热点追踪

基于三维高斯的动态人脸重建与几何纹理联合优化的语音驱动人脸

上海交通大学 李炫辰 程宇豪 晏轶超

本文是上海交通大学人工智能研究院晏轶超副教授团队的系列研究成果。高精度动态三维人脸资产的重建与驱动是三维数字人产业的核心问题，长期受到学术界和产业界的广泛关注。在动态人脸重建方面，当前业界广泛使用多目立体视觉和非刚性配准结合的方法。然而这种方法容易失效，且需要专业艺术家耗费大量时间精力进行手工修饰。针对这一痛点问题，团队提出了首个基于拓扑一致的三维高斯人脸表示的动态人脸重建框架Topo4D[1]，能够高效重建动态高保真人脸几何和8K分辨率纹理贴图，加速传统方法数十倍，且具有优秀的时序稳定性和面对极端表情的鲁棒性（如图1）。在语音驱动人脸方面，现有的工作都只聚焦于单独驱动人脸几何，而忽视了随几何一致变化的动态纹理贴图的生成。为了推动人脸几何与纹理的协同驱动生成研究，团队基于Topo4D的高效重建能力，兼顾效率与精度，构建了一个具有100个受试者的扫描级语音-模型-贴图对齐数据集TexTalk4D。基于该数据集，团队提出了首个基于扩散的语音协同驱动几何纹理框架TexTalker，能够根据任意语音输入同时生成与音频对齐的动态人脸模型和与面部运动一致变化的动态纹理贴图。大量的实验证明该方法在几何纹理精度和一致性方面都优于已有方法。论文[1]的项目网页公开在https://xuanchenli.github.io/Topo4D/.

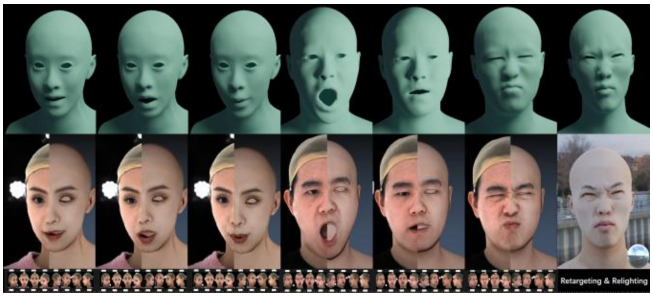

图 1　Topo4D 的结果示例。Topo4D 能够生成时序一致、拓扑一致的高保真网格模型和 8K 分辨率贴图。重建的资产可以直接应用于重定向和重打光。

一、研究背景与相关工作

1.1 高精度面部资产重建

获取具有预定义拓扑结构的高保真面部模型和贴图是业界一个长期存在的研究难点。随着3D可变形人脸被提出，许多人脸重建方法利用参数化模型实现基于单图或多目图像的面部资产重建。然而受限于参数化模型的表现力，这些方法难以忠实重建极端表情。在工业界常用非刚性ICP方法配准由多目立体视觉重建的高模。然而直接将该方法扩展到动态人脸的重建会导致时序上不稳定，导致纹理漂移等问题。为了实现稳定的动态重建，现有CG管线通常使用专业软件逐帧对高精度模型进行重拓扑，涉及经验丰富的艺术家投入大量时间和专业知识。为了解决这一问题，一些方法利用光流或其他技术作为监督来变形模板网格。然而，这些方法需要细致调整大量超参数，难以泛化到不同的个体，并且计算速度较慢。当前也有方法直接从图像中回归人脸模型，但这些方法需要大量昂贵的训练数据，并且难以扩展到新的采集系统，同时对于数据集之外的表情也容易失效。本文中我们提出的Topo4D能够在显著降低时间和人力成本的同时，获取通常需要艺术家手动制造的高质量面部模型和贴图。

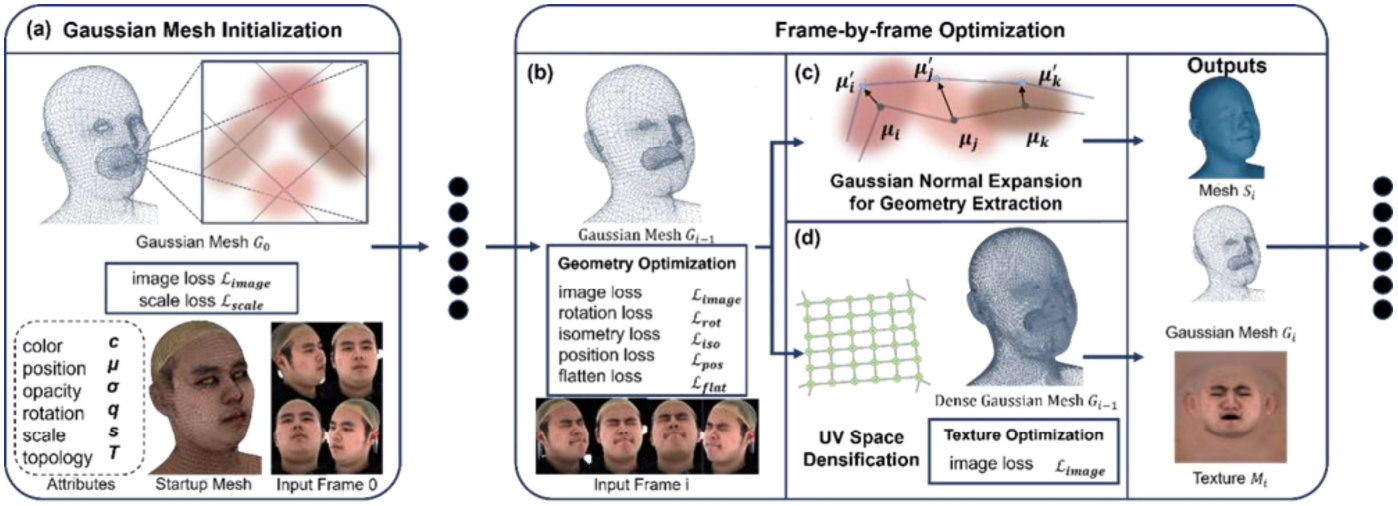

图 2　Topo4D 框架示意图。（a）初始化高斯网格并构建高斯点之间的拓扑关联。（b）利用物理先验与几何先验损失优化几何。（c）通过高斯法向偏移拉近高斯表面与渲染表面。（d）执行 UV 空间稠密化学习细致毛孔级纹理细节。

1.2 语音驱动三维人脸动画

语音驱动三维人脸动画的主要难点在于学习音素与视位素之间的复杂映射。传统的基于规则或数据驱动的方法利用手工设计的映射关系从音节中生成网格变形，该方法人力成本高，且无法泛化到新的语言。许多基于学习的方法能够从对齐数据中挖掘音频特征与人脸运动之间的关联。虽然这些工作已经能够生成生动的面部运动，却都忽视了动态纹理贴图的生成。如果没有动态纹理，则无法表现人脸运动过程中的毛孔挤压和褶皱变化，将会大大降低渲染真实感，甚至导致恐怖谷效应。

为了获取与面部肌肉运动一致变化的纹理贴图，传统CG管线通常从个性化的高精度表情基底中构建专属的动态纹理库，并根据顶点位移采样纹理库来实现纹理变化。该方法需要高成本的表情资产，无法泛化到不同的人，同时计算开销巨大。当前已有方法能够生成动态纹理，但是都忽视了时序稳定性。为了更好地研究几何与纹理之间的一致关联学习，促进几何与纹理的协同生成研究，我们提出了一个新的扫描级语音-模型-贴图对齐数据集TexTalk4D，包含100个个体的说话数据。我们基于此数据集提出的几何纹理协同生成框架TexTalker通过统一纹理与几何的表示，利用隐扩散模型在低维隐空间中学习几何与纹理的复杂关联，实现了高度一致和稳定的语音驱动人脸动画生成。

二、Topo4D

本节介绍提出的Topo4D，整体框架如图2所示。该方法旨在从多视角视频中实现时序稳定的头部重建和纹理生成。具体来说，给定来自K个视角的共F帧分辨率为$h \times w$的多视角图像序列$\{I_i^j \in \mathbb{R}^{h\times w\times 3} | 0 \le i \le F-1\}$，我们的方法能够重建具有固定拓扑T的头部网格模型序列$\{S_i := (V_i, T) | V_i \in \mathbb{R}^{n_v\times 3}\}_{i=0}^{F-1}$和8K纹理贴图序列$\{M_i \in \mathbb{R}^{8192\times 8192\times 3}\}_{i=0}^{F-1}$，其中$n_v$为顶点个数。

接下来本节分别介绍Topo4D中的高斯网格初始化、动态几何与纹理交替优化、以及从高斯网格中提取高质量资产的方法。最后我们通过展示实验结果证明Topo4D的优越性。

2.1 高斯网格初始化

原始三维高斯场景由于缺少拓扑关系，难以直接从中提取可用的网格模型。为了解决这个难点，我们提出了高斯网格人脸来将网格拓扑关系结合到三维高斯表示中。具体来说，我们将第i帧的人脸表示为一个三维高斯点集$G_i = \{G_{i,j}\}_{j=1}^{n_v}$，其中每个高斯点都有一组可学习参数：颜色（$\boldsymbol{c}$）、坐标（$\boldsymbol{\mu}$）、透明度（$\boldsymbol{\sigma}$）、旋转四元数（$\boldsymbol{q}$）、尺寸（$\boldsymbol{s}$）。为了获取预定义的拓扑结构，我们利用一个初始网格模型的顶点坐标和贴图颜色来初始化第一帧的高斯网格。并利用图像差异损失（$\boldsymbol{\mathcal{L}_{image}}$）和大小约束损失($\boldsymbol{\mathcal{L}_{scale}}$)在多视角图像上优化初始高斯网格的各个属性来更好地用高斯网格人脸拟合真实人脸：

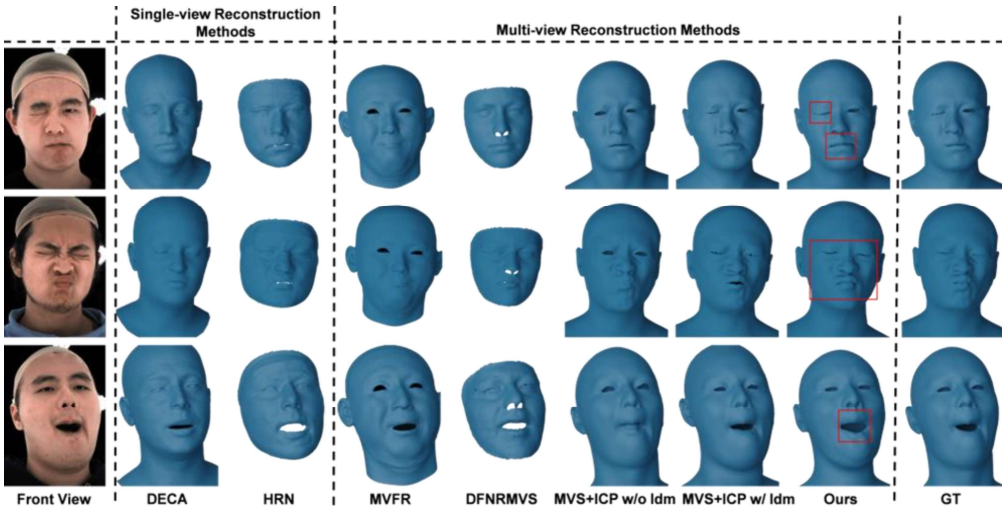

图 3　Topo4D 和其他拓扑一致的重建方法的重建结果比较。我们使用艺术家手工配准的网格作为 GT，红框突出了难以重建的区域。

$$\mathcal{L}_{\text{image}} = (1-\lambda_{image})\mathcal{L}_1(\mathbf{I}_0, \mathbf{I}_0') + \lambda_{image}\mathcal{L}_{D-SSIM}(\mathbf{I}_0, \mathbf{I}_0'),$$

$$\mathcal{L}_{\text{scale}} = \sum_{i\in G}(\|\boldsymbol{s}_{0,i}\|_{-\infty} + \boldsymbol{max}(\mathbf{0}, \boldsymbol{s}_{0,i} - \lambda_{init}\boldsymbol{s}_{init,i})),$$

其中图像差异损失比较真实照片$\mathbf{I}_0$和渲染结果$\mathbf{I}_0'$之间的差异，用$\mathcal{L}_1$损失和$\mathcal{L}_{D-SSIM}$损失加权表示。大小约束损失用于约束高斯点的尺寸，目的是限制高斯点法向上的尺寸接近于0，同时整体尺寸不超出初始值太多，使得高斯分布之间的重叠尽可能少，同时让高斯点贴合人脸表面。

2.2 几何纹理交替优化

在初始化得到第一帧的高斯网格后，我们提出了一种交替几何和纹理优化方法来在前一帧的基础上优化得到当前帧的高斯网格几何，并从多视角图像中学习细节的纹理颜色。

几何优化 直接优化高斯点的位置属性会导致模型顶点错乱。为了在优化过程中维持高斯点之间的拓扑结构和规则的网格排列，我们引入了基于物理先验和基于拓扑先验的损失函数来约束优化过程。该阶段损失包含三项：$\mathcal{L}_{\text{geo}} = \mathcal{L}_{\text{image}} + \mathcal{L}_{\text{phy}} + \mathcal{L}_{\text{topo}}$.

具体来说，面部的低频细节区域会导致高斯点跟踪错误，从而导致破坏网格的拓扑结构。在Luiten[2]等人提出的物理约束的基础上，我们向其中引入了拓扑关系，让每个高斯点受到其一环邻域中的点的约束：

$$\mathcal{L}_{\text{rot}} = \frac{1}{2n_e}\sum_{i\in G}\sum_{j\in K_i}\omega_{i,j}\|\hat{q}_{t,j}\hat{q}_{t-1,j}^{-1} - \hat{q}_{t,i}\hat{q}_{t-1,i}^{-1}\|_2,$$

其中$\hat{q}$为归一化四元数，n_e为边数，ω为影响权重，K_i为高斯点$G_{t,i}$的邻接点。影响权重考虑了初始边长：

$$\omega_{i,j} = exp(-\lambda_\omega\|\mu_{0,j} - \mu_{0,i}\|_2^2).$$

除了对约束高斯点的旋转相似性，我们发现对点于点之间距离施加刚性约束有助于保持长时间范围内的稳定高斯点跟踪：

$$\mathcal{L}_{\text{iso}} = \frac{1}{2n_e}\sum_{i\in G}\sum_{j\in K_i}\omega_{i,j}\left|\|\mu_{0,j} - \mu_{0,i}\|_2 - \|\mu_{t,j} - \mu_{t,i}\|_2\right|.$$

总之，物理先验损失为上述两项的加权组合：

$$\mathcal{L}_{\text{phy}} = \lambda_{rot}\mathcal{L}_{\text{rot}} + \lambda_{iso}\mathcal{L}_{\text{iso}}.$$

物理先验损失保证了高斯点动态跟踪的稳定性，但并不保证高斯网格表面和布线的规整。为了解决这个问题，我们进一步引入拓扑先验损失。具体来说，我们让每个高斯顶点向其邻接点的中心位置靠近：

$$\mathcal{L}_{\text{pos}} = \frac{1}{n_v}\sum_{i\in G}(\mu_{t,i} - \frac{\sum_{j\epsilon\ i}\mu_{t,j}}{|K_i|})^2.$$

Type	Methods	<0.2mm (%)↑	<0.5mm (%)↑	<1mm (%)↑	<2mm (%)↑	<3mm (%)↑	Mean (mm)↓	Med. (mm)↓
Single-view	DECA	2.055	5.136	10.264	20.075	29.046	8.104	5.929
	HRN	5.170	12.786	20.734	44.692	60.263	2.871	2.429
Multi-view	MVFR	4.139	10.130	19.407	34.629	43.661	7.800	4.357
	DFNRMVS	3.447	8.579	17.064	33.479	48.356	3.649	3.214
	Topo4D (Ours)	**22.485**	**52.856**	**87.376**	**94.379**	**97.697**	**0.686**	**0.471**

表 1　不同的人脸重建方法的定量比较。我们计算了在不同的误差水平内的顶点所占的百分比，并计算平均误差和中位数

为了获取平滑的表面，防止模型出现穿模和突刺，我们同时对面片之间的夹角进行约束：

$$\mathcal{L}_{\text{flat}} = \sum_{\theta_i \epsilon e_i}(1 - cos(\theta_{t,i} - \theta_{0,i})),$$

其中$\theta_{t,i}$为第t帧公共边为e_i的面片夹角。

总之，我们的拓扑先验损失为上述两项的加权组合：

$$\mathcal{L}_{\text{topo}} = \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{flat}}\mathcal{L}_{\text{flat}}.$$

纹理优化 在获得当前帧几何后，我们继续学习高细节的纹理。学习8K纹理需要大量的高斯点，与传统3DGS的致密化方法不同，我们提出UV空间稠密化。具体来说，在第t帧，我们首先用高斯网格G_t初始化稠密高斯网格G_t'。然后，我们通过双线性插值向每个四边形网格插入$N \times N$个高斯点，并初始化其属性。我们通过图像损失从4K原始图像中学习细节的高频纹理。

2.3 高质量资产提取

在当前帧优化结束后，可以从G_t和G_t'中提取网格和纹理。为了让高斯表面更加贴近渲染表面，我们通过高斯法向偏移让高斯中心偏移至高斯分布边缘，并直接从中提取顶点坐标。为了获取纹理，我们将G_t'中学习到的颜色属性通过预定义的UV坐标映射渲染至UV贴图空间。

2.4 相关实验

数据集 我们用光场相机采集了一个包含10个个体的数据集用于实验。每个受试者分别表演一段随机的表情和说一段话，包括各种极端和不对称的表情。每段视频长约10秒，每帧包含来自16个视角的分辨率为4096×3000的同步图像，帧率为60 fps。

几何精度评估 我们将Topo4D与三种拓扑一致的人脸重建方法进行比较：(1) 单目方法DECA[3]和HRN[4]；（2）多目方法MVFR[5]和DFNRMVS[6]；（3）使用人脸关键点引导的传统MVS和ICP结合方法。

图3展示了定性对比的结果，我们的方法可以忠实地捕捉不对称的极端表情和微小的面部变化，优于其他基于学习的方法。此外，我们将我们的方法与传统的基于优化的MVS+ICP管线进行了比较，同时在其中使用了关键点作为引导。即使是最先进的关键点检测器在极端表情下也会有较大的误差，导致配准错误，并且在精细的面部区域容易发生穿模。相比之下，我们的方法在不使用任何额外监督的情况下实现了正确的重建。

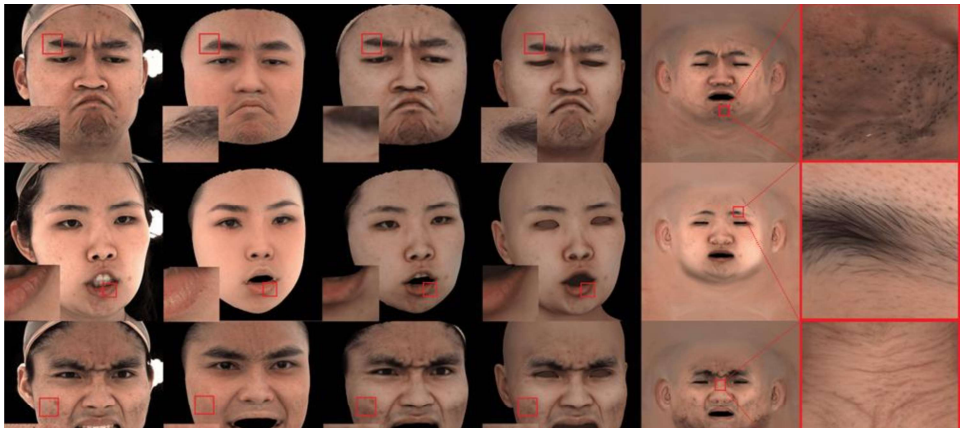

图 4　纹理质量比较。第 5 列和第 6 列展示了 Topo4D 生成的 8K 纹理贴图和其中的毛孔级细节

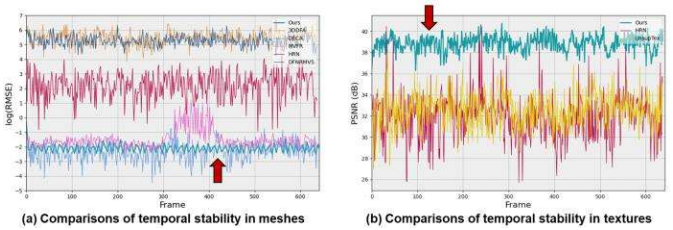

图 5　时间稳定性比较，Topo4D 用红色箭头标出

表1展示了定量比较的结果，我们计算了顶点到扫描模型表面的距离作为误差。我们的方法在所有指标上都显著优于其他拓扑一致的方法。由于我们在高斯初始化过程中引入了可靠的几何形状和颜色先验，大多数顶点都位于高精度范围内，在0.5mm精度范围内的顶点超过52.8%。

纹理质量评估 我们和最先进的面部纹理估计模型定性对比纹理质量，包括UnsupTex[7]和HRN[4]。图4显示了不同方法的渲染结果和Topo4D生成的8K纹理。UnsupTex和HRN的纹理缺乏真实细节且分辨率低。相比之下，我们的方法直接生成高质量的8K纹理，而无需进行上采样，忠实地捕捉面部皱纹、头发和毛孔，并实现了明显优越的渲染质量。

时序稳定性评估 我们比较了重建的网格和纹理的时间稳定性。在网格方面，我们通过计算相邻帧之间的RMSE来衡量稳定性。图5（a）展示了每种方法在一段视频上的重建曲线。Topo4D重建结果更加稳定，而其他方法有很大的波动。在某些情况下DECA的稳定性更好。这是因为DECA没有捕捉到一些极端的表情或微小的面部变化，因此倾向于保持网格不变。相反我们的方法可以忠实地捕捉到极端或细微的表情，同时保持稳定。

纹理方面，我们通过计算相邻帧之间贴图的PSNR来衡量稳定性。如图5（b）所示，我们的方法具有最高的平均值和最低的方差，表明其时序稳定。

消融实验 我们对方法中的核心设计进行了消融实验来评估其作用。图6展示了消融各个损失项、高斯法向偏移和改变输入视角数量的结果。物理先验损失有助于实现正确的密集对应，尤其嘴唇和鼻孔这些容易被遮挡且顶点密集的部位。虽然高斯尺寸损失似乎对几何精度的影响有限，但有助于学习毛孔级纹理细节和避免模糊，如图7所示。拓扑先验损失有效保证了不可见区域的稳定，同时避免了网格穿模和突刺问题。我们评估了输入视角的数量对重建的影响，如图6所示。即使只有一半的输入视图，我们的方法仍然可以产生有竞争力的结果，证明其也适用于只有较少视角的拍摄系统。然而，当视角数减少到4时，重建网格表现出明显的失真。

图7展示不同稠密化密度对纹理质量的影响。随着高斯点数量减少，纹理变得模糊并丢失细节。

三、TexTalk4D数据集构建

Topo4D能够兼顾效率和精度，高效重建扫描级动态人脸模型。为了实现具有动态纹理的语音驱动人脸动画模型，需要一个包含高质量纹理的音频-网格-纹理对齐数据集。虽然当前已经有许多可用的动态人脸数据集，但这些数据集大多不包含纹理贴图，或只提供低分辨率、稳定性差、缺少毛孔级细节的贴图，难以用于训练。因此我们基于Topo4D构建了一个包含8K分辨率动态纹理的扫描级精度数据集。

我们采集和构建数据的流程如图8所示。具体来说，

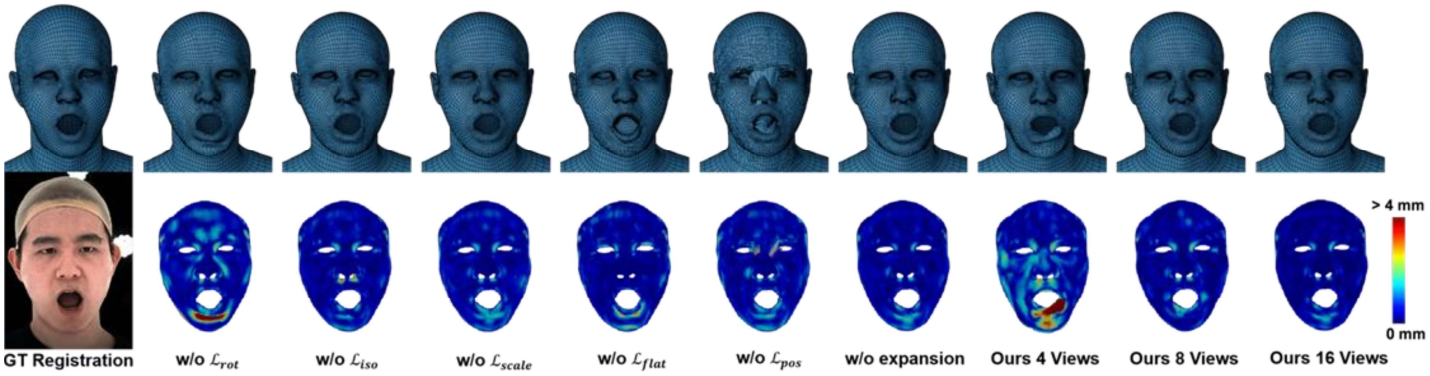

图 6　几何消融实验。第二行的热图展示了顶点距离真实表面的误差，颜色越亮代表误差越大。

我们手工重建了每段视频的第一帧模型作为初始化。由于原始的Topo4D难以重建高速的眨眼动作，我们通过引入人脸先验进行改进。在重建每一帧前，我们使用Mediapipe估计BS系数，并使用一套通用的表情基底作为当前帧的几何初始化。尽管估计的系数并不准确，但Topo4D可以在这个近似的基础上进行优化实现正确的重建。在获取了人脸几何后，我们将来自24个视角的4K图像颜色映射到UV空间中获取8K分辨率纹理贴图。我们通过将贴图与模板进行混合来消除头发等冗余区域。

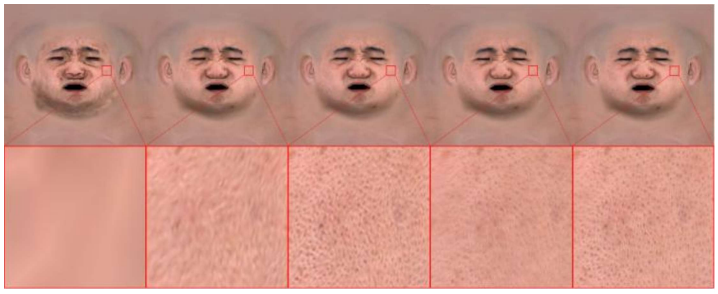

图 7　纹理消融实验

四、TexTalker

在TexTalk4D数据集的基础上，我们提出了一个能够音频同时驱动生成动态纹理与网格的模型TexTalker，如图9所示。我们的目标是从任意输入语音中生成具有个性化纹理的人脸动画资产。具体来说，给定T帧音频$A_{1:T}=(a_1,\dots,a_T)$，其中$a_t\in\mathbb{R}^D$的采样率为D，我们的方法能够生成面部运动序列$W_{1:T}=(w_1,\dots,w_T)$，和纹理序列$M_{1:T}=(m_1,\dots,m_T)$，其中$w_t\in\mathbb{R}^{n_f\times 3}$代表顶点位移，$m_t\in\mathbb{R}^{H\times W\times 3}$为用像素比值表示的纹理变化。除此之外，我们还希望能够控制生成的面部运动的说话风格和纹理变化的褶皱风格。

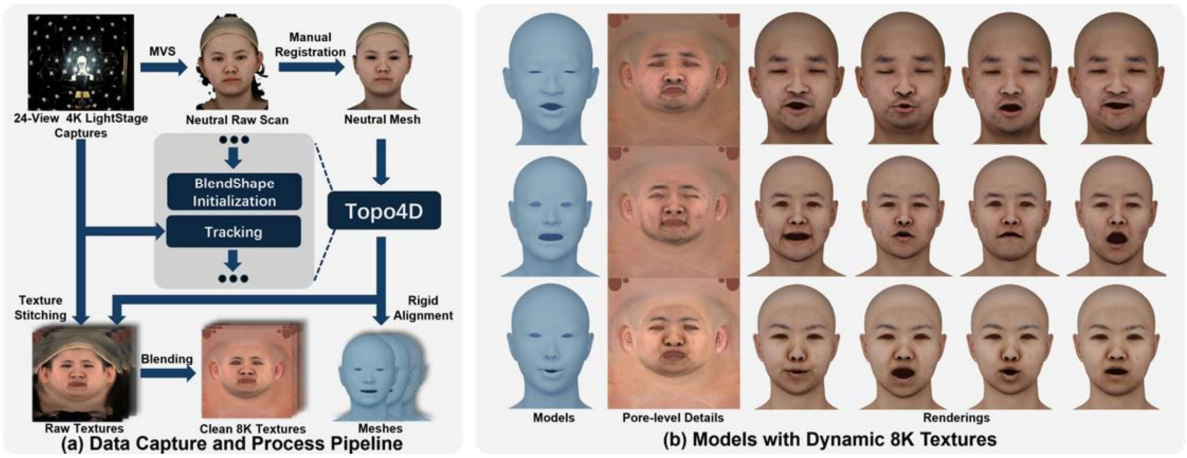

图 8　数据处理流程和数据集资产示例。我们使用 Topo4D 重建光场采集数据。在重建面部几何后，我们从 24 个视图的 4K 图像中映射贴图颜色来获取 8K 纹理。

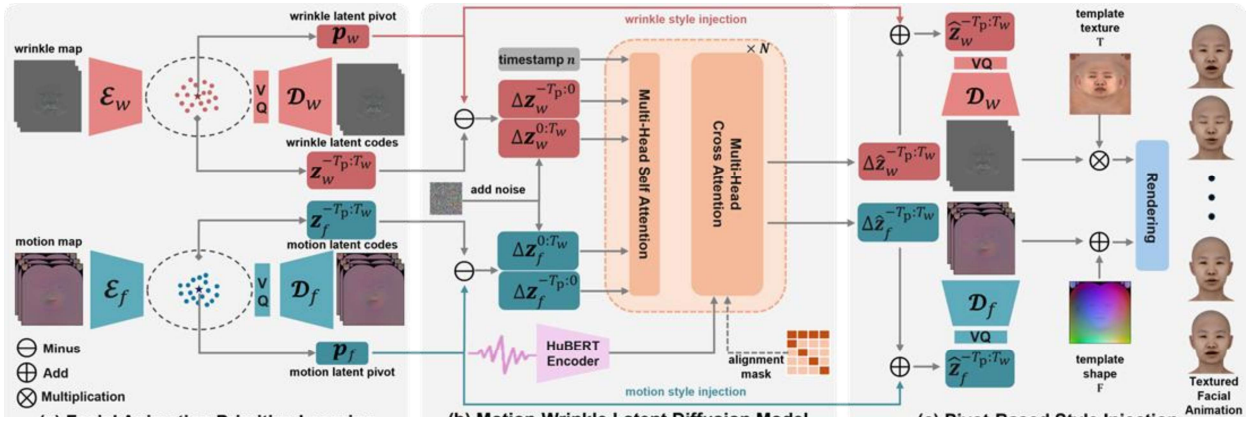

图 9　TexTalker 框架图。（a）我们训练了两个离散码本存储面部动画原语，并高效压缩贴图。（b）基于低维统一的几何与纹理表示，我们训练了隐扩散模型学习隐特征与均值特征的偏移。（c）通过将代表了风格的均值特征加回扩散模型的输出中，我们实现了解耦的说话和纹理风格控制。

为了实现这个任务，我们首先用相同的UV贴图空间统一表示面部运动和纹理变化，并在此基础上分别学习两个离散码本来存储面部动画原语并将贴图压缩至低维隐空间中。在此统一表征的基础上，我们训练了一个隐式扩散模型同时降噪生成几何和纹理隐式特征。最后，我们用隐空间中的平均特征来表示风格，并将其注入到扩散模型的输出中并解码得到所需的动画贴图，实现解耦的说话和纹理风格控制。整体框架如图9所示。

4.1 面部动画原语学习

为了统一几何与纹理的表示。在几何方面，我们通过UV映射将相对于无表情人脸的顶点位移映射到UV贴图空间得到面部运动贴图f。在纹理方面，我们使用和Zhang等人[8]同样的褶皱贴图w来表示纹理变化。

直接生成贴图会导致巨大的计算开销。受到离散先验表示的启发，我们通过训练离散自编码器将贴图压缩至低维空间中。我们分别训练一个几何自编码器$\mathcal{E}_f$和纹理自编码器$\mathcal{E}_w$来分别压缩几何运动和褶皱贴图，后续的生成模型也只需要以较低的开销学习生成隐特征即可。以几何自编码器为例，我们采用VQGAN[9]相同的方式联合训练编码器$\mathcal{E}_f$、码本$\mathcal{C}_f$、生成器$\mathcal{G}_f$、判别器$\mathcal{D}_f$。编码器将贴图压缩成低维隐特征$z_f = \mathcal{E}_f(f) \in \mathbb{R}^{h\times w\times d}$。通过对码本进行最近邻搜索，可以将将该隐特征变换为离散原语特征$\tilde{z}_f = \mathcal{Q}_f(z_f)$。最终，生成器可以从离散原语特征中重建出原始贴图$\hat{f} = \mathcal{G}_f(\tilde{z}_f)$。纹理自编码器以相同的方式实现压缩和重建。训练损失包括 (1) $\mathcal{L}_1$重建损失$\mathcal{L}_{\text{rec}}$，(2) VGG感知损失$\mathcal{L}_{\text{per}}$，(3) 对抗损失$\mathcal{L}_{\text{adv}}$，(4) 码本损失$\mathcal{L}_{\text{code}}$。总损失为以上损失项的加权组合：

$$\mathcal{L}_{\text{latent}} = \mathcal{L}_{\text{rec}} + \eta_{per}\mathcal{L}_{\text{per}} + \eta_{adv}\mathcal{L}_{\text{adv}} + \eta_{code}\mathcal{L}_{\text{code}}.$$

4.2 几何纹理协同优化隐式扩散模型

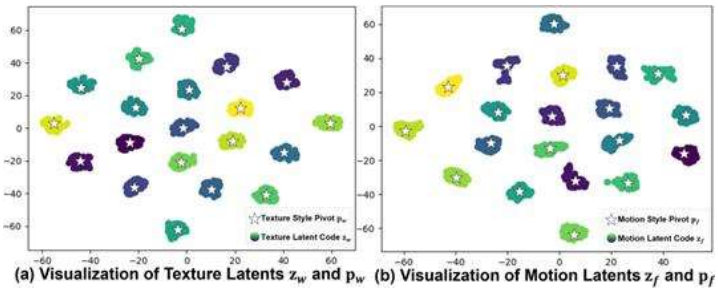

图 10　来自 20 个个体的隐特征空间 t-SNE 可视化

Method	LVE↓ 10^{-2}mm	MVE↓ 10^{-2}mm	FDD↓ 10^{-3}mm
FaceFormer	1.80	2.94	1.68
CodeTalker	1.83	2.80	1.38
FaceDiffuser	1.53	2.38	1.64s
TexTalker	**1.49**	**2.34**	**1.20**

表 2　几何精度比较

基于高效的低维隐式表示，我们训练一个基于Transformer的隐扩散模型$\mathcal{F}$学习语音驱动几何纹理。具体来说，z_f和z_w首先被拼接在一起构成一个动画样本$X^0 = [z_f, z_w]$。随后，向干净样本X^0中逐步加入高斯噪声得到$X^n (n = 1,2,\dots,N)$。$\mathcal{F}$学习在音频特征的引导下从噪声样本中降噪生成干净样本。我们使用预训练的HuBERT编码器生成音频特征。为了实现更好的时序稳定性，同时建模长距离关联关系，模型同时生成一段时序窗口内的特征：

$$\hat{X}^0_{-T_p:T_w} = \mathcal{F}\left(X^n_{0:T_w}, X^0_{-T_p:0}, A_{-T_p:T_w}, n\right).$$

其中我们不仅输入噪声样本$X^n_{0:T_w}$，还输入长度为T_p的来自上个窗口的干净样本$X^0_{-T_p:0}$和降噪时间步n。我们使用样本重建损失训练模型：

$$\mathcal{L}_{\mathcal{F}} = \left|\left|\hat{X}^0_{-T_p:T_w} - X^0_{-T_p:T_w}\right|\right|_2.$$

4.3 解耦的说话与褶皱风格注入

为了实现风格控制，我们的核心思想在于学习到的离散码本存储了动画原语，同时有相似风格的隐特征共享相似的原语特征，因此相同风格的隐特征应当在隐空间中聚类在一起。如图10所示，我们观察到来自相同受试者的面部运动特征与纹理变化特征都被聚类在一起。因此我们将聚类中心作为风格特征：$p = 1/T \sum_{t=1}^{T} z_t$.

这种设计的额外好处是风格特征和隐特征来自同一个空间，因此我们的动画生成模型$\mathcal{F}$只需要学习生成与身份无关但与音频相关的特征偏移量$\Delta z = z - p$，而风格可以被无缝注入，同时进行灵活的风格替换。在此基础上，$\mathcal{F}$生成偏移样本$X'_0 = [\Delta z_f, \Delta z_w]$表示为：

$$\hat{X}'^0_{-T_p:T_w} = \mathcal{F}\left(X'^n_{0:T_w}, X'^0_{-T_p:0}, A_{-T_p:T_w}, n\right).$$

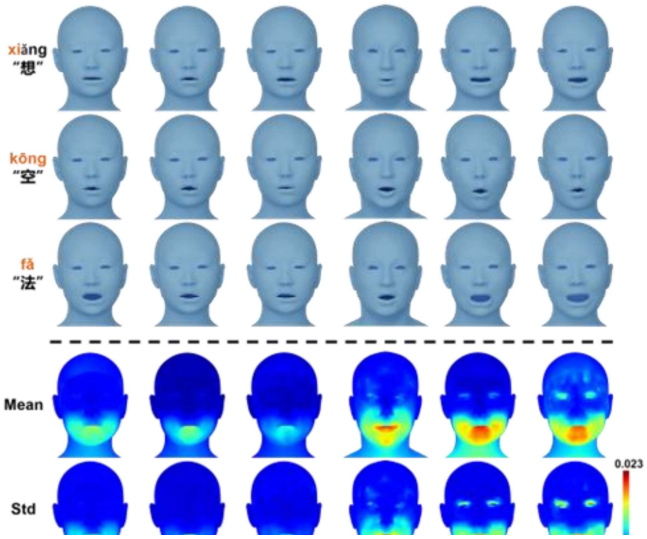

图 11　面部运动生成结果比较。下半部分展示了生成结果在整个序列上的动态统计信息。

最终在推理阶段，可以使用风格i的说话风格特征$p_{f,i}$来构建面部运动特征: $\hat{z}_f = p_{f,i} + \Delta\hat{z}_f$，用风格j的纹理风格特征$p_{w,j}$来构建面部纹理特征: $\hat{z}_w = p_{w,j} + \Delta\hat{z}_w$。利用训练好的生成器$\mathcal{G}_w$和$\mathcal{G}_f$可以从隐式特征重建出具有风格i的面部运动贴图序列和具有风格j的褶皱贴图序列，并从中恢复个性化的动态人脸模型和动态纹理贴图。

4.4 相关实验

数据集　我们将TexTalk4D分为来自85个受试者的80分钟训练集TexTalk4D-Train、来自5个受试者的5分钟验证集TexTalk4D-Valid、来自10个训练集中受试者的5分钟测试集TexTalk4D-Test-A、来自10个未知受试者的5分钟测试集TexTalk4D-Test-B。Test-A用于计算客观指标并衡量几何精度，Test-B用于定性分析。

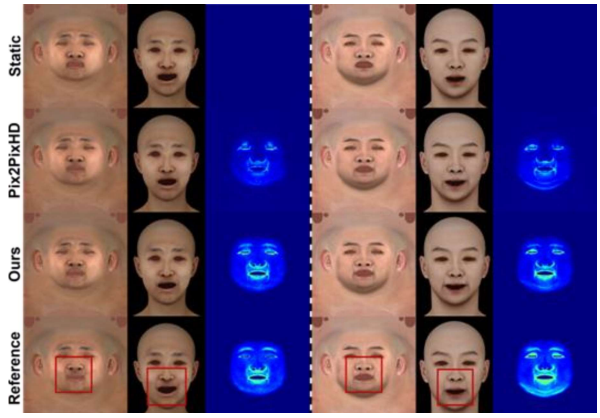

图 12　纹理贴图生成结果比较。第三列和第六列用热图展示了统计热图，颜色越亮代表动态范围越大。

几何质量评估　我们对比了FaceFormer[10]、CodeTalker[11]、FaceDiffuser[12]和DiffPoseTalk[13]。由于DiffPoseTalk只与Flame系数兼容，只进行定性对比。我们计算嘴唇顶点误差（LVE）、全脸顶点误差（MVE）、上脸动态变化（FDD）作为精度衡量指标。比较结果如表2所示，我们的方法优于其他方法。视觉比较结果如图11所示，我们的方法更加匹配真实嘴型。同时能够生成更大动态范围的结果，包括难以生成的自然的眨眼动作。

纹理质量评估　参考Zhao等人[14]，我们训练一个Pix2PixHD学习从面部运动贴图生成纹理贴图，如图12所示。表3展示了定量对比结果，我们方法显著优于静态纹理和Zhao等人的方法。

几何-纹理一致性评估　我们设计了一个user study来进行一致性比较。志愿者从真实感和一致性的角度评估生成的人脸动画的质量，从1-5进行打分。如表3所示，我们的方法生成的动画更加真实，一致性更强。

Meth.	Test-A			Test-B	
	PSNR	SSIM	LPIPS	Real.	Con.
Static	39.79	0.967	0.0146	3.10	2.65
P2PHD	42.34	0.981	0.0187	3.91	3.78
Ours	**44.13**	**0.985**	**0.0101**	**4.13**	**3.97**

表 3　纹理比较。我们在 TexTalk4D-Test-A 上计算指标，在 TexTalk4D-Test-B 上进行 User Study。

五、总结

本文提出了首个基于三维高斯的动态人脸重建方法Topo4D，能够高效重建扫描级几何资产和8K分辨率毛孔级细节纹理贴图。基于Topo4D，我们提出了一个全新的高多样性扫描级精度语音-网格-纹理对齐数据集TexTalk4D，包含100个受试者的说话数据，尤其提供8K动态贴图，能够表现动态面部褶皱变化。基于此数据集，我们提出了首个语音同时驱动几何和纹理的框架TexTalker，能够实现高一致性的说话人脸动画生成，同时能够解耦地控制说话风格和纹理变化风格。大量充分的实验证明了所提出的方法的有效性。该成果已发表于计算机视觉国际顶级会议ECCV 2024和CVPR 2025。

责任编委　张 青

参考文献

[1] Li X, Cheng Y, Ren X, et al. Topo4D: Topology-Preserving Gaussian Splatting for High-fidelity 4D Head Capture[C]//ECCV, 2025: 128-145.

[2] Luiten J, Kopanas G, Leibe B, et al. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis[J]. arXiv preprint arXiv:2308.09713, 2023.

[3] Feng Y, Feng H, Black M J, et al. Learning an animatable detailed 3D face model from in-the-wild images[J].TOG, 2021.

[4] Lei B, Ren J, Feng M, et al. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images[C]//CVPR. 2023: 394-403.\

[5] Xiao Y, Zhu H, Yang H, et al. Detailed facial geometry recovery from multi-view images by learning an implicit function[C]//AAAI. 2022.

[6] Bai Z, Cui Z, Rahim J A, et al. Deep facial non-rigid multi-view stereo[C]//CVPR. 2020: 5850-5860.

[7] Slossberg R, Jubran I, Kimmel R. Unsupervised high-fidelity facial texture generation and reconstruction[C]//ECCV, 2022.

[8] Zhang L, Zeng C, Zhang Q, et al. Video-driven neural physically-based facial asset for production[J].TOG, 2022.

[9] Esser P, Rombach R, Ommer B. Taming transformers for high-resolution image synthesis[C]//CVPR. 2021: 12873-12883.

[10] Fan Y, Lin Z, Saito J, et al. Faceformer: Speech-driven 3d facial animation with transformers[C]//CVPR. 2022: 18770-18780.

[11] Xing J, Xia M, Zhang Y, et al. Codetalker: Speech-driven 3d facial animation with discrete motion prior[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 12780-12790.

[12] Stan S, Haque K I, Yumak Z. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion[C]//Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games. 2023: 1-11.

[13] Sun Z, Lv T, Ye S, et al. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models[J]. TOG, 2024, 43(4): 1-9.

[14] Li J, Kuang Z, Zhao Y, et al. Dynamic facial asset and rig generation from a single scan[J]. TOG, 2020, 39(6): 215:1-215:18.

李炫辰

上海交通大学人工智能研究院博士研究生，导师为晏轶超副教授，主要研究方向为三维数字人脸重建、驱动、生成。

Email:lixc6486@sjtu.edu.cn

程宇豪

上海交通大学人工智能研究院博士研究生，导师为戴琼海院士和晏轶超副教授，主要研究方向为隐式数字人脸的生成/编辑以及基于网格的三维数字人脸的重建/驱动。

Email: chengyuhao@sjtu.edu.cn

基于三维高斯的动态人脸重建与几何纹理联合优化的语音驱动人脸

晏轶超

上海交通大学人工智能研究院副教授，博士生导师。主要研究方向为 AIGC 及三维数字人技术，发表包括 TPAMI、CVPR、NeurIPS 在内的论文 40 余篇。先后主持国家自然科学基金青年项目、CCF-阿里巴巴青年科学家基金等项目 8 项。曾入选上海市海外高层次人才计划，获 2020 年度中国图像图形学学会优秀博士论文奖。

Email:yanyichao@sjtu.edu.cn