征基，利用MLP学习从FLAME表情系数到特征混合系数的映射，以线性组合的方式得到表情特征图。身份特征图和表情特征图分别编码了每个高斯点表情无关和表情相关的特征，二者相加后得到最终的高斯特征图，并由高斯解码器解码得到UV局部坐标系下的高斯属性图。最后通过切线-副切线-法线(tangent-bitangent-normal，TBN)空间变换将高斯点转换到全局坐标系，并使用3DGS渲染高斯模型得到最终图像。

3.2 预处理与高斯模型表示

在预处理阶段，遵循GaussianAvatars等现有方法，TPAvatar从输入数据中跟踪得到目标建模人物的网格模型，包括FLAME几何参数(形状系数$\boldsymbol{\beta}$、表情系数$\boldsymbol{\psi}$和姿态系数$\boldsymbol{\theta}$)和FLAME纹理图$\boldsymbol{T}$。其中$\boldsymbol{\beta}\in\mathbf{R}^{100}$，$\boldsymbol{\psi}\in\mathbf{R}^{300}$，$\boldsymbol{\theta}\in\mathbf{R}^{36}$，$\boldsymbol{T}\in\mathbf{R}^{3\times \mathrm{H}_T\times \mathrm{W}_T}$，$\mathrm{H}_T$和$\mathrm{W}_T$分别表示纹理图的高和宽。

为了更好地利用预处理阶段得到的FLAME纹理，与FlashAvatar(Xiang等，2024)、RGCA(Saito等，2024)等工作相同，TPAvatar采用UV对齐的高斯点采样策略：假设高斯点均匀分布在纹理空间，每个纹素的中心对应一个高斯点，根据高斯点对应的UV坐标可以确定高斯点在3D空间中的初始位置。高斯点的数量N在训练和推理过程中保持不变。这种做法的好处是：1)将离散的高斯点转换为结构化的表示，便于使用卷积等神经网络架构处理高斯点特征的映射。2)将每个高斯点与一个网格三角形绑定，网格为高斯模型的位置等几何属性提供几何先验。在高斯模型初始化时，FLAME的拓扑保证了高斯点位置的合理性；在表情动画时，可以利用网格三

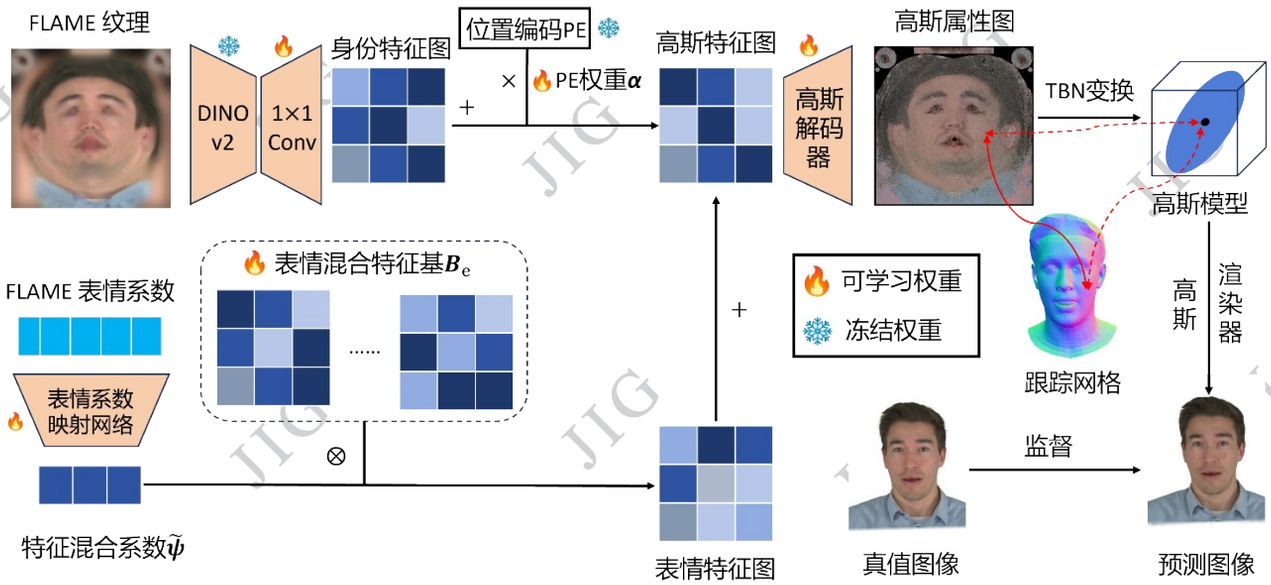


图2　TPAvatar的算法流程图

Fig. 2　Pipeline of TPAvatar

角形与对应高斯点的相对位置关系，对高斯点执行粗粒度的几何形变，模型只需要学习预测细粒度的高斯属性残差量，有效降低了模型预测形变的复杂度。3）高斯点和纹理图都处于同一规范的UV坐标表示下，有利于编码器将纹理图中身份相关的外观先验信息直接赋予对应位置的高斯点，特征提取和采样过程无需投影和跨模态的特征融合。

在这种表示下，整个高斯模型可以表示为一个属性图$\boldsymbol{G}$，$\boldsymbol{G}\in\mathbf{R}^{59\times H_T\times W_T}$。属性图中每个像素对应一个高斯点$\boldsymbol{g}$，$\boldsymbol{g}=\{\boldsymbol{x},\boldsymbol{r},\boldsymbol{s},o,\boldsymbol{c}\}$。此外，每个高斯点还拥有一个额外的特征向量$\boldsymbol{f}$，编码高斯点在网格局部空间中的几何和外观特征。模型先学习特征向量$\boldsymbol{f}$，再通过一个轻量级的解码器将特征向量$\boldsymbol{f}$映射为高斯点的5个属性。

对于特征向量$\boldsymbol{f}$的学习，TPAvatar将特征向量$\boldsymbol{f}$解耦为身份特征和表情特征，分别对应身份特征图$\boldsymbol{F}_{\mathrm{ID}}$和表情特征图$\boldsymbol{F}_{\mathrm{E}}$，其中$\boldsymbol{F}_{\mathrm{ID}},\boldsymbol{F}_{\mathrm{E}}\in\mathbf{R}^{|f|\times H_T\times W_T}$，$|f|$是特征向量$\boldsymbol{f}$的维度。FLAME纹理图$\boldsymbol{T}$、身份特征图$\boldsymbol{F}_{\mathrm{ID}}$、表情特征图$\boldsymbol{F}_{\mathrm{E}}$和最终解码得到的高斯属性图$\boldsymbol{G}$是空间对齐的，都定义在统一的FLAME的UV拓扑上，可以通过逐像素相加实现特征融合。

3.3　身份特征图的构建

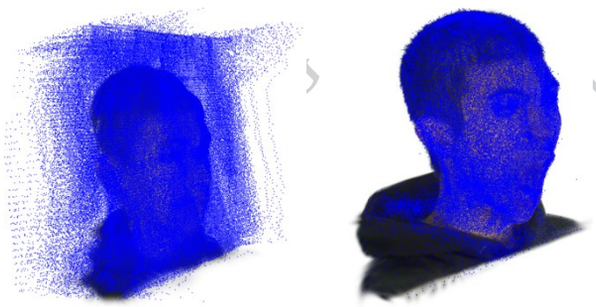
身份特征图的计算过程可以归纳为公式（3）：

$$\boldsymbol{F}_{\mathrm{ID}}=\mathrm{D}_{\mathrm{ID}}\left(\mathrm{E}_{\mathrm{ID}}(\boldsymbol{T})\right)+\boldsymbol{\alpha}\cdot PE \#(3)$$

式中，$\boldsymbol{F}_{\mathrm{ID}}$表示身份特征图，$\boldsymbol{T}$表示纹理图，$\mathrm{E}_{\mathrm{ID}}$表示预训练的DINOv2编码器，$\mathrm{D}_{\mathrm{ID}}$表示DINOv2特征上采样器，$PE$表示位置编码，$\boldsymbol{\alpha}$表示门控系数向量。

与现有的方法不同，为了充分利用网格模型的纹理信息，TPAvatar使用预训练的通用视觉特征提取器DINOv2（Oquab等，2024）从网格模型的纹理图中提取语义特征。需要说明的是，一些从单张图像中重建3D化身的工作也利用了DINOv2的特征来初始化高斯模型，但是在这些方法中DINOv2编码器的输入都是原始图像。例如，GAGAvatar（Chu等，2024）设定DINOv2特征图的每一个像素对应一个高斯点，利用特征图预测深度，将高斯从2D像素平面提升至3D空间。这导致大量高斯点的浪费（如图3（a）所示）。LAM（He等，2025）将高斯点绑定在FLAME顶点上，利用交叉注意力机制捕获DINOv2提取的图像特征。这种做法引入复杂的深度神经网络来实现高斯特征与图像语义特征的融合，进而导致庞大的计算量和昂贵的训练开销。GUAVA（Zhang等，2025）将DINOv2的特征图上采样到与原始图像相同的大小，将UV空间中的高斯点投影到特征图，通过双线性插值计算高斯点的特征。这意味着在训练过程中需要对每张原始图像都执行DINOv2编码，因此不适合多视角重建任务。此外，由于原始输入图像是相机视角相关的，模型对于原始图像中轮廓、头发、自遮挡区域特征的提取可能是

不稳定的。在特征采样过程中，遮挡、边界混叠和投影误差都会影响采样特征语义的稳定性。综上，这些方法要么导致严重的高斯点冗余，要么需要昂贵的计算将像素空间的语义特征融合到UV空间的高斯点，要么特征的提取和采样过程受到原始图像相机视角、光照、背景等因素的影响，存在一定程度的噪声干扰。



（a） GAGAvatar （b） TPAvatar（本方法）

（(a) GAGAvatar; (b) TPAvatar(Ours)）

图3 两种方法的高斯点位置的可视化结果

Fig. 3 Visualization of Gaussian positions of the two methods

与上述做法不同，TPAvatar直接从FLAME纹理图中提取语义特征，特征提取和采样都在UV空间中进行。高斯点在UV空间中是几何对齐的，保证了高斯点位置分布的合理性，而且不需要通过投影执行特征采样。高斯点的特征是视角无关的，在一个视角无关的表示下提取高斯点的特征是符合直觉的。VHAP跟踪器（Qian等，2024）输出的高质量纹理图使得特征提取器不受光照、遮挡等因素的影响。因为纹理是视角无关、姿态无关、背景无关的，只需要在训练开始前对训练数据中任意一张中性表情图像对应的FLAME纹理图执行一次DINOv2编码，无需对所有原始图像的纹理图重复执行。第4节的实验表明，DINOv2强大的跨域泛化能力使得编码器对于纹理图像同样具有良好的泛化能力，可以从中有效提取语义特征。综上，直接从FLAME纹理图采样特征是有效高效的，是有利于UV对齐的高斯表示的。

DINOv2编码器E_{ID}输出的特征是一组词元（token），需要通过一个上采样器D_{ID}映射为与纹理图分辨率相同、与高斯特征维度$|f|$相同的特征图。D_{ID}包含上采样和维度映射两步操作，其中上采样通过双线性插值实现，维度映射通过1×1卷积实现。

由于纹理特征是平滑的，为了增强D_{ID}的空间感知，TPAvatar为每一个高斯身份特征额外加上一个高斯点初始位置编码PE，编码维度与高斯点特征维度相同。另外模型需要额外学习一个门控系数向量α作为权重调节位置编码注入信号的强弱。

3.4 表情特征图的构建

为了建模FLAME表情参数ψ到表情特征图F_E的映射，朴素的做法是学习一个特征映射矩阵W_E，如公式（4）所示。

$$F_E = W_E\psi \#(4)$$

式中$W_E \in \mathbf{R}^{|f|\times H_T\times W_T\times 100}$代表特征映射矩阵。由于$W_E$的参数量太大，在有限的数据下模型很难学好这个高维的参数空间，容易过拟合，因此合理的做法是对W_E执行低秩分解：

$$W_E = W'_E W_M \#(5)$$

式中W'_E和W_M代表两个秩不大于m的低秩矩阵，m是超参数，其中$W'_E \in \mathbf{R}^{|f|\times H_T\times W_T\times m}$，$W_M \in \mathbf{R}^{m\times 100}$，$m \ll 100$。

TPAvatar通过学习一个结构为浅层MLP的表情系数映射网络M_e实现映射关系W_M。M_e将FLAME表情系数ψ从100维映射到m维，如公式（6）所示。在实践中，m的取值可以在10到30之间。

$$\tilde{\psi} = M_e(\psi) \#(6)$$

式中$\tilde{\psi}$代表m维的表情混合系数向量。

对于映射关系W'_E，TPAvatar通过维护m个表情特征基实现。表情特征基记作B_e：

$$B_e = \left[B_e^{(1)}, B_e^{(2)}, \ldots, B_e^{(m)}\right] \#(7)$$

式中$B_e^{(i)} \in \mathbf{R}^{|f|\times H_T\times W_T}$代表表情基分量。以$\tilde{\psi}$作为混合权重系数向量，通过表情特征基的线性组合得到表情特征图F_E：

$$F_E = \tilde{\psi} \otimes B_e = \tilde{\psi}^{(1)}\cdot B_e^{(1)} + \cdots + \tilde{\psi}^{(m)}\cdot B_e^{(m)} \#(8)$$

式中$\otimes$代表表情系数向量$\tilde{\psi}$和表情基B_e的点乘运算。

第4节的实验结果表明，因为表情特征图F_E建模的是高斯点在网格三角形定义的局部坐标系下的表情相关的高斯属性残差量，只需要少量特征基组成的低秩表示即可有效实现对表情特征的建模。

3.5 解码和渲染

在得到身份特征图F_{ID}和表情特征图F_E之后，将两者相加得到最终的特征图F_G：

$$F_G = F_{ID} + F_E \#(9)$$

然后将$\boldsymbol{F}_{\mathrm{c}}$送入解码器$\mathrm{D_f}$得到高斯属性图$\boldsymbol{G}$：

$$\boldsymbol{G}=\mathrm{D_f}(\boldsymbol{F}_{\mathrm{c}})\#(10)$$

解码器$\mathrm{D_f}$由一个轻量级的MLP实现，将每个高斯点对应的特征$\boldsymbol{f}$映射为高斯的5个属性。需要说明的是，解码器预测的高斯属性是网格三角形局部切线空间中的属性。所以在执行3DGS渲染前，需要通过TBN变换将局部空间中的高斯点映射到全局坐标系下，具体过程请参考(Li等，2025)。最后将预测的高斯点集合送入3DGS渲染器R，得到相机视角π下对应的预测图像$\boldsymbol{I}_r$。

3.6 训练与实现细节

1)损失函数的设计。为了加速训练，TPAvatar仅使用L1和D-SSIM作为损失项计算损失函数：

$$L=\lambda_1 L_1+\lambda_2 L_{\mathrm{D-SSIM}}\#(11)$$

式中L_1和$L_{\mathrm{D-SSIM}}$分别代表预测图像$\boldsymbol{I}_r$与真实图像$\boldsymbol{I}_{gt}$之间的L1和D-SSIM损失，λ_1和λ_2是权重系数，取值遵循原始的3DGS实现，$\lambda_1=0.8$，$\lambda_2=0.2$。实验表明，在不使用感知损失和其他正则项的情况下，TPAvatar同样可以实现高质量的重建与动画。

2)实现细节。纹理图高宽H_T，W_T均为300。每个高斯点的特征$\boldsymbol{f}$维度的大小为32。计算位置编码PE时先使用4阶傅里叶编码(Mildenhall等，2022)得到27维向量，再通过一个线性层W_{pos}将其映射为与特征$\boldsymbol{f}$相同的32维。位置编码权重系数$\boldsymbol{\alpha}$同样是一个32维的向量，每个分量初始值为0.001。表情系数映射网络$\mathrm{M_e}$是一个轻量级的MLP，包含两个128维的线性层，使用ReLU作为激活函数。高斯属性解码器$\mathrm{D_f}$的骨干部分由2个64维线性层组成，使用Leaky ReLU作为激活函数。$\mathrm{D_f}$有6个解码头，分别预测高斯的位置、旋转、缩放、不透明度、0阶球谐系数和高阶球谐系数。除高阶球谐系数以外，每个解码头都包含一个32维的中间层。为了训练的稳定性，旋转属性预测的是相对单位四元数的偏移。

3)训练细节。整个训练过程在一张RTX 3090 GPU上进行。训练图片使用随机颜色背景实现数据增强。模型使用gsplat渲染器(Ye等，2025)实现3DGS模型场景和视角的并行渲染，使用Adam优化器优化所有可学习参数。其中表情混合特征基$\boldsymbol{B}_{\mathrm{e}}$和位置编码权重系数向量$\boldsymbol{\alpha}$的学习率是0.00125；DINOv2特征上采样器$\mathrm{D_{ID}}$，位置编码投影层W_{pos}，表情系数映射网络$\mathrm{M_e}$，高斯属性解码器$\mathrm{D_f}$的学习率均为0.001。对于多视角重建，训练共进行30万步，其中表情驱动模块从2万步之后开始训练，大约需要1.5小时。训练批量大小为12，包括3个场景和4个视角。对于单目重建，训练共进行5万步，表情驱动模块从3千步之后开始训练，大约需要5~8分钟。训练批量大小为10。

4 实　验

4.1 实验设置

实验选用多视角同步的人头图像序列数据集NeRSemble(Kirschstein等，2023)和单目人像视频数据集INSTA(Zielonka等，2023)作为实验数据集。两个数据集都遵循GaussianAvatars(Qian等，2024)的数据划分方法：对于NeRSemble数据集，对每个身份随机取一个表情序列作为测试集。在剩余9个序列中，取8号相机的所有图片组成验证集，其余15个相机视角的图片作为训练集。对于INSTA数据集，取每个单目视频前70%的图片作为训练集，后30%的图片作为测试集，不设置验证集。综上，训练集和验证集对应相同的表情序列，只是相机视角不同，主要验证模型对训练集表情序列的拟合效果和对新视角的泛化能力。测试集则主要测试模型对新表情驱动信号的泛化能力。

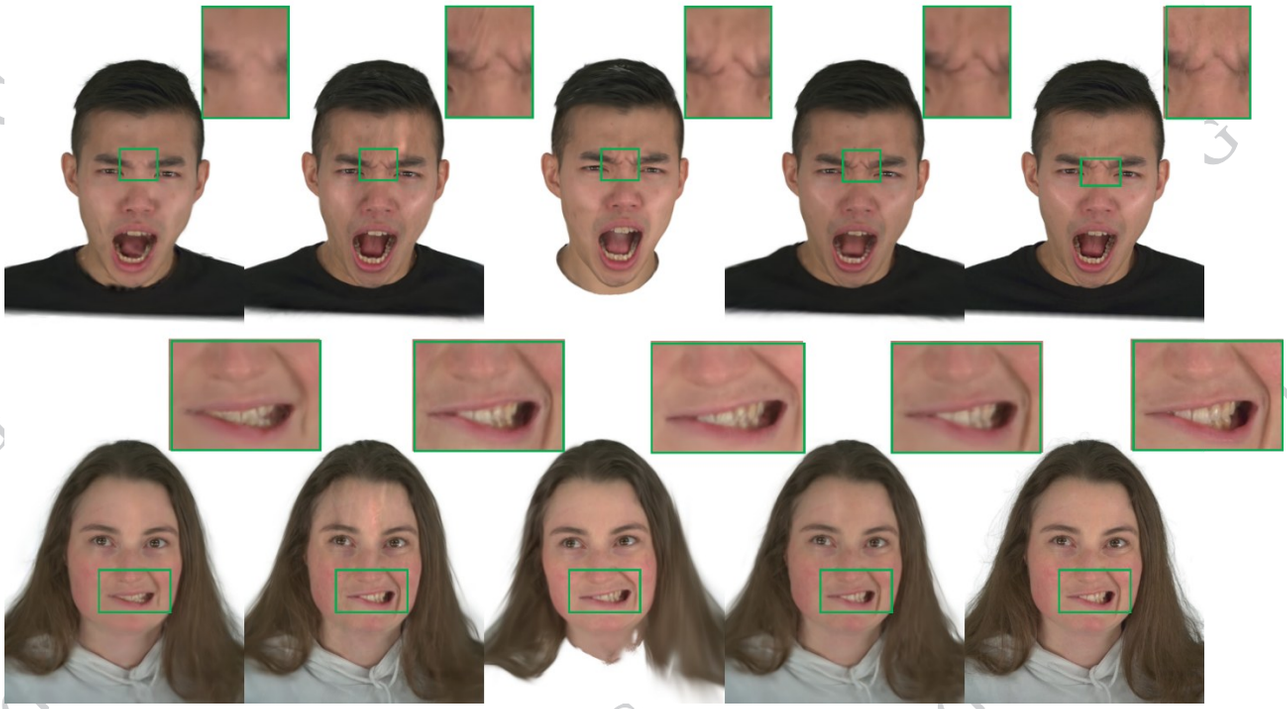
衡量重建质量的实验指标包括：峰值信噪比(peak signal to noise ratio，PSNR)，结构相似性(structural similarity，SSIM)，感知损失(learned perceptual image patch similarity，LPIPS)和L1损失。

4.2 对比实验

对比实验选择GaussianAvatars(Qian等，2024)，RGBAvatar(Li等，2025)，GEM(Zielonka等，2025a)作为基线方法。实验包含新视角合成(验证集)和新表情驱动(测试集)两项任务。

三种基线方法均采用官方开源代码进行实验，均可以从单目或多视角数据中重建可实时动画的3D人头化身模型。需要额外说明的是，GEM只能实现头部和脖子的重建。为了比较的公平性，所有方法在计算定量指标时去除肩膀区域，只计算头部和脖子区域。

从图4至图5展示的定性实验结果可以看到，GaussianAvatars的渲染结果趋于平滑，缺少细粒度的表情细节。这是因为GaussianAvatars的高斯点动



(a) GaussianAvatars (b) RGBAvatar (c) GEM (d) TPAvatar(本方法) (e)真值图像GEM无法重建化身的肩膀区域,因此结果中只有头部和脖子区域。
((a)GaussianAvatars;(b) RGBAvatar;(c) GEM;(d)TPAvatar(Ours);(e)Ground Truth)

图4　在NeRSemble验证集(新视角合成任务)上的部分定性实验结果。

Fig. 4　Qualitative comparisons of different methods on the NeRSemble validation set(novel-view synthesis task).

画完全由网格驱动,难以建模细粒度表情形变。对于多视角建模的场景,RGBAvatar在训练集没有的8号相机视角渲染时会出现明显的伪影,说明方法对于新视角的泛化能力较差。GEM利用主成分分析(PCA)构建表情特征基,导致外观和表情细节的丢失:如图5所示,GEM重建的表情与原图表情存在明显的差异。TPAvatar可以准确重建训练集中的表情,对于测试集中没有见过的新表情也具有良好的泛化能力。

从表1至表3展示的定量实验结果可以看到,无论是多目重建任务还是单目重建任务,无论新视角合成还是新表情驱动任务,TPAvatar在各个指标上均取得了最优的结果。

与RGBAvatar相比,TPAvatar不但可以做到高质量的单目重建,而且显著提高了多视角重建的质量。GaussianAvatars允许高斯点在训练时执行分裂、克隆等自适应密度控制策略,因此能够较好地拟合训练集图片中的表情细节,在验证集上取得了较好的结果。但是模型存在过拟合训练集表情序列的问题,对于新表情的泛化能力不足。TPAvatar和另外两个基线方法都采用了UV对齐的高斯点采样策略,模型的高斯点数量在训练过程中保持不变。这三种方法都在网格驱动高斯移动变形的基础上额外学习细粒度的高斯偏移量,实现更精细的动画控制。GEM基于统计学习构建的高斯特征基保留了原始高斯属性图中的关键信息,有效去除噪声干扰,但是主成分分析在滤除噪声的同时丢失了许多表情细节信息,损害了模型对头部姿态的拟合效果,导致较差的定量实验结果。

表4展示了四种方法的训练与推理成本。尽管RGBAvatar是四种方法中训练和推理速度最快的,但是对于新视角的泛化能力则有所欠缺;而TPAvatar在有限的代价下超越了RGBAvatar的重建动画效果,同时实现了高帧率和高保真的动画。

4.3　消融实验

4.3.1　身份特征构建模块

为了证明TPAvatar身份编码器设计的有效性,本小节对比了三种设计方案:(a)使用位置编码初始化身份特征图,训练过程中通过优化的方式学习;(b)使用DINOv2提取的纹理特征作为身份特征图;

（a） GaussianAvatars （b） RGBAvatar （c） GEM （d） TPAvatar（本方法） （e） 真值图像

GEM无法重建化身的肩膀区域，因此结果中只有头部和脖子区域。
((a)GaussianAvatars；(b) RGBAvatar；(c) GEM；(d)TPAvatar(Ours)；(e)Ground Truth)

图5 在NeRSemble和INSTA测试集（新表情合成任务）上的部分定性实验结果。

Fig. 5 Qualitative comparisons of different methods on the NeRSemble and INSTA test set (novel-expression synthesis task).

(c)在(b)的基础上额外增加位置编码增强空间感知(TPAvatar采用的方案)。

从图6和表5的结果可以看到，引入DINOv2提取的纹理外观先验可以有效提高重建质量；引入位置编码可以有效增强解码器的空间感知，进一步提高模型对新表情的泛化能力

4.3.2 表情驱动模块

为了证明采用表情特征基的线性组合作为表情驱动方法的有效性，本小节验证了使用MLP代替表情特征基执行动画的方案，将高斯特征向量、位置编码和表情向量三个向量拼接后作为MLP的输入，输出对每个高斯点5个属性的预测。对于高斯特征向量，本小节分别测试了使用位置编码初始化和使用

表1　NeRSemble验证集(新视角合成任务)上的定量结果
Table 1　Quantitative comparison on the NeRSemble validation dataset (Novel viewpoint synthesis task)

	PSNR↑	SSIM↑	LPIPS↓	L1↓
GaussianAvatars	36.0578	0.9772	0.0577	0.0065
GEM	24.6180	0.9146	0.1042	0.0165
RGBAvatar	35.6082	0.9724	0.0680	0.0067
TPAvatar(本方法)	**36.1221**	**0.9779**	**0.0525**	**0.0061**

注:本节所有定量实验表格均用黑色加粗字体表示该指标下的最优结果,下划线表示次优结果;↑表示该指标值越高越高,↓表示值越低越好;后面不再重复说明。

表2　NeRSemble测试集(新表情合成任务)上的定量结果
Table 2　Quantitative comparison on the NeRSemble test dataset (Novel expression synthesis task)

	PSNR↑	SSIM↑	LPIPS↓	L1↓
GaussianAvatars	32.4721	0.9598	0.0727	0.0089
GEM	23.6773	0.9106	0.1158	0.0182
RGBAvatar	33.0322	0.9610	0.0821	0.0077
TPAvatar(本方法)	**34.0329**	**0.9668**	**0.0690**	**0.0074**

表3　INSTA测试集(新表情合成任务)上的定量结果
Table 3　Quantitative comparison on the INSTA test dataset (Novel expression synthesis task)

	PSNR↑	SSIM↑	LPIPS↓	L1↓
GaussianAvatars	33.0254	0.9613	0.0488	0.0075
GEM	24.7057	0.8809	0.0982	0.0191
RGBAvatar	34.1444	0.9692	0.0409	0.0064
TPAvatar(本方法)	**34.2620**	**0.9698**	**0.0393**	**0.0062**

表4　NeRSemble数据集上四种方法的训练与推理成本
Table 4　The training and inference costs of the four methods on the NeRSemble dataset

	平均训练时间↓	推理速度↑
GaussianAvatars	8小时	90FPS
GEM	12小时	118FPS
RGBAvatar	**1小时**	**335FPS**
TPAvatar(本方法)	1.5小时	164FPS

注:所有结果均是在同一块RTX3090 GPU上测得的。

DINOv2初始化两种方案。除输入层以外,MLP的整体结构与高斯属性解码器D_f完全相同。定量实验结果如表6所示。定性实验结果如图7所示。

从实验结果可以看到,两种使用MLP的方案重建质量都发生了严重的退化。原因是模型的学习目标是预测局部网格空间中的细粒度形变。这一形变是低秩的,是与高斯点上下文高度相关的,因此使用局部可学习向量的线性组合优于使用以全局表情向量和位置编码为引导条件的MLP网络。当然,使用MLP的方案在测试集上某些指标的结果优于特征基的线性组合方案。这主要是因为MLP的归纳偏置引导输出结果趋于平滑,有利于PSNR、L1等像素级指标的计算。但是对于更能反映人类视觉感知相似度的感知指标LPIPS,使用线性组合方案的结果仍然显著优于使用MLP的结果。值得注意的是,在使用MLP作为驱动方式的情况下,使用DINOv2代替位置编码反而进一步造成了重建质量的退化。这可能是由于解码器网络在学习过程中过早地依赖DINOv2特征,抑制了网络对于空间位置信号和全局表情信号的感知。综上,实验结果充分说明了使用特征基线性组合作为表情驱动方法的合理性。

最后图8展示了部分表情特征基的作用效果。与FLAME等网格统计模型不同,TPAvatar建模的是每个高斯点在网格三角形局部坐标系中的细粒度形变,是高斯点相对于网格模型的偏移和残差。为了更清晰展示面部微表情的变化,图8使用PSNR热力图展示激活不同特征基后的人脸与原始“中性人脸”之间的差异。可以看到不同的表情基可以控制不同的人脸部位产生细粒度的表情形变。这表明特征基与部位形变、表情语义是对齐的,实现了面部表情的解耦建模。

5　结　论

1)总结。TPAvatar是一种新的低成本高保真的人头化身建模方法,可以从多视角数据或单目数据中快速重建高斯模型。TPAvatar是第一个利用网格纹理图初始化高斯化身模型的方法,有效提升了化身的重建与动画质量,为化身重建的研究工作引入了新思路。TPAvatar构建了一组可学习的表情特征基,通过线性组合的方式实现了有效的表情动画。与现有方法相比,TPAvatar同时做到了高效建模、实

表5　身份特征构建模块消融实验结果

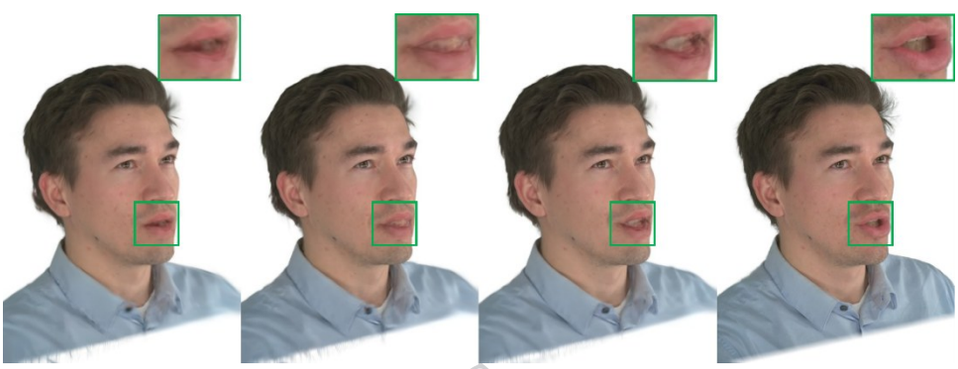
Table 5　Ablation results of the Identity Feature Module

数据	方法	PSNR ↑	SSIM ↑	LPIPS ↓	L1 ↓
验证集	位置编码(PE)	32.4188	0.9587	0.1211	0.0115
	DINOv2	32.9274	0.9644	0.1039	0.0108
	DINOv2+PE (本方法)	**33.2555**	**0.9653**	**0.0998**	**0.0103**
测试集	位置编码(PE)	**25.7081**	0.9245	0.1447	**0.0165**
	DINOv2	25.6105	**0.9247**	0.1359	**0.0165**
	DINOv2+PE (本方法)	25.4306	0.9234	**0.1336**	0.0168

表6　表情驱动模块消融实验结果

Table 6　Ablation results of the Expression driving Module

数据	方法	PSNR ↑	SSIM ↑	LPIPS ↓	L1 ↓
验证集	MLP+PE	31.9045	0.9508	0.1411	0.0106
	MLP+DINOv2	29.8322	0.9393	0.1835	0.0128
	表情特征基混合(本方法)	**33.2555**	**0.9653**	**0.0998**	**0.0103**
测试集	MLP+PE	**25.8048**	0.9258	0.1539	**0.0165**
	MLP+DINOv2	25.7808	**0.9260**	0.1913	0.0175
	表情特征基混合(本方法)	25.4306	0.9234	**0.1336**	0.0168



(a) 位置编码(PE) (b) DINOv2 (c) DINOv2+PE (本方法) (d)真值图像

((a)Positional Encoding(PE);(b) DINOv2;(c) DINOv2+ PE(Ours);(d)Ground Truth)

图6　身份特征构建模块的消融实验结果(第一行是验证集上的结果,第二行是测试集上的结果)

Fig. 6　Ablation study of the Identity Feature Module(First row for validation set, the second row for test set)

时动画和高保真重建三项人头化身建模研究的核心目标。

2)对于小样本重建场景的讨论。虽然TPAvatar是一种基于优化的逐实例重建方法,主要针对多视角重建场景设计,但方法可以利用合成数据实现小样本场景(少量甚至单张输入图像)下的化身重建:借助生成模型生成不同视角下目标人物不同表情的合成图像,构建合成数据集,将问题转换为多视角重建问题。因此TPAvatar与CAP4D(Taubner等,2024),MVP4D(Taubner等,2025)等工作是完全正交的,可以在它们生成的合成数据基础上完成小样本重建任务。所以TPAvatar不仅适用于标准的多视角重建任务,也可以扩展至更一般的场景,具有良好的前景和潜力。

(a) MLP+PE (b) MLP+DINOv2 (c) 表情特征基混合(本方法) (d)真值图像
((a) MLP+PE; (b) MLP+DINOv2; (c) Expression Feature Blendshapes(Ours); (d)Ground Truth)

图7　表情驱动模块的消融实验结果(第一行是验证集上的结果,第二行是测试集上的结果)

Fig. 7　Ablation study of the Expression driving Module(First row for validation set, the second row for test set)

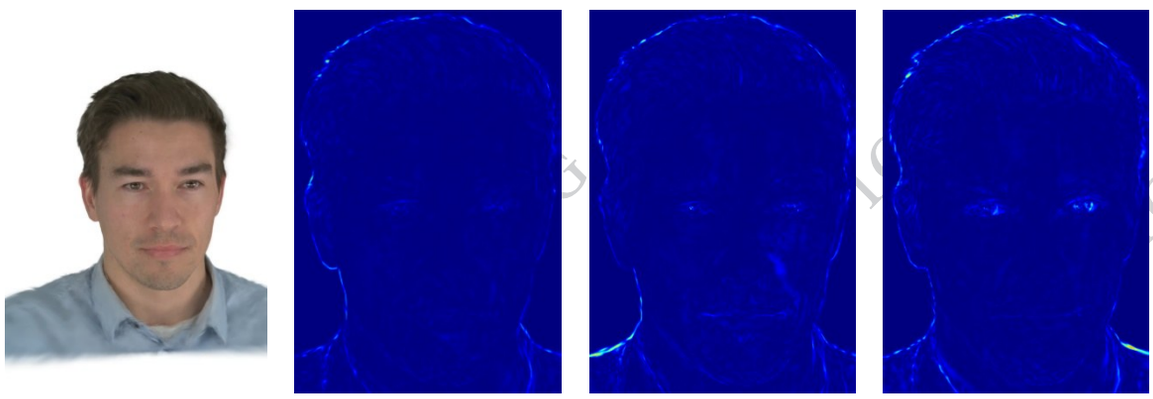


图8　TPAvatar不同表情特征基的可视化结果。

Fig. 8　Visualization of the effects of different expression feature blendshapes of TPAvatar.

3)局限性。TPAvatar仍存在一些局限性。例如,模型对于发丝、细粒度皱纹等细节的重建质量仍可以进一步提高。在表情泛化性方面,对于一些极端的表情在渲染时仍可能出现伪影。

4)未来的工作。未来将通过引入局部非线性基函数、引入物理先验约束等方式来进一步提升模型建模极端表情下皮肤滑动的复杂非线性动力学特征的能力。此外,未来将通过在合成数据上训练通用先验模型将方法拓展到前馈式的单张图像重建任务。另一方面,后续研究将与重光照等下游任务相结合,使模型适应不同的光照条件,从而应用于更复杂的场景。

参考文献(References)

Aneja S, Weiss S, Baeza I, Chandran P, Zoss G, Niessner M, et al. 2025. ScaffoldAvatar: high-fidelity Gaussian avatars with patch expressions//Proceedings of ACM SIGGRAPH 2025 Conference Papers. Vancouver, BC, Canada: ACM: 90: 1-90: 11 [DOI: 10.1145/3721238.3730729]

Blanz V and Vetter T. 1999. A morphable model for the synthesis of 3D faces//Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co.: 187-194 [DOI: 10.1145/311535.311556]

Chen Y, Wang L, Li Q, Xiao H, Zhang S, Yao H, et al. 2024. MonoGaussianAvatar: monocular Gaussian point-based head avatar//Proceedings of ACM SIGGRAPH 2024 Conference Papers. Denver, CO, USA: ACM: 58: 1-58: 9 [DOI: 10.1145/3641519.3657499]

Chu X G and Harada T. 2024. Generalizable and animatable Gaussian head avatar//Advances in Neural Information Processing Systems 37. Vancouver, Canada: Neural Information Processing Systems Foundation, Inc.: 57642-57670 [DOI: 10.52202/079017-1838]

Feng Y, Feng H, Black M J and Bolkart T. 2021. Learning an animatable detailed 3D face model from in-the-wild images. ACM Transactions on Graphics, 40(4): 88: 1-88: 13 [DOI: 10.1145/3450626.

3459936]

Gao X, Liu D, Zhang J. 2024. Multi-modal digital human modeling, synthesis, and driving: a survey. Journal of Image and Graphics, 29(09):2494-2512 (高玄, 刘东宇, 张举勇. 2024. 多模态数字人建模、合成与驱动综述. 中国图象图形学报, 29(09):2494-2512) [DOI: 10.11834/jig.230649.]

Gao X, Zhong C, Xiang J, Hong Y, Guo Y and Zhang J. 2022. Reconstructing personalized semantic facial NeRF models from monocular video. ACM Transactions on Graphics, 41(6): 200:1-200:12 [DOI: 10.1145/3550454.3555501]

Giebenhain S, Kirschstein T, Rünz M, Agapito L and Nießner M. 2024. NPGA: neural parametric Gaussian avatars//Proceedings of ACM SIGGRAPH Asia 2024 Conference Papers. Tokyo, Japan: ACM: 127:1-127:11 [DOI: 10.1145/3680528.3687689]

Hao Conghui, Du Youyang, Wang Lu, Wang Beibei. 2024. Survey of digital face rendering and appearance recovery methods. Journal of Image and Graphics, 29(09):2513-2540 (郝琮晖, 杜悠扬, 王璐, 王贝贝. 2024. 数字人脸渲染与外观恢复方法综述. 中国图象图形学报, 29(09):2513-2540) [DOI: 10.11834/jig.230683.]

He Y, Gu X, Ye X, Xu C, Zhao Z, Dong Y, et al. 2025. LAM: large avatar model for one-shot animatable Gaussian head//Proceedings of ACM SIGGRAPH 2025 Conference Papers. Vancouver, BC, Canada: ACM: 27:1-27:13 [DOI: 10.1145/3721238.3730706]

Kerbl B, Kopanas G, Leimkuehler T and Drettakis G. 2023. 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42 (4): 139: 1-139: 14 [DOI: 10.1145/3592433]

Kirschstein T, Qian S, Giebenhain S, Walter T and Nießner M. 2023. NeRSemble: multi-view radiance field reconstruction of human heads. ACM Transactions on Graphics, 42 (4): 161: 1-161: 14 [DOI: 10.1145/3592455]

Li L, Li Y, Weng Y, Zheng Y and Zhou K. 2025. RGBAvatar: reduced Gaussian blendshapes for online modeling of head avatars//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 10747-10757 [DOI: 10.1109/CVPR52734.2025.01004]

Li T, Bolkart T, Black M J, Li H and Romero J. 2017. Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, 36(6): 194:1-194:17 [DOI: 10.1145/3130800.3130813]

Ma S, Weng Y, Shao T and Zhou K. 2024. 3D Gaussian blendshapes for head avatar animation//Proceedings of ACM SIGGRAPH 2024 Conference Papers. Denver, CO, USA: ACM: 60:1-60:10 [DOI: 10.1145/3641519.3657462]

Mildenhall B, Srinivasan P P, Tancik M, Barron J T, Ramamoorthi R and Ng R. 2022. NeRF: representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1): 99-106 [DOI: 10.1145/3503250]

Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. 2023. DINOv2: learning robust visual features without supervision[EB/OL].[2026-03-29]. https://arxiv.org/abs/2304.07193 [DOI: 10.48550/ARXIV.2304.07193]

Pan Ye, Li Shaoxu, Tan Shuai, Wei Junjie, Zhai Guangtao, Yang Xiaokang. 2025. Advancements in digital character stylization, multimodal animation, and interaction. Journal of Image and Graphics, 30(02):0334-0360 (潘烨, 李韶旭, 谭帅, 韦俊杰, 翟广涛, 杨小康. 2025. 数字人风格化、多模态驱动与交互进展. 中国图象图形学报, 30(02):0334-0360) [DOI: 10.11834/jig.230639.]

Qian S, Kirschstein T, Schoneveld L, Davoli D, Giebenhain S and Nießner M. 2024. GaussianAvatars: photorealistic head avatars with rigged 3D Gaussians//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 20299-20309 [DOI: 10.1109/CVPR52733.2024.01919]

Saito S, Schwartz G, Simon T, Li J and Nam G. 2024. Relightable Gaussian codec avatars//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 130-141 [DOI: 10.1109/CVPR52733.2024.00021]

Shao Z, Wang Z, Li Z, Wang D, Lin X, Zhang Y, et al. 2024. SplattingAvatar: realistic real-time human avatars with mesh-embedded Gaussian splatting//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 1606-1616 [DOI: 10.1109/CVPR52733.2024.00159]

Taubner F, Zhang R, Tuli M and Lindell D B. 2025. CAP4D: creating animatable 4D portrait avatars with morphable multi-view diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 5318-5330

Taubner F, Zhang R, Tuli M, Bahmani S and Lindell D B. 2025. MVP4D: multi-view portrait video diffusion for animatable 4D avatars//Proceedings of ACM SIGGRAPH Asia 2025 Conference Papers. New York, NY, USA: ACM: 125:1-125:11 [DOI: 10.1145/3757377.3763889]

Teotia K, Kim H, Garrido P, Habermann M, Elgharib M and Theobalt C. 2024. GaussianHeads: end-to-end learning of drivable Gaussian head avatars from coarse-to-fine representations. ACM Transactions on Graphics, 43(6): 264:1-264:12 [DOI: 10.1145/3687927]

Wang Y, Wang X, Yi R, Fan Y, Hu J, Zhu J, et al. 2025. 3D Gaussian head avatars with expressive dynamic appearances by compact tensorial representations//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 21117-21126

Xiang J, Gao X, Guo Y and Zhang J. 2024. FlashAvatar: high-fidelity head avatar with efficient Gaussian embedding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 1802-1812 [DOI: 10.1109/

CVPR52733.2024.00177］

Xu Y, Chen B, Li Z, Zhang H, Wang L, Zheng Z, et al. 2024. Gaussian head avatar: ultra high-fidelity head avatar via dynamic Gaussians//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 1931-1941［DOI: 10.1109/CVPR52733.2024.00189］

Xu Y, Wang L, Zhao X, Zhang H and Liu Y. 2023. AvatarMAV: fast 3D head avatar reconstruction using motion-aware neural voxels//Proceedings of ACM SIGGRAPH 2023 Conference Papers. Los Angeles, CA, USA: ACM: 47:1-47:10［DOI: 10.1145/3588432.3591567］

Ye V, Li R, Kerr J, Turkulainen M, Yi B, Pan Z, et al. 2025. gsplat: an open-source library for Gaussian splatting. Journal of Machine Learning Research, 26(34): 1-17

Zhang D, Liu Y, Lin L, Zhu Y, Li Y, Qin M, et al. 2025. GUAVA: generalizable upper body 3D Gaussian avatar//Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu, HI, USA: IEEE: 14205-14217

Zhao Baoquan, Fu Yiyu, Su Zhuo, Wang Ruomei, Lyu Chenlei, Luo Xiaonan. 2024. A survey on multimodal information-guided 3D human motion generation. Journal of Image and Graphics, 29(09): 2541-2565 (赵宝全, 付一愉, 苏卓, 王若梅, 吕辰雷, 罗笑南. 2024. 多模态信息引导的三维数字人运动生成综述. 中国图象图形学报, 29(09): 2541-2565)［DOI: 10.11834/jig.230626.］

Zielonka W, Bolkart T, Beeler T and Thies J. 2025. Gaussian eigen models for human heads//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 15930-15940

Zielonka W, Bolkart T and Thies J. 2023. Instant volumetric head avatars//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada: IEEE: 4574-4584［DOI: 10.1109/CVPR52729.2023.00444］

Zielonka W, Garbin S J, Lattas A, Kopanas G, Gotardo P, Beeler T, et al. 2025. Synthetic prior for few-shot drivable head avatar inversion//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 10735-10746［DOI: 10.1109/CVPR52734.2025.01003］

作者简介

毛宇翔，男，博士研究生，主要研究方向为计算机图形学、虚拟现实、人工智能。E-mail: maoyuxiang22z@ict.ac.cn

叶永竞，男，助理研究员，主要研究方向为计算机图形学、虚拟现实、人工智能。E-mail: yeyongjing@ict.ac.cn