

基于多分支空间变化卷积网络的多层次三维人脸表情编辑

任明国^{1,2)}, 迟静^{1,2)*}, 沈天舒^{1,2)}, 陈玉妍^{1,2)}

¹⁾ (山东财经大学计算机科学与技术学院 济南 250014)

²⁾ (山东省数字媒体技术重点实验室 济南 250014)

(chijing@sdufe.edu.cn)

摘要: 针对三维人脸表情编辑中模型的操作繁复、生成表情的真实感低、细节信息缺乏的问题, 提出一种基于多分支空间变化卷积网络的多层次三维人脸表情编辑方法, 可根据面部模型上控制点的位移, 生成真实自然且细节丰富的新表情. 首先, 面部网格经过网络模型的高频区域划分模块, 识别出高频区域; 然后, 将整个面部网格和控制点约束输入粗略编辑模块, 生成基本表情, 同时, 将高频区域和控制点约束输入精细编辑模块, 生成丰富的表情细节; 最后, 将基本表情与表情细节融合得到新表情. 这种分层次处理可使网络以较快的速度生成精细的细节; 通过引入空间变化卷积, 可更好地捕获不规则网格的空间特征, 提高了表情编辑的准确性. 提出融合顶点空间邻近性和运动一致性的时空相关性准则, 并改进 K-Means 聚类算法, 极大地提高了高频区域自动划分的合理性和准确性; 通过构建结合顶点位置和顶点法向约束的新损失函数, 有效地提升了整个网络的精度. 对 4 个角色模型分别使用所提方法进行表情编辑, 与其他方法相比, 所提方法能够根据用户的编辑要求, 生成细节丰富的高真实感表情.

关键词: 表情编辑; 多分支网络; 空间变化卷积

中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024.20008

Multi-Level 3D Facial Expression Editing Based on Multi-Branch Spatially Varying Convolutional Network

Ren Mingguo^{1,2)}, Chi Jing^{1,2)*}, Shen Tianshu^{1,2)}, and Chen Yuyan^{1,2)}

¹⁾ (School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014)

²⁾ (Shandong Provincial Key Laboratory of Digital Media Technology, Jinan 250014)

Abstract: Aiming at the problems in 3D facial expression editing, such as the complicated operation of models, the low reality of generated expressions and the lack of detailed information, this paper proposes a multi-level 3D facial expression editing method based on multi-branch spatially varying convolutional network. The method can generate new expressions that are both realistic and rich in detail based on the displacement of control points on the facial model. Firstly, the facial mesh undergoes a high-frequency region segmentation module in the network model to identify the high-frequency regions. Then, the entire facial mesh and control point constraints are input to the coarse editing module to generate basic expressions. Simultaneously, the high-frequency regions and control point constraints are input to the fine editing module to generate rich expression details. Finally, the basic expressions and expression details are fused to obtain the new expression. This hierarchical processing allows the network to generate fine details at a relatively fast

收稿日期: 2022-10-14; 修回日期: 2023-10-13. **基金项目:** 国家自然科学基金面上项目(61772309); 山东省高等学校青创科技支持计划(2020KJN007); 济南市“新高校 20 条”科研带头人工作室(2021GXRC092); 山东省重点研发计划(2021SFGC0102). 任明国(1997—), 男, 硕士研究生, 主要研究方向为三维表情动画、数字几何处理; 迟静(1980—), 女, 博士, 教授, 博士生导师, 论文通信作者, 主要研究方向为计算机动画、图形图像智能处理、大数据建模; 沈天舒(1998—), 女, 硕士研究生, 主要研究方向为图像处理、微表情识别; 陈玉妍(1997—), 女, 硕士研究生, 主要研究方向为人脸图像合成.

speed. By introducing spatially varying convolution, spatial features of irregular mesh can be better captured, thus improving the accuracy of expression editing. A spatiotemporal correlation criterion that combines vertex spatial proximity and motion consistency is proposed, and the *K*-Means clustering algorithm is improved to greatly enhance the rationality and accuracy of automatic high-frequency region segmentation. A new loss function combining vertex position and vertex normal constraints is constructed, effectively improving the overall accuracy of the network. Using the proposed methods, expression editing was performed on four different character models. Compared to other methods, the proposed method can generate highly realistic expressions with rich details according to the user's editing requests.

Key words: expression editing; multi-branch network; spatially varying convolution

三维人脸表情编辑一直是计算机图形学、计算机动画和虚拟现实等方向的研究热点和难点。人类表情丰富多样,即使细微的表情变化在信息传递过程中也起到重要作用,同时人类对自身的熟悉,又使得任何不自然的表情都极易被察觉。因此如何合成真实自然的表情是三维人脸表情编辑研究中的一个极具挑战性的问题。传统的表情动画制作方式需要动画师操作模型上的大量控制元素,工作量巨大。因此如何为用户提供便捷的操作方式,使用户能够简单、实时且高效地完成表情编辑工作,从而降低表情动画编辑的专业门槛,提高表情合成的效率,是三维人脸表情编辑研究中的另一个重点问题。

传统的三维人脸表情编辑方法可分为基于物理模型的表情编辑、基于参数化的表情编辑和基于样本的表情编辑。基于物理模型的三维人脸表情编辑主要是对三维人脸搭建骨骼或肌肉结构^[1],通过控制相关结构的位移变化,实现表情的变化。但是这些方法需要动画师进行繁杂的操作,工作量较大,且难以模拟皮肤的细节,如皱纹等。线性混合蒙皮(linear blend skinning, LBS)方法是一种简单且快速的面部网格变形方法,也称为骨骼子空间变形(skeleton subspace deformation, SSD)^[2],它通过线性分解的方式将面部皮肤顶点分解为多个由骨骼影响的部分,根据所施加的影响,计算运动时面部皮肤顶点的新位置。但是, LBS 方法多应用于较大网格的变形,对于局部区域变形需要设置较多的骨骼,不利于动画师的操作,也容易产生皮肤塌陷和皮肤扭曲现象。球面混合蒙皮(spherical blend skinning, SBS)算法^[3]和双四元数线性混合蒙皮(dual quaternion linear blending, DLB)算法^[4]在一定程度上克服了 LBS 方法易导致皮肤失真的缺陷。Le 等^[5]提出一种线性蒙皮分解(smooth skinning

decomposition with rigid bones, SSDR)算法,从样本数据中分解出 LBS,继而通过骨骼和顶点权重图近似拟合出面部变形。由于骨骼和权重数据是根据面部变形自动优化得到的,因此生成的骨骼系统可能与动画师的要求不符。基于参数化的三维人脸表情编辑是根据面部表情的变化情况构建面部表情变化函数,并借助函数调整面部表情。Feng 等^[6]提出用于面部变形的基于核典型关联分析的变形回归方法和基于泊松变换的求解技术,其从表情样本数据中学习适当的变形子空间,以便于使用控制点简单、快速地生成表情动画。基于样本的三维人脸表情编辑多采用形状融合变形 blendshape^[7]的思想,将人脸模型表示为若干个已知样本模型的加权线性组合,通过修改样本模型的形状或调整各样本模型的融合权值来生成新的表情;该方法实现简单,但需要庞大的样本量。Chi 等^[8]根据顶点的时空相关性自动地对人脸网格进行区域划分,并根据各控制点的贡献图以及形变模型对各个区域进行变形融合,生成新的人脸表情,且结合表情传递方法,克服了对大规模数据库的依赖。

随着深度学习技术的发展,研究者将其扩展应用到三维人脸表情合成中,如人脸重建^[9-10]、人脸动画生成^[11-12]和表情编辑^[13-14]等,通过提取样本数据库中面部网格变形的空间特征和影响因素,实现面部表情的合成。经典卷积神经网络(convolutional neural network, CNN)被广泛应用于捕获规则网格的空间特征,但是对于不规则的面部网格,其通常含有不规则的拓扑结构,因此仅适用于规则二维或三维网格数据的卷积核,无法直接应用于面部网格。一种常见的折中方法是三维网格数据映射到预定义的 UV 空间中,通过训练经典的二维 CNN 学习 UV 空间中的特征,再将其

恢复到三维网格。例如, Bailey 等^[15]借助 UV 图对三维人脸网格进行处理。但是这种方法不可避免地受到参数化失真, 以及 UV 图中接缝的影响。与经典 CNN 不同, 图卷积网络(graph convolutional network, GCN)通过定义网格卷积运算将其应用到不规则三维网格中。Litany 等^[16]提出一种 GCN 自动编码器, 用来学习非刚性网格的外部形状。Ranjan 等^[17]使用类似的方法生成三维人脸模型。Li 等^[18]通过改变 GCN 的感受野, 提出了多列 GCN 结构来合成三维面部网格。但是 GCN 可能存在过度平滑问题, 导致细节信息丢失, 也可能存在振荡问题。因此, Zhou 等^[19]提出一种使用空间变化卷积的方法, 即在不规则网格上进行卷积操作时, 通过对卷积核进行线性采样, 学习网格的空间特征。空间变化卷积不仅能够直接处理不规则三维网格, 而且在一定程度上能够提取到网格上更多的细节特征信息。此外, Shi 等^[20]提出 liftedGAN 网络框架, 将预先训练好的 StyleGAN2 进行分解和提升, 以生成三维人脸模型。Otterdout 等^[21]结合中性三维人脸模型和表情标签生成动态三维人脸表情, 先利用 Motion3DGAN 学习三维人脸关键点的位移特征, 再利用 S2D 解码器将关键点的运动特征映射为三维人脸网格顶点的位移, 得到三维人脸表情。Tran 等^[22]改进了线性三维形变模型(3D morphable model, 3DMM), 提出了非线性 3DMM 用于三维人脸重建。闫衍英等^[23]根据二维人脸特征点, 利用 CNN 回归出表情 AU 参数, 然后基于 blendshape 模型, 利用 AU 参数和与之对应的三维人脸模型生成表情动画。张睿思等^[24]通过 CNN 学习人脸图像与三维面部骨骼参数之间的映射关系, 并根据人脸图像几何关系优化该映射关系, 然后根据人脸图像和该映射关系生成三维面部表情。虽然上述基于深度学习的表情合成方法取得了较好的进展, 但依然存在结果的精确性依赖样本数据集的大小、训练时间较长和表情的真实感存在不同程度的损失等问题。

在表情编辑中, 三维面部网格的细节处理对表情的真实感有着重要作用。于志平等^[25]在完成源模型与目标模型的基本表情迁移之后, 利用细节特征向量调整策略重建目标模型的细节, 提高了面部表情的真实感。Deng 等^[26]通过对面部皱纹的连续空间的有效采样和插值, 生成面部皱纹。Gecer 等^[27]提出 Fast-GANFIT 网络模型, 采用生成对抗网络(generative adversarial network, GAN)和深度卷积神经网络(deep convolutional neural net-

work, DCNN)重建三维面部形状和纹理。程轩等^[28]提出一种统一框架下的协同补全模型 TDGAN, 根据人脸纹理特征分布和几何细节特征分布的双向关联特性进行三维面部重建, 获得了具有高质量细节的三维人脸, 但是其对于不对称人脸重建的效果较差。万贤美等^[29]提出一种基于时空的人脸表情动画编辑方法, 利用基于拉普拉斯的网格变形技术来保存面部细节特征, 但其需使用大型线性系统, 当网格结构复杂时, 求解速度慢, 占用内存大。Bailey 等^[15]通过增加卷积层数进一步提取 UV 图中的潜在特征, 以此来完善面部表情的细节, 但是由于在网络结构中需借助 UV 图, 不可避免地会丢失空间特征信息。

综上, 现有的三维人脸表情编辑方法在表情细节模拟精度的提升、表情真实感的增强以及用户操作便捷性的提高等方面仍存在较大的研究空间。针对这些问题, 本文提出一种基于多分支空间变化卷积网络的多层次表情编辑方法。该方法允许用户自由指定人脸网格上的少量控制点, 并拖拽控制点, 即可直观、简捷地实现对面部表情的实时编辑, 且生成的新表情既符合用户的编辑要求, 又真实自然、细节丰富。

1 本文方法

本文提出一种基于多分支空间变化卷积网络的三维人脸表情编辑方法。给定一个三维人脸网格, 允许用户自由选择人脸网格上的控制点, 并通过改变控制点的位置, 轻松地实现对面部表情的实时编辑。本文方法的整体网络架构如图 1 所示, 由高频区域划分模块、粗略编辑模块和精细编辑模块 3 部分构成。用户选定并调整控制点后, 将三维面部网格以及控制点的位置变化情况输入到网络中。这里, 控制点的位置变化是对整个面部网格整体变形的约束和指导, 也反映了用户对所生成的新表情的期望, 本文称之为控制点约束。首先高频区域划分模块利用高斯曲率和肘部法则自动确定面部大曲率顶点的位置, 再依据大曲率顶点, 使用以新时空相关性为准则的改进 K-Means 聚类算法和肘部法则, 自动地获取高频区域的个数和每个区域的范围; 然后, 将整个网格和控制点约束输入到粗略编辑模块, 得到粗颗粒的网格变形, 可生成满足用户期望的基本表情, 且执行速度快; 同时, 将面部高频区域和控制点约束输入到精细编辑模块, 得到细颗粒的网格变形, 可生成满足用户期望

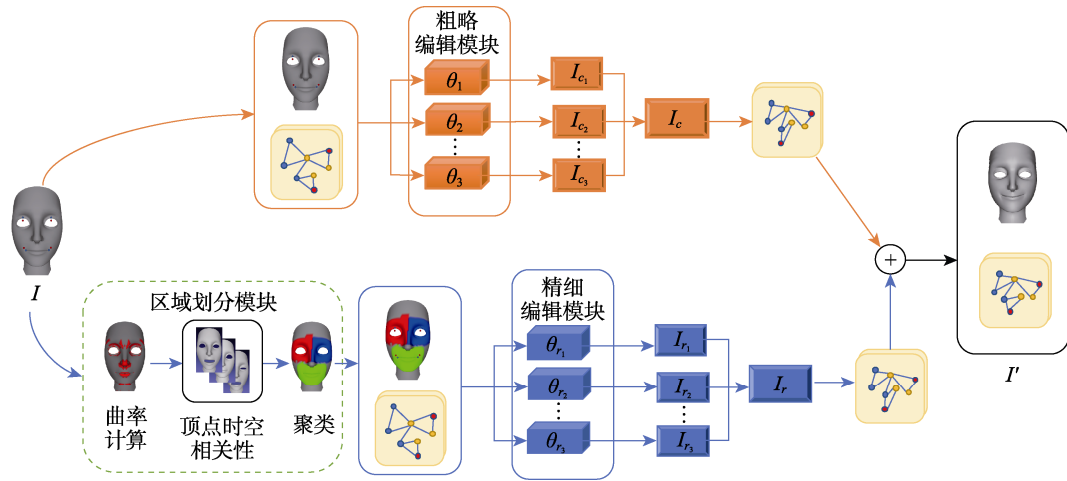


图 1 本文方法框架

的精细的表情细节, 如皮肤褶皱等, 从而丰富表情细节信息; 最后, 将粗略编辑模块得到的基本表情和精细编辑模块得到的细节信息融合, 得到最终的符合用户期望的且真实自然的新表情。

本文方法的数学表达为: 给定一个含有控制点的三维面部网格, 设 $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ 为面部网格 I 的所有顶点位置, n 表示顶点个数; 设 $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_m\}$ 为面部网格 I 的所有控制点位置, m 表示控制点的个数; 设 $\mathbf{P}' = \{\mathbf{p}'_1, \mathbf{p}'_2, \dots, \mathbf{p}'_m\}$ 为用户调整后所有控制点的新位置. 令 $\bar{\mathbf{P}} = \{\bar{\mathbf{p}}_1, \bar{\mathbf{p}}_2, \dots, \bar{\mathbf{p}}_m\}$ 表示控制点的位置变化, 即控制点约束条件, 其中, $\bar{\mathbf{p}}_i = \mathbf{p}'_i - \mathbf{p}_i$ 表示控制点 $\mathbf{p}_i$ 的位移向量. 将面部网格 I 以及控制点约束条件 $\bar{\mathbf{P}}$ 输入到网络模型中, 本文的目标就是经过网络模型计算出面部网格 I 中的各个顶点的新位置, 记作 $\mathbf{V}' = \{\mathbf{v}'_1, \mathbf{v}'_2, \dots, \mathbf{v}'_n\}$. 即估算一个函数 $\mathbf{V}' = F(\mathbf{V}, \bar{\mathbf{P}})$, 使得顶点位置的改变带动整个网格的变形, 且由顶点新位置 $\mathbf{V}'$ 确定的变形后的网格 I' 所表示的新表情, 既符合用户的编辑期望、富有丰富的细节, 又真实自然。

1.1 高频区域划分模块

上述估算网格顶点新位置的过程也可以理解成估算网格的一个变形, 它能够使得新网格所呈现的表情符合用户的期望, 且真实自然. 由于不同尺度的网格可以反映不同的特征, 且处理效率不同, 因此本文将网格的变形划分为粗颗粒和细颗粒 2 个不同尺度的变形. 其中, 粗颗粒的变形是针对整个网格的变形, 反映的是基本表情, 如喜怒哀乐等的变化, 可以大致逼近用户期望的表情; 细颗粒的变形是针对网格上高频区域的变形, 反映的是

表情细节, 如皮肤褶皱等的变化, 可以精确地逼近用户期望的表情细节. 前者的运算量小, 逼近速度快; 后者的精度高, 可实现对表情细节的精确模拟. 因此, 分层次的网格变形可使整个网络模型在生成精细的表情细节的同时, 保持较好的运行效率。

本文采用粗略编辑模块估算粗颗粒的网格变形, 即处理整个网格; 用精细编辑模块估算细颗粒的网格变形, 即处理网格上的高频区域. 显然, 高频区域是否准确识别, 将对后续的分支处理和最终的表情生成结果产生重要影响. 基于此, 本文提出新的时空相关性聚类准则和高频区域划分方法, 以有效地提高区域划分的合理性和准确性。

高频区域划分的步骤如下. 首先求解输入网格各个顶点的高斯曲率, 依据肘部法则自动筛选出大曲率顶点, 并增加大曲率顶点在聚类时的权重; 然后, 使用改进的 K -Means 聚类算法和肘部法则对网格上的所有顶点进行聚类, 自动确定高频区域的个数和每个区域的范围。

具体地, 对输入的三维面部网格 I , 计算顶点的高斯曲率. 顶点 $\mathbf{v}_i$ 的高斯曲率为

$$K_{\mathbf{v}_i} = \frac{1}{A(\mathbf{v}_i)} \left(2\pi - \sum_{\mathbf{v}_{i,j} \in R(\mathbf{v}_i)} \alpha_{i,j} \right).$$

其中, $R(\mathbf{v}_i)$ 表示顶点 $\mathbf{v}_i$ 的一阶邻接顶点集合; $\alpha_{i,j}$ 表示第 j 个一阶邻接三角形中顶点 $\mathbf{v}_i$ 对应的角度; $A(\mathbf{v}_i)$ 表示由顶点 $\mathbf{v}_i$ 的一阶邻接边的垂直平分线相交所构成的多边形面积。

将曲率按绝对值大小排序后, 利用肘部法则自动选取肘部位置处的曲率, 并将大于肘部位置处曲率的顶点标记为大曲率顶点. 这里的肘部法则用于计算曲率的变化程度, 因此将肘部法则的

成本函数设定为曲率的绝对值, 肘部位置即为曲率绝对值下降幅度最大的位置. 由于网格上曲率较大的地方往往存在局部尖锐的变形, 这些变形反映了网格的细节信息, 因此大曲率顶点大概率是高频区域中的一部分. 基于此, 本文在聚类时将增加大曲率顶点的权重, 以增加这些顶点划入高频区域的概率.

在确定大曲率顶点并为其设置大权重后, 对面部网格顶点进行聚类, 自动确定高频区域的个数和每个区域的范围. 本文采用肘部法则自动选取最佳的聚类个数, 这里的肘部法则需要考虑类的畸变程度, 成本函数应该反映网格顶点与类中心的相关性. 传统的相关性衡量往往采用网格顶点与类中心之间的欧几里得距离, 但对于面部网格, 顶点之间的相关性不仅取决于两者之间的空间距离, 还应取决于两者在表情变化过程中的运动一致性. 显然, 变化过程中运动越一致的网格顶点, 它们的关系越密切, 它们属于同一个区域的概率就越高. 基于此, 本文借鉴文献[8]的思想, 提出新的时空相关性准则, 用于衡量网格顶点与类中心的密切程度, 并基于新的时空相关性准则改进 K-Means 聚类算法, 利用肘部法则和 K-Means 聚类算法自动确定高频区域的个数, 以及每个高频区域的范围.

本文提出的新的时空相关性准则在衡量顶点之间的关系密切程度时, 既考虑顶点之间的空间邻近性, 又考虑顶点在表情变化过程中的运动一致性. 但与文献[8]中采用的欧几里得距离不同, 本文采用测地线距离来计算顶点之间的空间邻近性. 测地线距离是沿着人脸网格表面测得的顶点之间的最短距离, 显然, 相比欧几里得距离, 它能更准确地反映网格曲面上的顶点在空间位置上的邻近关系. 新的时空相关性准则具体表示如下.

设输入的面部网格的已知训练样本共有 T 帧, 这些样本构成了该角色模型的表情动画序列. 不失一般性, 将输入的面部网格看作序列中的第 1 帧, 则对于该网格上的任意 2 个顶点 $\mathbf{v}_i$ 和 $\mathbf{v}_j$, 它们之间的测地线距离定义为

$$d(\mathbf{v}_i, \mathbf{v}_j) = \min L(o(\mathbf{v}_i, \mathbf{v}_j)).$$

其中, $o(\mathbf{v}_i, \mathbf{v}_j)$ 为顶点 $\mathbf{v}_i$ 与 $\mathbf{v}_j$ 的所有路径; $L(o(\mathbf{v}_i, \mathbf{v}_j))$ 表示测地线距离, 是相邻点之间的欧几里得距离之和. 将顶点 $\mathbf{v}_i$ 和 $\mathbf{v}_j$ 在动画序列中第 t 帧的相关程度值定义为

$$\begin{cases} z_{i,j}^t(\mathbf{v}_i, \mathbf{v}_j) = d_{i,j}^t(\mathbf{v}_i, \mathbf{v}_j) r_{i,j}^t(\mathbf{v}_i, \mathbf{v}_j) s_{i,j}^t(\mathbf{v}_i, \mathbf{v}_j), \\ t = 1, 2, \dots, T, i = 1, 2, \dots, n, j = 1, 2, \dots, n, i \neq j \\ d_{i,j}^t(\mathbf{v}_i, \mathbf{v}_j) = 1 - \frac{d(\mathbf{v}_i, \mathbf{v}_j)}{\max_{i,j} d(\mathbf{v}_i, \mathbf{v}_j)} \\ r_{i,j}^t(\mathbf{v}_i, \mathbf{v}_j) = \frac{1}{2} \left(1 + \frac{(\mathbf{v}_{i,t} - \mathbf{v}_i) \cdot (\mathbf{v}_{j,t} - \mathbf{v}_j)}{\|\mathbf{v}_{i,t} - \mathbf{v}_i\| \cdot \|\mathbf{v}_{j,t} - \mathbf{v}_j\|} \right) \\ s_{i,j}^t(\mathbf{v}_i, \mathbf{v}_j) = \min \left\{ \frac{\|\mathbf{v}_{i,t} - \mathbf{v}_i\|}{\|\mathbf{v}_{j,t} - \mathbf{v}_j\|}, \frac{\|\mathbf{v}_{j,t} - \mathbf{v}_j\|}{\|\mathbf{v}_{i,t} - \mathbf{v}_i\|} \right\} \end{cases} \quad (1)$$

其中, $\mathbf{v}_{i,t}$ 和 $\mathbf{v}_{j,t}$ 分别表示顶点 $\mathbf{v}_i$ 与 $\mathbf{v}_j$ 在表情动画序列的第 t 帧网格上的位置. $d_{i,j}^t(\mathbf{v}_i, \mathbf{v}_j)$, $r_{i,j}^t(\mathbf{v}_i, \mathbf{v}_j)$ 和 $s_{i,j}^t(\mathbf{v}_i, \mathbf{v}_j)$ 的取值范围均为 $[0, 1]$.

如图 2 所示, 在式(1)中, $d_{i,j}^t(\mathbf{v}_i, \mathbf{v}_j)$ 利用测地线距离计算顶点 $\mathbf{v}_i$ 与 $\mathbf{v}_j$ 之间的空间相关性, 其值越大, 表示其空间邻近性越高; $r_{i,j}^t(\mathbf{v}_i, \mathbf{v}_j)$ 利用顶点 $\mathbf{v}_i$ 与 $\mathbf{v}_j$ 在 2 帧之间的位移向量的夹角, 计算它们沿着表情动画序列在运动方向上的相近程度; $s_{i,j}^t(\mathbf{v}_i, \mathbf{v}_j)$ 利用顶点 $\mathbf{v}_i$ 与 $\mathbf{v}_j$ 在 2 帧之间的移动距离的比值, 计算它们移动速率的相近程度. $r_{i,j}^t(\mathbf{v}_i, \mathbf{v}_j)$ 和 $s_{i,j}^t(\mathbf{v}_i, \mathbf{v}_j)$ 体现了顶点 $\mathbf{v}_i$ 与 $\mathbf{v}_j$ 在角色模型发生表情变化时的运动一致性, 值越大, 表示其运动一致性越高.

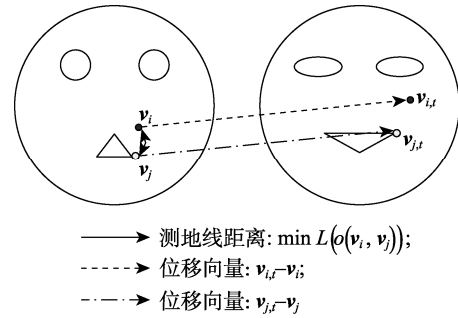


图 2 时空相关性示意图

利用式(1)计算得到顶点 $\mathbf{v}_i$ 与 $\mathbf{v}_j$ 在表情动画序列中每一帧的相关程度值后, 对所有值取平均, 得到顶点 $\mathbf{v}_i$ 与 $\mathbf{v}_j$ 的时空相关性系数

$$z(\mathbf{v}_i, \mathbf{v}_j) = \frac{1}{T-1} \sum_{t=2}^T z_{i,j}^t(\mathbf{v}_i, \mathbf{v}_j).$$

其中, $t = 1, 2, \dots, T; i = 1, 2, \dots, n; j = 1, 2, \dots, n; i \neq j$.

根据新的时空相关性准则求得网格上任意 2 个顶点之间的时空相关性系数之后, 对面部网格顶点执行聚类操作, 以获得高频区域. 本文基于时空

相关性准则改进 K-Means 聚类算法, 将顶点与聚类中心的时空相关性作为聚类过程中衡量顶点与聚类中心相关性的指标. 时空相关性系数越大, 表明顶点与聚类中心的相似度越高, 分配到该类中的可能性就越高. 该指标综合考量了网格顶点在整个动画序列上的空间邻近性和运动一致性, 可使得聚类结果更加准确.

将肘部法则与改进的 K-Means 聚类算法相结合, 以自动确定聚类的个数. 将肘部法则的成本函数定义为

$$E_{\text{elbow}} = \sum_{i=1}^k \sum_{v \in C_i} (1 - z(v, u_i)) \quad (2)$$

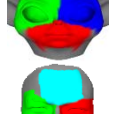
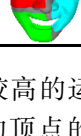
其中, $z(\cdot, \cdot)$ 为顶点之间的时空相关性系数; k 为聚类的个数; C_i 为第 i 个类; u_i 为第 i 个类的中心.

对于不同的聚类数目, 利用改进的 K-Means 聚类算法获得每个类的范围, 并计算式(2)中的成本函数值. 成本函数值下降幅度突然变小的位置即为肘部位置, 该位置处所对应的聚类数目即为高频区域的个数; 该数目下由改进 K-Means 聚类算法得到的聚类结果即为各个高频区域的范围. 聚类过程中, 本文限定了所形成的类的最大范围, 以避免全部或多余的顶点被划分到高频区域中, 进而保证高频区域的精确性. 这个最大范围值也可由用户指定. 本文实验中, 将每个类的最大边界范围限定为不超过中性面部网格宽度值的二分之一.

为了验证本文所提出的高频区域划分方法的合理性, 将其与使用 K-Means 聚类算法的高频区域划分方法进行了对比实验, 部分实验结果如表 1 所示.

从表 1 中可以看出, 基于 K-Means 聚类算法的高频区域划分属于硬性划分, 高频区域的个数是在划分前确定好的, 本实验中前 4 个模型的区域划分个数定义为 3, 第 5 个模型的区域划分个数定义为 4. 而本文提出的高频区域划分方法属于动态划分, 能够根据面部网格的顶点关系自适应地确定高频区域的个数和每个区域的范围. 另外, 本文方法不仅考虑了网格顶点的空间邻近性, 也考虑了顶点在整个表情序列上的运动一致性, 能够更加合理地确定高频区域的个数和范围. 例如, 表 1 中的 Zhang 角色模型, 在表情变化过程中, 其额头部

表 1 不同方法高频区域划分结果对比

角色模型名称	本文	K-Means 聚类
Ray ^①		
Yu ^②		
Monarudo ^③		
Yoda ^④		
Zhang		

分的顶点在整个表情序列上具有较高的运动一致性, 但是这些运动与左右眼部分的顶点的运动差异较大. K-Means 聚类算法划分高频区域时, 将额头的顶点错误地归类于左眼和右眼区域, 而本文方法可以将额头的顶点准确地归类于一个单独区域. 另外, 其他一些经验人脸区域划分方法一般也是将人脸划分为眼睛、鼻子、嘴巴等区域, 但往往只考虑了顶点之间的空间邻近性, 所以存在着与 K-Means 聚类算法划分高频区域方法相同的问题. 因此, 本文方法能够更准确、合理地划分出高频区域.

1.2 粗略编辑模块

进行表情编辑时, 在考虑面部表情精度的同时, 也应确保整个网络的运行速度. 本文方法中设计的粗略编辑模块用于估算粗颗粒的网格变形, 以使变形后的整个面部网格能大致逼近用户期望的基本表情. 该模块的网络结构简单, 参数量较少, 能够快速计算出网格变形后的顶点位置. 在此基础上再进行细颗粒的变形, 可使整个模型具有较好的运算效率. 该模块引入空间变化卷积层^[19]和采样残差层, 很好地克服了传统 CNN 无法直接处理不规则三维网格的缺陷, 因此能够捕获到三维网格上更多的特征信息, 且使用的网络结构相对简单, 故而可快速地逼近用户期望的基本表情.

① <https://www.cgtrarian.com/maya-character-rigs/download-free-3d-character-ray.html>

② <https://www.cgtrarian.com/maya-character-rigs/download-free-3d-character-yu-lon.html>

③ <https://www.character-rigs.com/monarudo-rig>

④ <https://www.turbosquid.com/3d-models/baby-yoda-model-1693451>

如图 3 所示, 粗略编辑模块由 2 个上采样残差块和 2 个下采样残差块组成. 该模块的输入是面部网格 I 与控制点约束 $\bar{P}$. 首先, 将控制点约束 $\bar{P}$ 施加到面部网格 I 的控制点上, 使其发生位移变化, 以引导面部网格根据控制点约束进行变形, 逼近符合用户要求的参考网格表情. 即将面部网格上的控制点由原始位置 p_i 移动到控制点约束所指定的新位置 p'_i , $i=1,2,\dots,m$. 然后, 对控制点位置已变化后的面部网格, 将其所有顶点位置 V 先经过 2 个空间变化卷积核半径为 4、步长为 4 的下采样残差块, 从 $\mathbb{R}^{n \times 3}$ 压缩至 $\mathbb{R}^{7 \times 9}$, n 是面部网格中的顶点数. 再经过 2 个相同核半径和步长的上采样残差

块, 恢复至网格顶点的原维度, 并得到各个顶点的新位置, 进而得到变形后的面部网格 I_c . 该粗颗粒的变形过程可表示为函数 $I_c = f_c(I, \bar{P}; \theta)$. 其中, f_c 表示由空间变化卷积和采样残差层组成的粗略编辑模块神经网络, 并将其参数化为 θ . 将面部网格 I 与控制点约束 $\bar{P}$ 输入到粗略编辑模块中, 通过训练模块参数 θ , 使网格顶点位置 V 发生变化, 逼近用户所期望的基本表情; 再根据输入面部网格 I 的拓扑结构即可得到变形后的面部网格 I_c . 本文网络学习的是从控制点约束到面部网格最终顶点位置的端到端映射, 故无需监督网格顶点在网络中的变化过程.

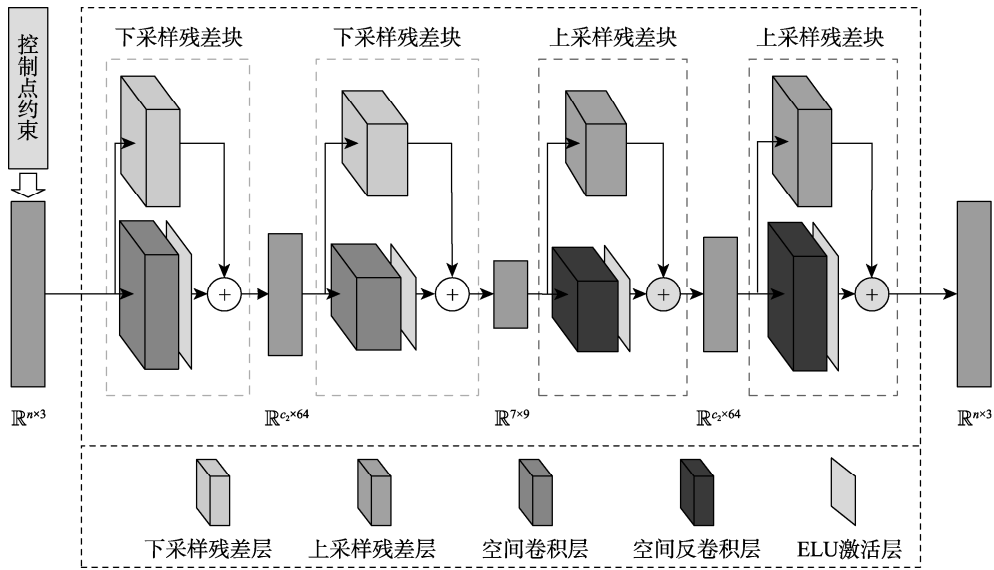


图 3 粗略编辑模块

三维面部网格的不规则性导致经典 CNN 无法对其进行直接处理, 并且 GCN 存在振荡问题以及过度平滑问题. 因此, 本文在粗略编辑模块的所有采样残差块中均引入空间变化卷积和采样残差层, 用以处理不规则的三维面部网格. 另外, 该卷积操作能够提取到面部网格中更多的空间特征信息, 因此可使得整个粗略编辑模块拥有较强的逼近能力. 如图 4 所示, 每个采样残差块均由空间变化卷积层、激活层 ELU 和采样残差层组合而成, 其处理过程是将输入网格分别通过空间变化卷积层和激活层 ELU, 以及采样残差层, 再将结果相加得到输出. 经典 CNN 对规则网格进行卷积操作时, 卷积核是结构规则且固定的, 对网格的采样权重也是相同且全局共享的, 这种全局共享的采样权重称为权重基. 然而, 面部三维网格是不规则的, 无法使用固定的卷积核进行卷积操作. 因此, 与经

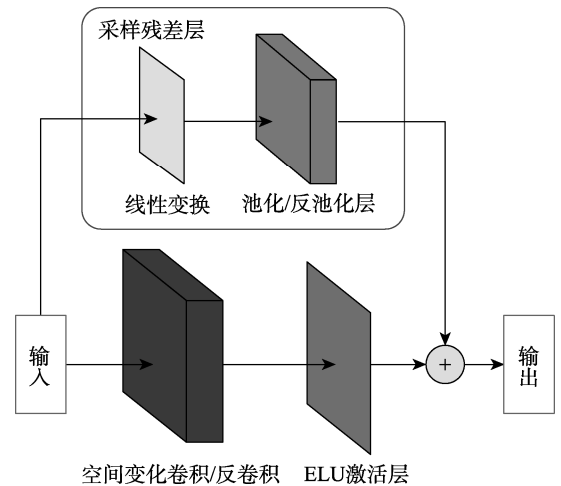


图 4 采样残差块

典卷积不同, 空间变化卷积对面部网格进行卷积操作时, 其卷积核是通过一组采样参数对权重基进

行线性采样实现的, 这组参数在不同的面部网格区域是不同的, 因此可很好地处理不规则的面部网格, 提取到更多的空间特征信息. 训练网络时, 需同时训练权重基和各个局部区域的采样参数. 设面部网格中, 利用空间变化卷积核进行特征提取的局部区域为 $\mathcal{N}$, $\mathbf{v}_i$ 是该局部区域 $\mathcal{N}$ 中的顶点位置, 则面部网格上该局部区域的空间变化卷积的特征提取函数为 $\mathbf{y} = \sum_{\mathbf{v}_i \in \mathcal{N}} \mathbf{W}_i \mathbf{v}_i + \mathbf{b}$. 其中, $\mathbf{y}$ 是该局部区域

的特征输出; $\mathbf{b}$ 是可学习的偏置向量; $\mathbf{W}_i$ 是面部网格上该局部区域中每个顶点的权重系数. $\mathbf{W}_i$ 由权重基 $\mathbf{B} = \{\mathbf{B}_\alpha\}_{\alpha=1}^\eta$, $\mathbf{B}_\alpha \in \mathbb{R}^{l \times o}$ 与顶点位置 $\mathbf{v}_i$ 在不同局部区域中的采样系数 $\mathbf{E}_i = \{e_{i,\alpha}\}_{\alpha=1}^\eta$, $e \in \mathbb{R}$ 两者的线性组合求得, 即 $\mathbf{W}_i = \sum_{\alpha=1}^\eta e_{i,\alpha} \mathbf{B}_\alpha$. 其中, η 表示面部网格中局部区域的个数; $\mathbf{B}$ 和 $\mathbf{E}_i$ 都是需要训练的参数.

由于面部网格的不规则性, 导致每个点的采样密度不同, 而传统的池化操作无法处理面部网格, 因此借助蒙特卡罗积分进行池化操作, 其操作函数为 $\mathbf{y} = \sum_{\mathbf{v}_i \in \mathcal{N}} \rho'_i \mathbf{v}_i$. 其中, $\rho'_i = |\rho_i| / \left| \sum_{\mathbf{v}_i \in \mathcal{N}} |\rho_i| \right|$ 为每个顶点的密度系数, ρ_i 是训练参数, 在整个数据集中共享. 类似地, 定义残差层为 $\mathbf{y} = \sum_{\mathbf{v}_i \in \mathcal{N}} \rho'_i \mathbf{G} \mathbf{v}_i$. 当

残差层的输入和输出特征维度相同时, $\mathbf{G}$ 是一个单位矩阵, 否则 $\mathbf{G}$ 是一个在整个面部网格顶点中共享并需要由训练得出的 $X \times Y$ 矩阵, 其中 X 为输入矩阵的维度, Y 为输出矩阵的维度. 基于上述操作设计实现的上采样和下采样残差块, 可很好地处理不规则的面部网格, 可通过整个粗略编辑模块得到较好的面部表情逼近效果.

1.3 精细编辑模块

粗略编辑模块可使面部网格快速地逼近用户期望的大致表情, 但难以精确地模拟皮肤皱纹等高频变形; 而精细编辑模块只关注对高频区域变形的逼近, 可很好地实现对表情细节变化的高精度重建. 两者结合, 可快速、准确地逼近用户期望的、真实自然且细节丰富的表情.

如图 5 所示, 精细编辑模块由 3 个下采样残差块和 3 个上采样残差块组成. 残差块采用与粗略编辑模块中的残差块一样的设计. 设面部网格 I 经高频区域划分后, 得到的高频区域集合为 $I_s = \{I_1, I_2, \dots, I_k\}$. 其中, k 为高频区域的个数;

I_s 中的顶点是整个网格 I 中顶点的子集. 面部网格的高频区域集合 I_s 与控制点约束 $\bar{\mathbf{P}}$ 为精细编辑模块的输入. 首先, 将控制点约束 $\bar{\mathbf{P}}$ 施加到高频区域集合 I_s 中的控制点上, 使其发生位移变化, 以进一步逼近参考网格高频区域中的表情细节信息. 即将面部网格高频区域集合中的控制点由原始位置 $\mathbf{p}_i$ 移动到控制点约束所指定的新位置 $\mathbf{p}'_i$, $i=1, 2, \dots, q$, $q < m$. 然后, 对于控制点位置已变化后的高频区域集合 I_s , 将其所有顶点经过 3 个空间变化卷积核半径为 4、步长为 3 的下采样残差块后, 从 $\mathbb{R}^{r \times 3}$ 压缩至 $\mathbb{R}^{7 \times 9}$, 其中 r 为 I_s 中的顶点数. 再经过 3 个相同核半径和步长的上采样残差块后, I_s 中的顶点恢复至原维度, 得到新的顶点位置, 进而得到变形后的高频区域. 该细颗粒的变形过程可表示为函数 $I_r = f_r(I_s, \bar{\mathbf{P}}; \theta_r)$. 其中, I_r 是变形后的高频区域集合; θ_r 是精细编辑模块的网络参数. 将高频区域集合 I_s 与控制点约束 $\bar{\mathbf{P}}$ 输入到精细编辑模块中, 通过训练网络参数 θ_r 来改变 I_s 中的顶点位置, 实现各个高频区域的精细变形, 以逼近用户期望的表情细节. 精细编辑模块神经网络的损失函数与粗略编辑模块相同. 这里要说明的是, 虽然精细编辑模块是将面部网格的所有高频区域合到一起进行细颗粒的变形处理, 但是, 利用本文提出的基于聚类思想的高频区域划分方法事先识别出面部网格多个高频区域的工作是具有重要意义的. 这是由于不同的角色模型其高频区域可能是不同的, 并且高频区域不一定为连续区域. 如表 1 中, Ray 和 Zhang 角色模型的高频区域是不连续的. 本文的高频区域划分方法不仅保证了高频区域的准确性, 而且每个高频区域仅包含时空相关性最密切的顶点, 其大小范围是最合适的, 因此区域之间可以不连续. 这避免了未经聚类处理所得的大片连续高频区域输入精细编辑模块导致计算复杂度增加的问题, 从而保证了整个网络模型的运行速度.

三维面部网格 I 经粗略编辑模块进行粗颗粒变形后, 得到含有基本表情的网格 I_c ; 网格 I 中的高频区域集合 I_s 通过精细编辑模块进行细颗粒变形后, 得到含有表情细节信息的高频区域 I_r ; 将两者的表情信息进行融合, 即可得到最终的满足用户期望的、含有表情细节信息的三维面部网格 I' . 具体地, 将网格 I_c 中高频区域内的顶点的位置替换为 I_r 中对应顶点的位置, 得到最终的各个

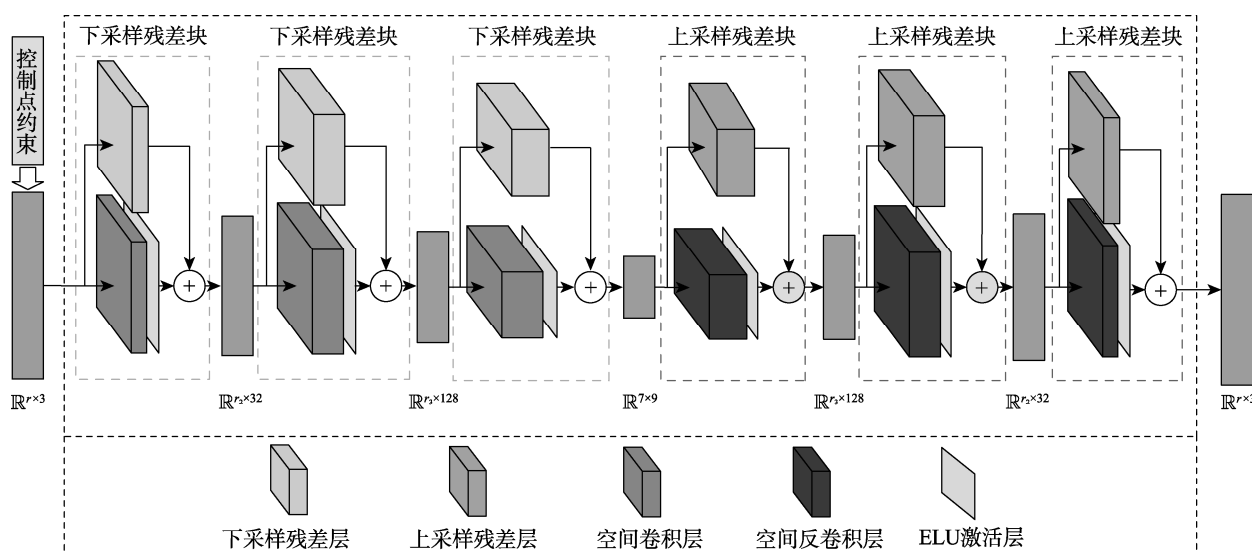


图 5 精细编辑模块

顶点的新位置 $\mathbf{V}'$, 进而得到最终变形后的三维面部网格 I' 。

表 2 展示了对表情进行多层次编辑的结果, 这里平均顶点误差定义为 2 个网格之间对应顶点位置的欧几里得距离平均值。从表 2 可以看出, 仅经过粗略编辑生成的网格具有与参考网格一致的基本表情, 但缺乏表情细节信息。而经过精细编辑后, 表情细节得到了很好的丰富, 且平均顶点误差显著降低。

表 2 表情的多层次编辑结果

中性 网格	参考 网格	粗略编辑		精细编辑	
		基本 表情	平均顶点 误差	最终 表情	平均顶点 误差
			0.47		0.29
			0.72		0.38

注: 基本表情由对中性网格进行粗略编辑后生成的, 对应的平均顶点误差是粗略编辑生成的网格与参考网格的; 最终表情是将精细编辑生成的表情细节与基本表情融合后得到的, 对应的平均顶点误差是融合后的网格与参考网格的。

1.4 损失函数

为提高模型的逼近精度, 本文提出一个新的损失函数用于求解最优网络参数。新损失函数既约束生成网格与参考网格之间的顶点位置的差异, 又约束两者顶点之间的顶点法向量的差异。显然, 对于 2 个网格, 如果它们对应顶点的位置相同, 对应顶点的法向量也相同, 则它们的形状, 即面部表

情一定相同。因此, 新损失函数定义为

$$L = L_{\text{pose}} + L_{\text{normal}}.$$

其中, $L_{\text{pose}} = \frac{1}{n} \sum_{i=1}^n \|\mathbf{v}_i - \mathbf{v}'_i\|_1$ 为约束参考网格与生成

网格顶点位置的损失函数; $L_{\text{normal}} = \frac{1}{n} \sum_{i=1}^n \|\mathbf{n}_i - \mathbf{n}'_i\|_1$

为约束两者顶点法向量的损失函数; $\mathbf{v}_i$ 为参考网格的顶点位置, $\mathbf{v}'_i$ 为生成网格的顶点位置; $\mathbf{n}_i$ 为参考网格的顶点法线向量, $\mathbf{n}'_i$ 为生成网格的顶点法线向量。本文使用 L_1 损失而不是 L_2 损失, 是因为 L_1 损失能产生鲜明的特征, 且鲁棒性更好。而顶点法向量损失的引入, 可进一步提升模型的精确度。另外, 由于本文的网络模型对面部高频区域进行单独处理, 在将精细编辑模块得到的高频区域变形信息融合到粗略编辑模块得到的整体变形后的网格上时, 会不可避免地出现边界和接缝问题, 而顶点法向量损失通过约束生成网格, 使其顶点法线与参考网格上对应顶点的法线尽量一致, 正好解决了这一问题。

2 实验及结果分析

2.1 实验环境及参数

本文训练粗略编辑模块中的模型参数 θ 与精细编辑模块中的模型参数 θ_r , 以使损失函数 L 最小化。模型使用 Adam 优化器进行训练, 批处理大小设置为 64。优化过程的学习率初始为 0.000 1, 每个周期将减小 1%。整个网络的前向和后向过程由 Pytorch 实现, 并在 GPU 上并行化运行。实验环

境为 Ubuntu 20.04, GeForce RTX 2080, CUDA 10.1 和 Pytorch 1.4.0.

2.2 数据集

本文采用 Ray, Monarudo, Yu 和 Yoda 4 个角色模型进行实验验证. 为构建角色模型的表情样本数据集, 使用乘性噪声增强角色模型的动画数据, 并应用数据平衡以防止某种常见表情(如中性表情)在动画数据中过度出现. 对每个角色模型, 均匀随机选取 h 个控制点, 指定每个控制点位移的变化范围, 使控制点在该范围中随机均匀变化, 得到控制点约束 $\bar{P}$. 该模型的数据集构建为

$$\mathcal{D} = \{I \odot \bar{P} | I \sim \bigcup (0.25, 3.0)^h, \bar{P}\}.$$

其中, $I \in \mathbb{R}^h$ 为一个随机值向量, $I \sim \bigcup (0.25, 3.0)^h$ 表示其 h 个分量在 $[0.25, 3.0]$ 范围内均匀随机变化; $\odot$ 表示向量的分量相乘. 将 I 与 $\bar{P}$ 的每一个随机相乘结果作为控制点约束输入 Maya, 可得到该随机约束条件下角色网格模型的变形, 即得到一个随机的表情样本. 这些随机的表情样本即构成一个动画序列. 这里, 2 个随机向量 I 与 $\bar{P}$ 的相乘可保证得到更多随机的控制点约束, 从而保证了所

生成的动画数据中表情的多样性. 在本文实验中, 角色模型 Ray, Monarudo, Yu 和 Yoda 的控制点个数 h 分别设置为 42, 29, 36 和 38.

上述表情数据集 $\mathcal{D}$ 是由角色模型的动画序列构成的, 因此某些表情可能比其他表情更频繁地出现. 利用这样的数据集训练网络模型, 会过度拟合频繁出现的表情, 使得在编辑其他表情时表现不佳. 为解决该问题, 受 Feng 等^[30]方法的启发, 本文将数据集划分到多个黑盒中, 随机均匀地选取一个黑盒, 再在黑盒内随机均匀地选取一个样本, 将这些样本构成最终的数据集, 以达到平衡数据集的目的.

本文利用上述操作分别为每个角色模型生成了包含 10 000 个表情样本的数据集. 由于每个角色的网格模型的复杂度不同, 因此在具体实验中, 又为每个模型从其数据集中均匀随机选择了不同数量的表情样本数据, 将其作为实验数据集. 其中, Ray, Yu, Monarudo 和 Yoda 模型分别选择了 3 000, 8 000, 5 000, 8 000 个表情样本. 每个角色的实验数据集均按照 8 : 1 : 1 的比例划分为训练、验证和测试数据. 图 6 展示了 Monarudo 实验数据集的部分表情样本.



图 6 Monarudo 数据集中的部分表情样本

2.3 消融实验

本文提出了基于多分支空间变化卷积网络的多层次表情编辑模型, 该模型包含粗略编辑模块和精细编辑模块 2 个分支, 在保证模型整体运行速度的同时, 又保证了模型的逼近精度. 为了验证本文网络中各分支的有效性, 设计以下分支消融实验: (1) 仅使用粗略编辑模块处理面部网格; (2) 仅使用高频区域划分方法和精细编辑模块处理面部网格; (3) 使用本文完整方法处理面部网格.

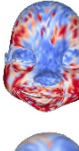
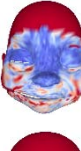
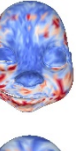
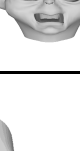
表 3 展示了对 4 个角色模型进行上述分支消融实验的定性和定量对比结果. 由第 4 列仅使用粗略编辑模块得到的结果可见, 角色模型的部分细节信息丢失, 如 Ray 模型眉头部分的细节, Yu 模型嘴角部分的细节等; 第 5 列的热力图更加直观地展示了粗略编辑的结果与参考网格之间的误差, 其中, 红色越深表示参考网格与生成网格之间的误差越大, 蓝色越深表示误差越小. 从表 3 可见, Ray, Yu 和 Yoda 角色模型存在着大量的深红色区域; 第 6

列为参考网格与粗略编辑生成的网格之间的平均顶点误差. 第 7~9 列分别为仅使用高频区域划分和精细编辑模块得到的结果、结果的热力图 and 平均顶点误差, 可见生成网格与参考网格之间存在着较大误差. 第 10~12 列分别为本文完整方法得到的结果、结果的热力图 and 平均顶点误差, 可见, 本文完整方法可以得到高精度的、细节丰富且真实自然的三维人脸表情.

为了验证本文高频区域划分方法对本文整体模型的增益, 分别其和 K-Means 聚类算法的高频区域划分方法作为高频区划分模块, 训练生成三维面部表情, 并进行对比实验.

图 7 展示了对 Ray 和 Yoda 角色模型采用不同的高频区域划分方法训练网络, 并生成的新表情对比结果. 从第 1~2 行中可以看出, 本文方法很好地生成了 Ray 角色模型眼角处的皱纹, 而使用 K-Means 聚类算法的高频区域划分方法进行表情编辑时, 却丢失了这一细节特征. 其中, 细节放大

表 3 分支网络消融实验

角色模型名称	中性网格	参考网格	粗略编辑			高频区域划分+精细编辑			本文		
			结果表情	热力图	平均顶点误差	结果表情	热力图	平均顶点误差	结果表情	热力图	平均顶点误差
Ray					0.56			1.22			0.27
Yu					5.12			9.25			2.51
Monarudo					0.49			3.83			0.31
Yoda					7.03			8.16			4.12

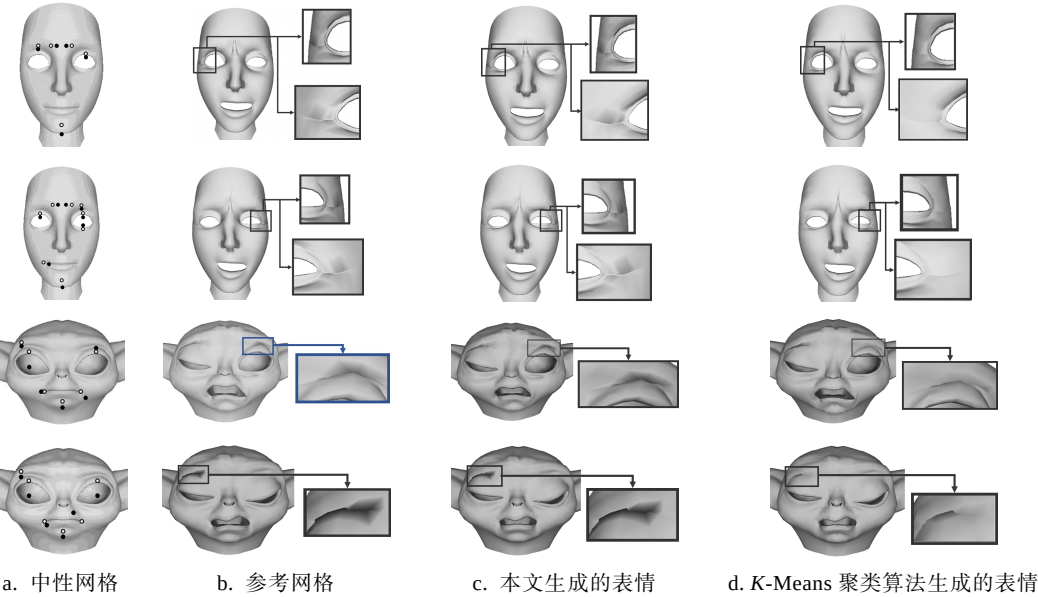


图 7 使用不同高频区域划分方法生成表情对比实验

框对同一细节分别从正面和侧面角度进行了展示. 对于第 3~4 行的 Yoda 角色模型的眉头处皱纹, 本文方法能够更精确地生成该细节, 细节放大框对该细节进行了展示. 本文分别对 Ray 和 Yoda 角色模型编辑生成了 200 个新表情, 表 4 展示了采用 2 种不同高频区域划分方法得到的所有新表情网格与参考网格的顶点误差的平均值. 从图 7 和表 4 可见, 本文提出的高频区域划分方法可以使整个网络生成的表情更加精细、真实.

表 4 用不同高频区域划分方法生成表情的平均顶点误差

角色模型名称	本文	K-Means 聚类
Ray	0.35	0.42
Yoda	2.79	3.13

本文还提出了新的损失函数用于模型训练, 函数 L_{pose} 不仅约束生成网格与参考网格的顶点位置, 而且函数 L_{normal} 同时约束顶点法向量. 为验证

顶点法向量约束的引入对提升模型精度的有益作用, 本文设计如下约束项消融实验. 对不同的角色模型, 分别采用仅利用顶点位置约束以及同时利用顶点位置约束和顶点法向量约束作为损失函数训练网络模型进行实验, 比较角色模型的测试集在 2 种不同情况下生成网格和参考网格之间的误差的平均值.

表 5 展示了对各角色模型进行约束项消融实验的误差比较结果. 从表 5 可见, 顶点法向量约束的引入明显降低了误差, 有效地提高了网格逼近的精度. 图 8 展示了约束项消融实验的一个直观视觉效果, 可以看出, 未采用顶点法向量约束时, 粗略编辑生成的网格与精细编辑生成的网格在合成时产生了明显的接缝, 而采用顶点法向量约束后, 很好地避免了接缝问题.

表 5 不同损失函数约束项消融实验的平均顶点误差比较结果

角色模型名称	损失函数	
	L_{pose}	$L_{pose} + L_{normal}$
Ray	0.45	0.37
Yu	2.79	2.55
Monarudo	0.48	0.38
Yoda	2.92	2.77

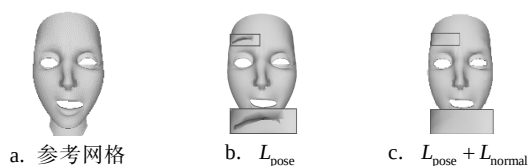


图 8 接缝问题示意图

2.4 对比实验结果

本文的网络模型一旦按照第 2.1 节所述步骤训练完成, 即可允许用户在角色模型上自由选择控制点, 并通过改变控制点的位置, 使角色模型发生表情变化, 得到用户期望的表情. 图 9 展示了对不同角色模型, 用户选择控制点并编辑得到的部分新表情; 其中白色点为用户选择的控制点, 黑色点为用户将控制点移动到的新位置. 由图 9 可以看出, 本文方法在表情编辑中取得了较好的结果, 生成网格不仅与参考网格的表情相同, 而且具有丰富的细节特征.

为进一步验证本文方法的有效性, 将本文的网络模型与 Bailey 等^[15]提出的网络模型以及 LBS 模型^[2]进行了比较. Bailey 等^[15]的方法使用经典 CNN 搭建的网络模型, 其根据逼近误差划分高频

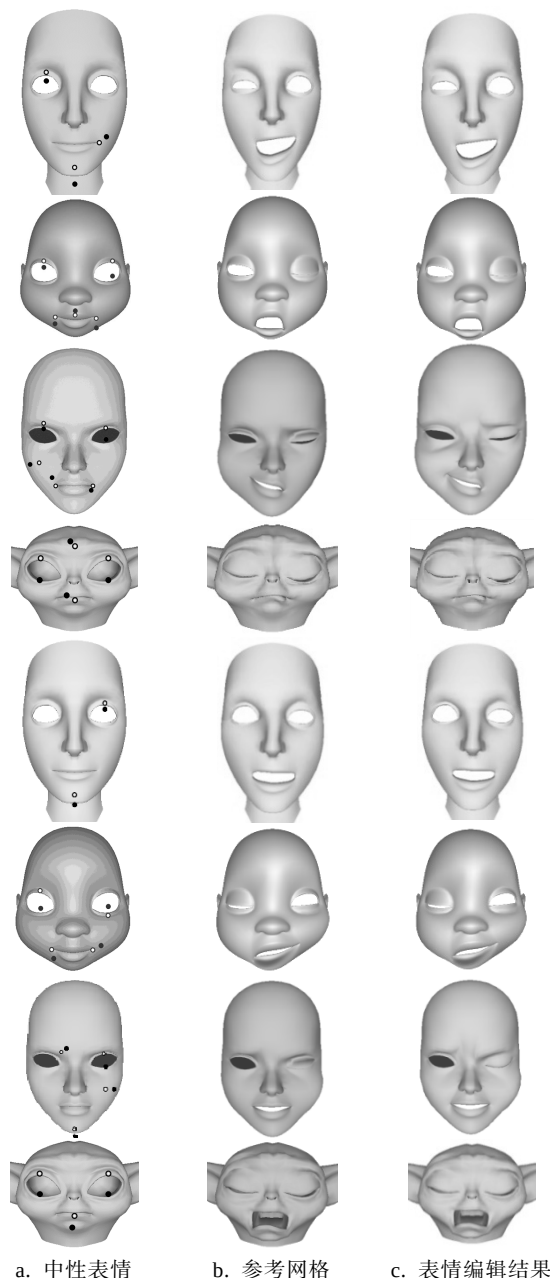


图 9 本文方法的部分表情编辑结果

区域, 并对三维网格的 UV 图进行处理; 采用与本文方法相同的数据集, 按照 Bailey 等^[15]方法的训练策略训练其网络模型. LBS 方法是对面部网格设置骨骼和顶点权重以引导网格变形, 本文使用 Le 等^[5]的方法为 LBS 模型生成骨骼和顶点权重, 共生成 16 块骨骼, 每个顶点赋予其 8 个非零权重.

图 10~图 13 展示了对各个角色模型分别使用本文方法、Bailey 等^[15]方法和 LBS 方法产生的部分表情编辑结果. 4 个角色模型中各个表情的初始表情均为中性表情, 然后通过变换不同控制点的位置生成不同的表情. 在图 10 的 Ray 模型表情生成结果中, LBS 方法得到的生成网格丢失了很多

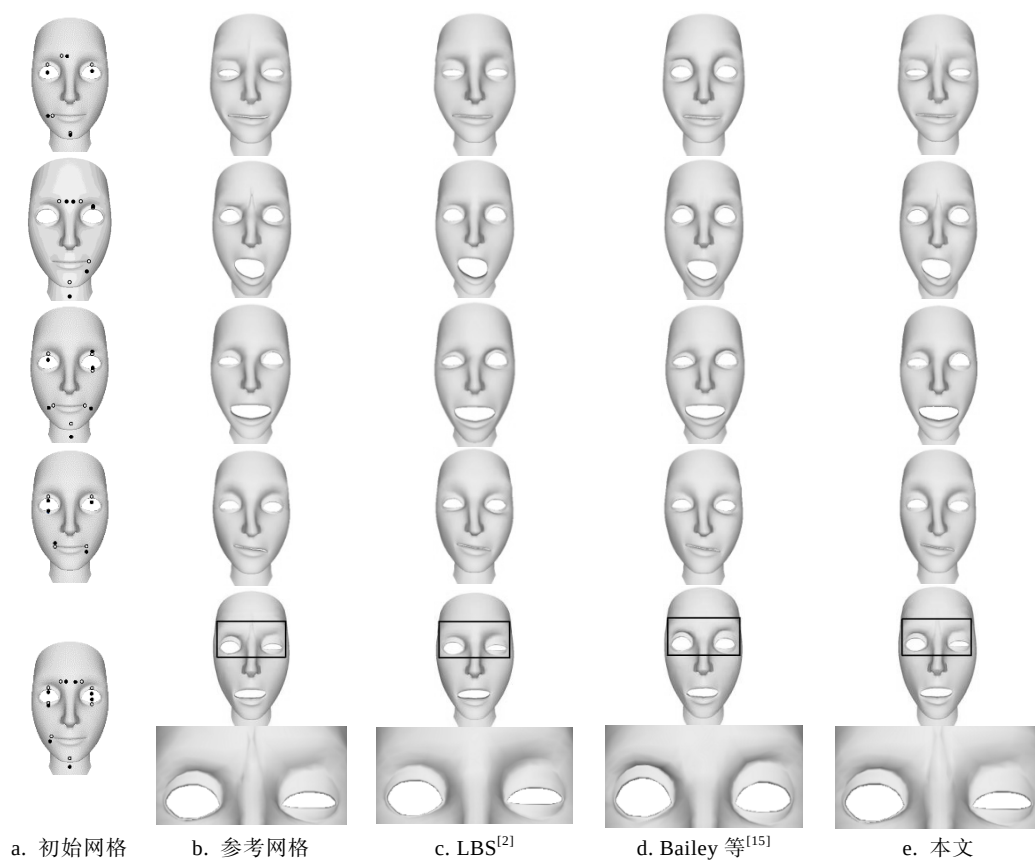


图 10 不同方法对 Ray 角色模型的面部表情编辑效果对比图

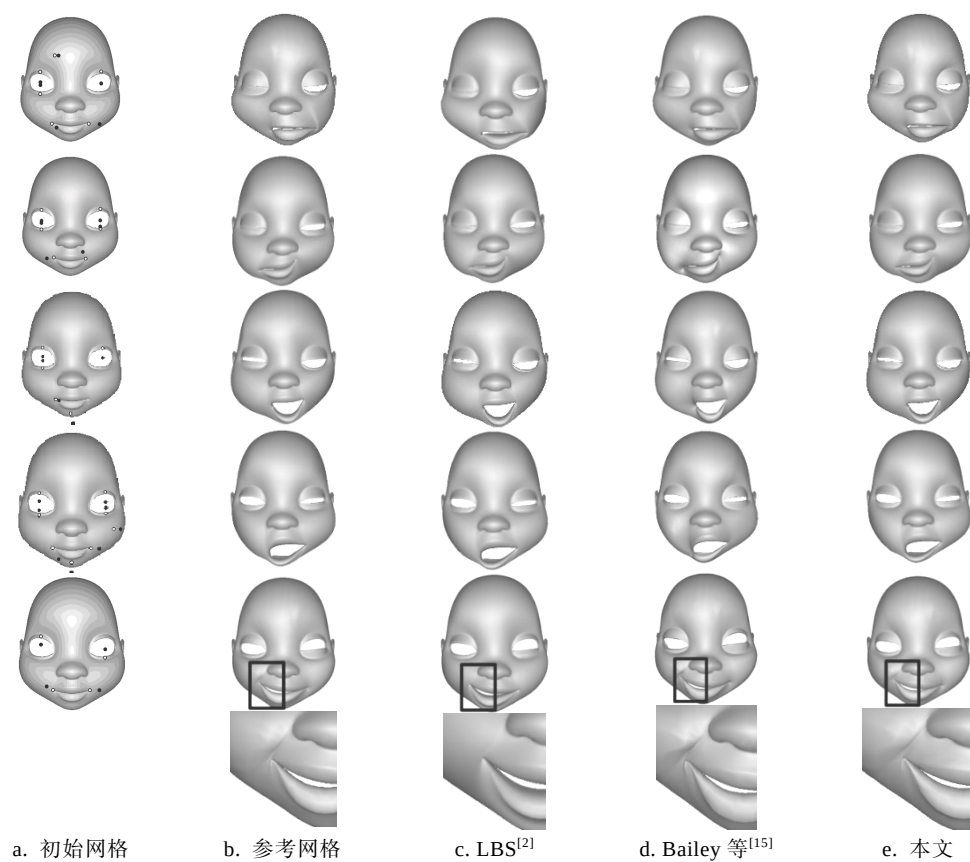


图 11 不同方法对 Yu 角色模型的面部表情编辑效果对比图

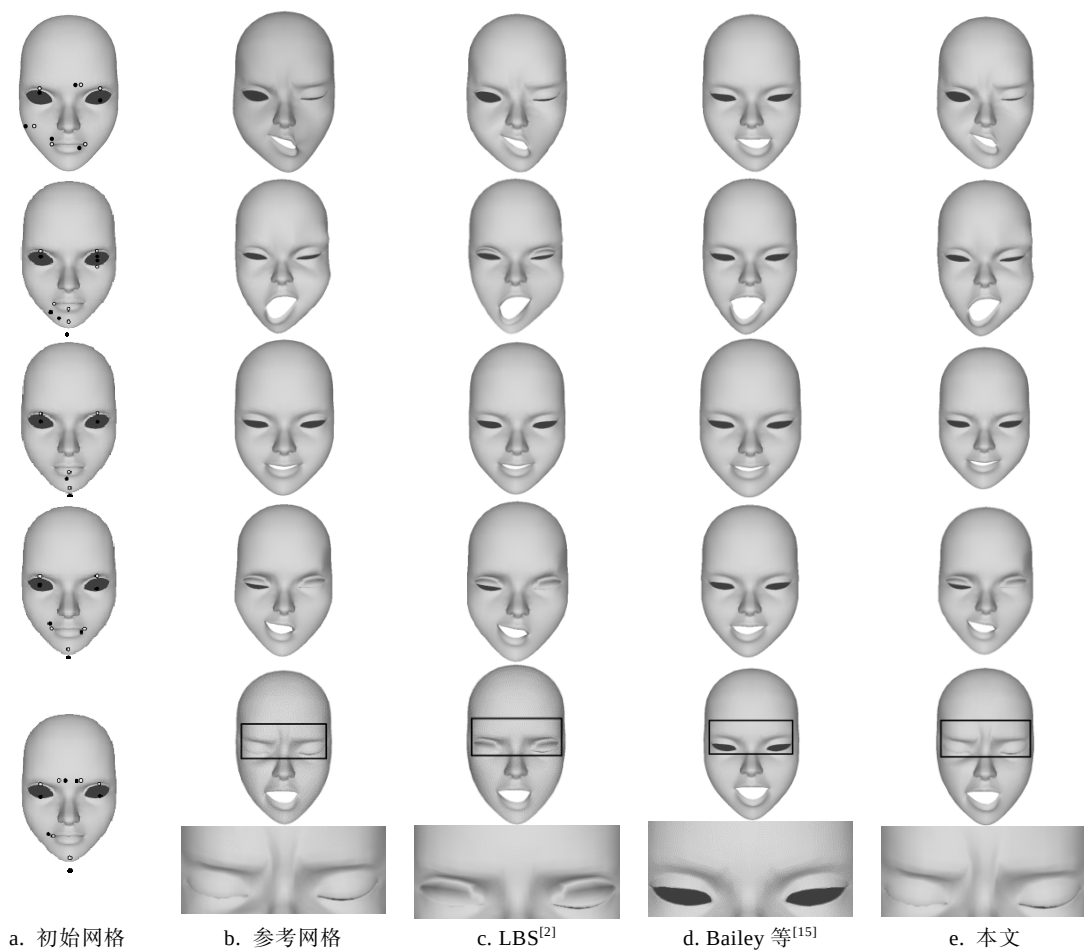


图 12 不同方法对 Monaruda 角色模型的面部表情编辑效果对比图

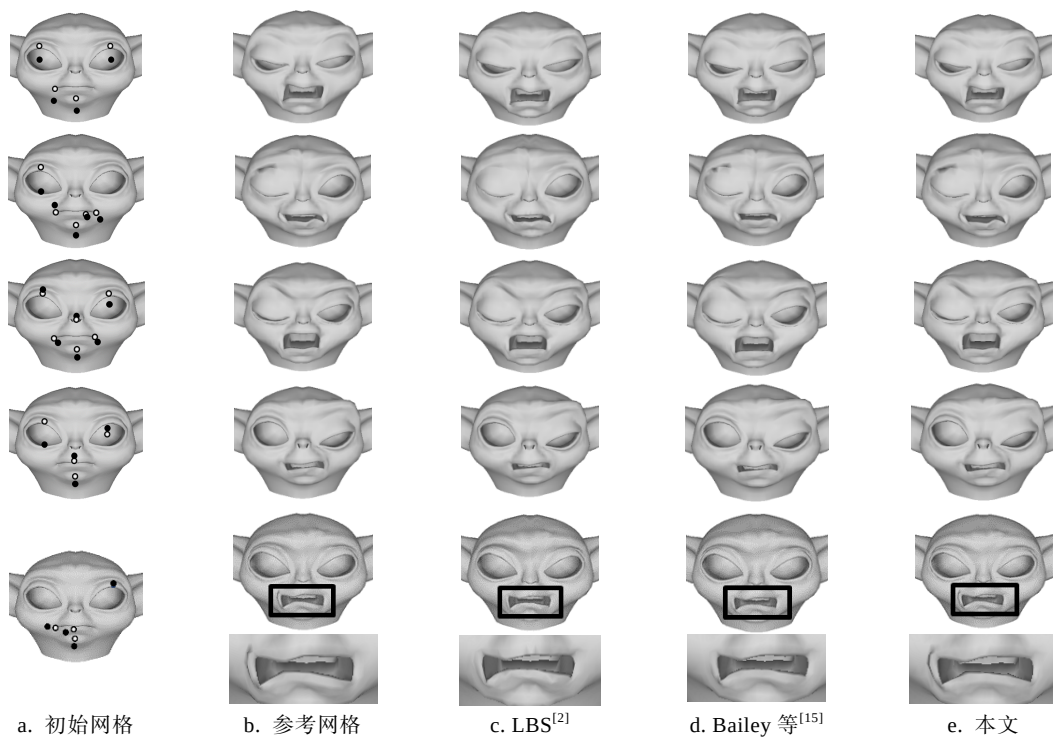


图 13 不同方法对 Yoda 角色模型的面部表情编辑效果对比图

表情细节, 如第 1 个、第 2 个和第 5 个表情眉头的皱纹部分; Bailey 等^[15]方法生成的网格存在较大的误差, 如第 1 个和第 5 个表情模型的眼睛闭合程度较小, 第 1 个、第 2 个和第 5 个表情的眉头皱纹丢失, 与参考网格差异较大; 而本文方法生成的表情与参考网格极其相近, 且具有丰富的细节. 在图 11 的 Yu 模型表情生成结果中, 使用 LBS 方法逼近的表情细节仍有所缺失, 如第 5 个表情右嘴角的皱纹; Bailey 等^[15]方法使部分生成网格发生了扭曲变形, 如第 3 个表情眉头的凹陷部分, 第 2 个表情的左嘴角和第 5 个表情的右嘴角的折痕部分. 在图 12 的 Monarudo 模型表情生成结果中, Bailey 等^[15]方法的整体效果较差, 只有嘴巴部分产生了微小的形变, 眼睛区域几乎没有变化; LBS 方法生成的表情仍然存在部分细节信息的丢失, 如第 5 个表情, 其眉头皱纹丢失, 且眼睛的变形扭曲. 在图 13 的 Yoda 模型表情生成结果中, 3 种方法虽然整体都取得了较好的效果, 但是 LBS 方法与 Bailey 等^[15]方法在对

一些表情细节的逼近中仍存在不准确的情况, 如第 4 个和第 5 个表情的嘴角皱纹, 第 2 个表情中的右眼皮上方和嘴角皱纹. 为更好地呈现细节差异, 对部分表情的细节区域进行了放大处理(如矩形框所示), 从细节图可以看出, 本文方法所生成的新表情与参考表情更接近, 细节特征更精确. 综上, 本文方法可编辑生成高精度的表情模型, 在生成真实自然的表情的同时, 也生成了精细的表情细节.

为了更直观地展示不同表情编辑方法在生成面部表情精度上的差异, 本文将各角色模型的表情生成网格与参考网格的对应顶点位置的欧几里得距离作为误差, 制作误差热力图. 图 14 给出了对每个模型的中性表情, 分别利用 3 种方法调整控制点所产生的新表情网格与参考网格的误差热力图. 由图 14 可见, 本文方法生成的网格, 其红色区域明显少于其他方法, 说明本文方法编辑得到的表情能够更加精确地逼近用户所期望的表情.

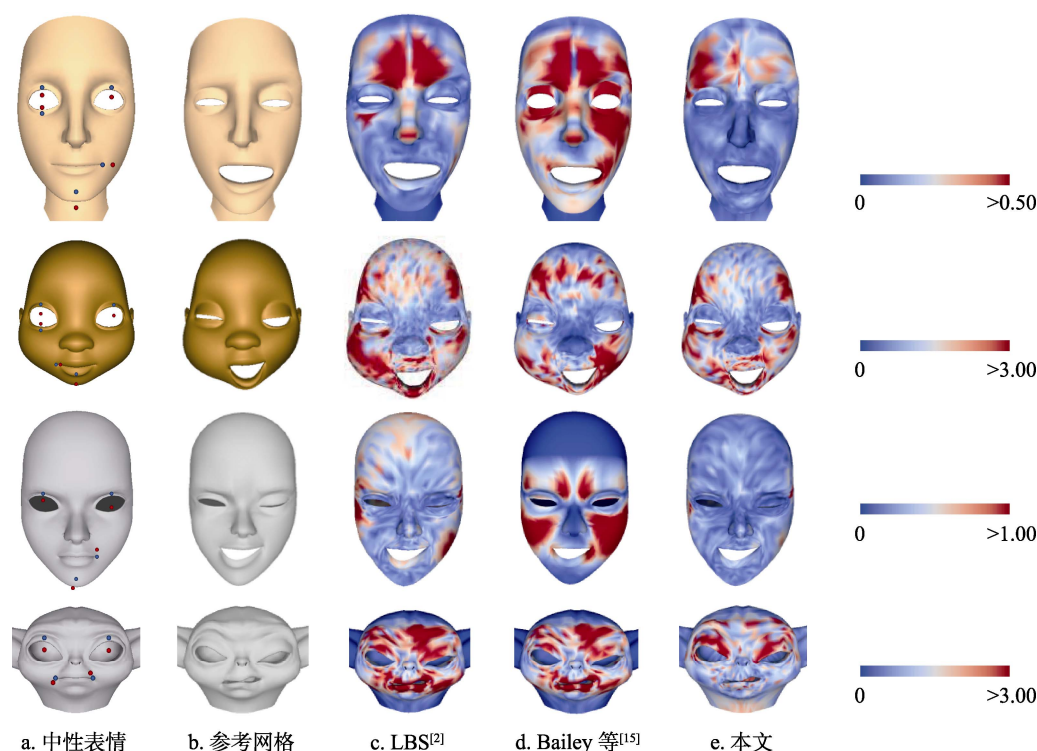


图 14 不同方法的表情编辑结果的热力图

表 6 中列出了 Ray, Yu, Monarudo 和 Yoda 这 4 个角色模型在测试集上得到的生成网格和参考网格之间顶点误差的平均值. 从表 6 可以看出, 对所有角色模型, 本文方法与 LBS 和 Bailey 等^[15]方法相比, 误差都有较大程度的降低, 表明本文方法所生成的变形网格与参考网格更接近, 逼近精度更高.

表 6 不同方法平均顶点位置误差

角色模型名称	LBS ^[2]	Bailey 等 ^[15]	本文
Ray	0.64	0.85	0.37
Yu	5.21	3.84	2.55
Monarudo	0.95	3.94	0.38
Yoda	7.34	5.11	2.77

图 15 展示了利用本文方法从中性表情编辑生成任意表情的过程. 这些表情没有参考网格, 是根据控制点约束生成的全新表情. 从图 15 可以看出, 本文可以通过逐步改变控制点的位置(白色点为控制点的原位置, 黑色点为控制点移动到的新位置), 使面部网格逐步变形, 生成连续的面部表情. 因此, 本文方法可为用户提供直观、简便且快捷的表情编辑工具, 生成满足用户期望的面部表情, 且表情真实自然, 含有丰富的细节特征.

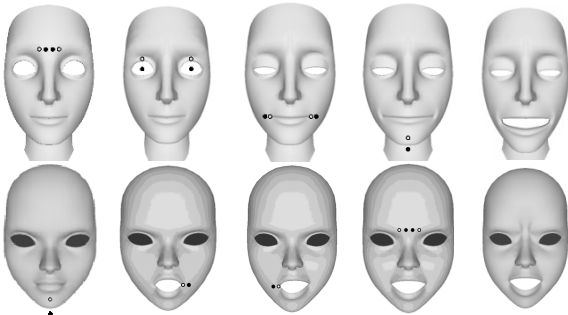


图 15 改变控制点位置生成连续表情效果图

2.5 运行时间测试

本文所有的实验均在配置为 Intel Xeon Bronze 3106 和 NVIDIA GeForce RTX 2080 的机器上运行. 为了进一步评估本文方法的性能, 将其与 LBS 和 Bailey 等^[15]方法在 4 个角色模型上进行了运行时间的对比, 如表 7 所示. 这里的运行时间是对每个模型计算了连续变换生成 300 个表情所用的时间. 从表 7 中可以看出, 本文方法与其他方法相比, 在运行速度上略有优势; 与速度提升相比, 更重要的是本文方法所生成的三维人脸表情更加准确, 并富含精细的细节特征.

表 7 不同方法在 4 个角色模型上的运行时间 s

角色模型名称	LBS ^[2]	Bailey 等 ^[15]	本文
Ray	64.12	48.13	43.34
Yu	93.26	84.51	80.64
Monarudo	39.18	24.91	23.77
Yoda	59.32	47.68	46.14

3 结 语

本文的主要贡献如下.

(1) 提出一种用于三维人脸表情编辑的多分支空间变化卷积网络. 对人脸的基本表情和表情细节(即面部网格的高频区域)采用不同的分支网络模块进行处理, 粗略编辑模块用于生成基本表

情, 精细编辑模块用于丰富和完善表情细节, 两者结合, 既保证了表情编辑的精度, 又保证了模型的运行效率. 另外, 将空间变化卷积引入分支网络结构, 可直接提取三维网格的空间特征, 大大提高了面部表情编辑的准确性.

(2) 提出一种面部高频区域识别方法. 用高斯曲率和肘部法则自动地筛选出面部网格上的大曲率顶点, 并使用基于新的时空相关性准则的改进 K-Means 聚类算法和肘部法则, 自动地确定高频区域的个数和每个区域的范围. 顶点时空相关性这一聚类准则的提出, 可在整个表情序列上从空间邻近性和运动一致性 2 个方面综合衡量顶点之间的相关性, 极大地提高了区域划分的合理性和准确性.

(3) 提出一个新的损失函数, 用于同时约束顶点位置和顶点法向量. 该损失函数可有效地提高整个模型的精度, 且解决了高频区域融合时的接缝问题.

传统的三维人脸表情编辑方法多是借助 Maya 等大型软件实现, 对操作人员的专业要求较高, 需要熟悉面部生理结构, 且需要对模型搭建骨骼并调整骨骼生成新表情, 该过程操作复杂, 且在变形的过程中容易产生穿模等问题. 本文方法无需搭建骨骼, 只需选择少量控制点并拖动控制点, 即可很好地生成满足用户编辑要求、真实自然且细节丰富的表情, 降低了操作复杂度, 对普通用户友好, 同时方法可避免穿模问题. 但是, 本文方法在泛化能力上具有一定的局限性, 对具有不同拓扑结构的角色模型需要进行重新训练. 如何提高网络模型的泛化能力, 是本文今后的研究重点.

参考文献(References):

- [1] Zou Bei, Peng Yongjin, Wu Lihua, et al. Facial expression animation based on physical model[J]. Chinese Journal of Computers, 2002, 25(3): 331-336(in Chinese)
(邹北骥, 彭永进, 伍立华, 等. 基于物理模型的人脸表情动画技术研究[J]. 计算机学报, 2002, 25(3): 331-336)
- [2] Magnenat-Thalmann N, Laperrière R, Thalmann D. Joint-dependent local deformations for hand animation and object grasping[C] //Proceedings of the Graphics Interface '88. Toronto: Canadian Information Processing Society, 1989: 26-33
- [3] Kavan L, Žára J. Spherical blend skinning: a real-time deformation of articulated models[C] //Proceedings of the Symposium on Interactive 3D Graphics and Games. New York: ACM Press, 2005: 9-16
- [4] Kavan L, Collins S, Žára J, et al. Skinning with dual quaternions[C] //Proceedings of the Symposium on Interactive 3D

- Graphics and Games. New York: ACM Press, 2007: 39-46
- [5] Le B H, Deng Z G. Smooth skinning decomposition with rigid bones[J]. ACM Transactions on Graphics, 2012, 31(6): Article No.199
 - [6] Feng W W, Kim B U, Yu Y Z. Real-time data driven deformation using kernel canonical correlation analysis[J]. ACM Transactions on Graphics, 2008, 27(3): 1-9
 - [7] Lewis J P, Anjyo K, Rhee T, *et al.* Practice and theory of blendshape facial models[C] //Proceedings of the EUROGRAPHICS (State of the Art Reports). Eindhoven: The Eurographics Association, 2014: 199-218
 - [8] Chi J, Gao S S, Zhang C M. Interactive facial expression editing based on spatio-temporal coherency[J]. The Visual Computer, 2017, 33(6-8): 981-991
 - [9] Lin J K, Yuan Y, Shao T J, *et al.* Towards high-fidelity 3D face reconstruction from in-the-wild images using graph convolutional networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 5891-5900
 - [10] Feng Y, Feng H W, Black M J, *et al.* Learning an animatable detailed 3D face model from in-the-wild images[J]. ACM Transactions on Graphics, 2021, 40(4): Article No.88
 - [11] Koujan M R, Roussos A, Zafeiriou S. DeepFaceFlow: in-the-wild dense 3D facial motion estimation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 6618-6627
 - [12] Cao C, Agrawal V, De La Torre F, *et al.* Real-time 3D neural facial animation from binocular video[J]. ACM Transactions on Graphics, 2021, 40(4): Article No.87
 - [13] Song S L, Shi W Q, Reed M. Accurate face rig approximation with deep differential subspace reconstruction[J]. ACM Transactions on Graphics, 2020, 39(4): Article No.34
 - [14] Wang M J, Bradley D, Zafeiriou S, *et al.* Facial expression synthesis using a global-local multilinear framework[J]. Computer Graphics Forum, 2020, 39(2): 235-245
 - [15] Bailey S W, Omens D, Dilorenzo P, *et al.* Fast and deep facial deformations[J]. ACM Transactions on Graphics, 2020, 39(4): Article No.94
 - [16] Litany O, Bronstein A, Bronstein M, *et al.* Deformable shape completion with graph convolutional autoencoders[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 1886-1895
 - [17] Ranjan A, Bolkart T, Sanyal S, *et al.* Generating 3D faces using convolutional mesh autoencoders[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 704-720
 - [18] Li K, Liu J Y, Lai Y K, *et al.* Generating 3D faces using multi-column graph convolutional networks[J]. Computer Graphics Forum, 2019, 38(7): 215-224
 - [19] Zhou Y, Wu C L, Li Z M, *et al.* Fully convolutional mesh autoencoder using efficient spatially varying kernels[C] //Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2020: 9251-9262
 - [20] Shi Y C, Aggarwal D, Jain A K. Lifting 2D StyleGAN for 3D-aware face generation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 6258-6266
 - [21] Othbertout N, Ferrari C, Daoudi M, *et al.* Sparse to dense dynamic 3D facial expression generation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 20385-20394
 - [22] Tran L, Liu X M. On learning 3D face morphable model from in-the-wild images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 157-171
 - [23] Yan Yanfu, Lyu Ke, Xue Jian, *et al.* Facial animation method based on deep learning and expression AU parameters[J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(11): 1973-1980(in Chinese)
(闫衍芙, 吕科, 薛健, 等. 基于深度学习和表情 AU 参数的人脸动画方法[J]. 计算机辅助设计与图形学学报, 2019, 31(11): 1973-1980)
 - [24] Zhang Ruisi, Pan Ye. Stylized avatar animation based on deep learning[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(5): 675-682(in Chinese)
(张睿思, 潘烨. 用于角色表情动画生成的深度学习技术[J]. 计算机辅助设计与图形学学报, 2022, 34(5): 675-682)
 - [25] Yu Zhiping, Chi Jing, Ye Yanan, *et al.* Detailed features-preserving 3D facial expression transfer[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(2): 186-198(in Chinese)
(于志平, 迟静, 叶亚男, 等. 细节特征保持的三维面部表情迁移方法[J]. 计算机辅助设计与图形学学报, 2021, 33(2): 186-198)
 - [26] Deng Q X, Ma L M, Jin A B, *et al.* Plausible 3D face wrinkle generation using variational autoencoders[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(9): 3113-3125
 - [27] Gecer B, Ploumpis S, Kotsia I, *et al.* Fast-GANFIT: generative adversarial network for high fidelity 3D face reconstruction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(9): 4879-4893
 - [28] Cheng Xuan, Liu Renshuai, Zheng Yinglin, *et al.* Collaborative completion of facial texture and displacement in UV space[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(1): 113-120(in Chinese)
(程轩, 刘仁帅, 郑英林, 等. UV 空间中的人脸颜色纹理和几何细节协同补全[J]. 计算机辅助设计与图形学学报, 2022, 34(1): 113-120)
 - [29] Wan Xianmei, Jin Xiaogang. Spacetime facial animation editing[J]. Journal of Computer-Aided Design & Computer Graphics, 2013, 25(8): 1183-1189(in Chinese)
(万贤美, 金小刚. 时空人脸表情动画编辑[J]. 计算机辅助设计与图形学学报, 2013, 25(8): 1183-1189)
 - [30] Feng Z H, Kittler J, Awais M, *et al.* Wing loss for robust facial landmark localisation with convolutional neural networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 2235-2245