计算机科学
COMPUTER SCIENCE

基于区域显著性与空间特征提取的说话人像合成方法

王邢波, 张浩, 高浩, 翟明亮, 谢九成

引用本文

王邢波, 张浩, 高浩, 翟明亮, 谢九成. 基于区域显著性与空间特征提取的说话人像合成方法[J]. 计算机科学, 2025, 52(3): 58-67.

WANG Xingbo, ZHANG Hao, GAO Hao, ZHAI Mingliang, XIE Jiucheng. Talking Portrait Synthesis Method Based on Regional Saliency and Spatial Feature Extraction [J]. Computer Science, 2025, 52(3): 58-67.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240300030

基于区域显著性与空间特征提取的说话人像合成方法

王邢波　张　浩　高　浩　翟明亮　谢九成
南京邮电大学自动化学院、人工智能学院　南京 210023
(sinbowang@163.com)

摘　要　音频驱动的说话人像合成技术致力于将任意的输入音频序列转换为逼真的说话人像视频。近期,基于神经辐射场(NeRF)的多个说话人像合成工作取得了优秀的视觉效果。但是,此类工作仍普遍存在着语音-嘴唇同步欠佳、躯干抖动和合成视频清晰度较低等不足。为了解决上述问题,提出了一种基于区域显著特征和空间体积特征的高保真说话人像合成方法。具体而言,一方面,开发了一个区域显著性感知模块用于头部建模。它利用多模态输入信息动态调整头部空间点的体积特征,同时优化基于哈希表的特征存储,从而提高面部细节表征的精确度和渲染效率。另一方面,设计了一个空间特征提取模块用于躯干的独立建模。不同于现有方法普遍采用的直接基于躯干表面空间点坐标估计其颜色和密度的方式,该模块利用参考图像构建躯干场以提供对应的纹理和几何先验,从而实现更清晰的躯干渲染和自然的躯干运动。应用于多个人物主体的实验结果表明,在自我重建场景中,所提方法相较于当前最优的基线模型,在图像质量上(PSNR,LPIPS,FID,LMD)分别取得了 10.15%,12.12%,0.77%和 1.09%的提升,在嘴唇同步精度上(AUE)提高了 14.20%。此外,在交叉驱动(使用非训练集音频)的场景下,该算法在嘴唇同步精度(AUE)上提升了 4.74%。

关键词:说话人像合成;三维重建;音视频同步;神经辐射场;注意力机制
中图分类号　TP391

Talking Portrait Synthesis Method Based on Regional Saliency and Spatial Feature Extraction

WANG Xingbo,ZHANG Hao,GAO Hao,ZHAI Mingliang and XIE Jiucheng
College of Automation & College of Artificial Intelligence,Nanjing University of Posts and Telecommunications,Nanjing 210023,China

Abstract　Audio-driven talking portraits synthesis endeavors to convert arbitrary input audio sequences into realistic talking portrait videos. Recently,several works on synthesizing talking portraits leveraging neural radiance fields(NeRF) have achieved superior visual results. However,such works still generally suffer from poor audio-lip synchronization,torso jitter,and low clarity in the synthesized videos. To address these issues,a method based on regional saliency features and spatial volume features is proposed to achieve high-fidelity synthesis of talking portraits. On one hand,a regional saliency-aware module is developed,dynamically adjusting the volumetric attributes of spatial points in the head region with multimodal input data and optimizing feature storage through hash tables,thus improving the precision and efficiency of facial detail representation. On the other hand,a spatial feature extraction module is designed for independent torso modeling. Unlike conventional methods that estimate color and density directly from torso surface spatial points,this module constructs a torso field using reference images to provide relevant texture and geometric priors,thereby achieving more precise torso rendering and natural movements. Experiments applied to multiple subjects demonstrate that,in self-reconstruction scenarios,the proposed method improves image quality(PSNR,LPIPS,FID,LMD) by 10.15%,12.12%,0.77%,and 1.09% respectively,and enhances lip-sync accuracy(AUE) by 14.20% compared to the current state-of-the-art baseline model. Concurrently,there is a notable increase of 14.20% in lip synchronization accuracy as measured by Sync metrics. Under cross-driving conditions with out-of-domain audio sources,the lip synchronization accuracy is achieved improvements of 4.74%.

Keywords　Talking portrait synthesis,3D reconstruction,Audio-video synchronization,Nerual radiance fields,Attention mechanism

到稿日期:2024-03-05　返修日期:2024-10-08
基金项目:国家自然科学基金(62301278,62371254,61931012);江苏省自然科学基金(BK20230362,BK20210594)
This work was supported by the National Natural Science Foundation of China(62301278,62371254,61931012)and Natural Science Foundation of Jiangsu Province,China(BK20230362,BK20210594).
通信作者:谢九成(jiuchengxie@njupt.edu.cn)

1 引言

音频驱动的说话人像合成技术致力于实现从任意输入音频到动态面部图像的跨模态映射，旨在合成具有真实细节和自然说话风格的面部图像。早期研究工作[1]主要采用生成对抗网络[2](Generative Adversarial Networks，GAN)来学习音频到面部的映射关系，并在合成过程中引入二维或三维的中间面部表征[3]以提高合成结果的质量和可控性。然而，由于对抗生成模型本身的局限性[4]，训练不稳定、模式崩溃或细节难以精确捕捉等问题依然存在。

神经辐射场[5](Neural Radiance Fields，NeRF)作为新兴的计算机图形学技术，由 Guo 等[6]首次应用于合成说话人像，提出了 AD-NeRF(Audio Driven Neural Radiance Fields for Talking Head Synthesis)。该方法通过多层感知器(Multi-Layer Perceptron，MLP)学习音频特征到空间点密度和颜色的映射，进而使用体渲染技术生成对应的说话人像。与基于 GAN 的渲染技术相比，基于 NeRF 的工作[7-8]能够生成更加清晰且自然的结果，但其在训练与推理过程中所需的漫长时间是制约其广泛部署与应用的主要障碍。近期，Instant-NGP(Instant Neural Graphics Primitives)[9]使用一种基于哈希表的编码方式来表示空间点坐标，这种方式能够有效捕捉 3D 空间中的体积特征，从而对 NeRF 训练进行加速。随后，RAD-NeRF(Real-time Neural Talking Portrait Synthesis)[10]将该加速技术应用于说话人像合成领域，并针对领域中的具体问题进行了专门优化，从而提升了合成人像的效率和视觉质量。

然而，这些端到端的模型在实际应用中仍面临两大挑战。首先，在头部合成方面，这些模型依赖于哈希表等网格编码器来存储三维场景中空间点的体积特征。由于三维场景中空间点的数量远远超过网格编码器的存储容量，因此将不可避免地使有效且稀疏的面部表面点与无效但密集的非表面点在编码器中发生哈希碰撞。这些碰撞均匀地散布在占用的网格中，最终导致合成结果过于平滑，缺少高频细节。其次，在躯干部分的合成中，模型往往依赖于面部跟踪技术获得的头部姿势来指导躯干的运动。该方法将复杂且迅速的头部运动与相对稳定的躯干运动一同建模，往往会导致两种运动模式相互干扰，最终出现模糊的躯干合成结果与不自然的躯干运动。

本文在 AD-NeRF[6]的基础上，针对动态头部表示和躯干合成中的挑战，进行了一系列的改进与优化。首先，针对头部合成中的问题，采用三平面哈希表示法代替传统的哈希编码器。该机制将三维空间点坐标投影至 3 个正交平面上的二维坐标，并对这些二维坐标进行哈希编码，从而将三维空间区域压缩至二维平面。相较于直接对三维坐标进行哈希编码，它能将哈希碰撞限制在低维子空间中并降低其发生的频率。其次，考虑到体渲染过程中面部表面点直接影响说话人的嘴型同步精度与面部细节，本文提出了一个区域显著性感知模块。该模块通过多模态输入信息预测显著性向量，并利用该向量对采样点的体积特征进行动态调整。显著性向量的应用不仅针对性地强化了表面点对应的体积特征，从而提升了渲染精度和效率，还通过反向传播机制确保了哈希表中优先存储表面点的体积特征。此外，对于躯干的合成，考虑到躯干相较于头部具有更稳定的运动模式，且其在不同头部姿势下具有相同的拓扑结构，本文算法采用了一种编码器-解码器架构[11]从单张躯干图像中提取躯干特征，从而更好地捕捉躯干的细节和动态变化。进一步地，算法通过更高效的三特征平面作为躯干场的表征，在不增加计算复杂度的情况下，允许每个特征平面以更高分辨率进行捕捉和存储。最后，通过单独的神经辐射场来建模躯干，有效地避免了复杂头部运动对躯干合成结果的干扰，实现了更平滑且自然的躯干运动。简而言之，本文的主要贡献如下：

1)提出了一个轻量级的区域显著性感知模块，通过多模态输入信息对空间点的体积特征进行动态调整，从而实现更高的渲染效率和面部建模精确度。

2)提出了一个空间特征提取模块，从单张躯干图像中提取以三平面为表征的躯干场，并通过单独的神经辐射场对躯干进行建模。在避免复杂头部运动对躯干合成产生干扰的同时，更好地捕捉躯干的细节和动态变化。

3)经过严格的实验，验证了本文算法在提升训练和推理效率的同时，显著提高了生成人像的质量和嘴唇同步精度，在客观评价指标和用户研究测试中均展现出先进的性能。

2 相关工作

2.1 音频驱动的说话人像合成

音频驱动的说话人像合成技术致力于利用任意输入音频序列生成特定人物的高质量说话视频。早期工作[1]采用传统方法，通过建立音频与嘴型之间的映射规则，在静态面部图像上合成相应动作的嘴唇。随着深度学习技术的发展，生成对抗网络[2]等深度学习模型被应用于音频驱动的说话人像合成，这些方法[12]旨在利用面部关键点作为中间表征，采用二阶段合成策略来增强合成结果的可控性。然而，二维关键点无法准确表征三维头部结构的缺点，限制了此类说话人像合成方法的合成质量和准确度。为了弥补这一缺陷，一些研究[13-14]转向使用三维可变形模型[15](3D Morphable Model，3DMM)进行说话人像合成，得益于三维模型在表征头部结构方面的优势，这些方法能够生成更准确且自然的结果。但如何解决 3DMM 模型固有的误差和信息损失以及其在训练过程中引入的额外误差，仍是一大挑战。

近期，NeRF 技术在音频驱动的说话人像合成领域展现出了巨大的应用潜力，特别是在处理三维头部结构的精确建模方面。在这一领域的早期尝试中，AD-NeRF[6]开创性地将动态 NeRF 技术引入说话人像合成领域，使用两个独立的神经辐射场分别建模头部和躯干。SSP-NeRF[16]提出了一个语义感知的动态射线采样模块以提升采样效率，并引入了一个躯干变形模块以稳定大范围的非刚性躯干运动。DFRF[17]关注到了基于 NeRF 的说话人像合成技术在泛化性上的不足，并引入面部扭曲模块来实现对未见过主体的快速泛化。DFA-NeRF[18]则采取了一种解耦唇部运动特征和个性化属性的方法，以实现高保真度和个性化的说话人像生成。最近，RAD-NeRF[10]通过引入哈希编码机制，大幅度提升了合成结果的质量和渲染速度。然而，上述方法大多依赖于一个复杂

的多层感知机隐式地学习音频到面部运动的跨模态映射，这限制了其收敛速度和重建质量。本文提出了一种更高效的网格 NeRF 方法，利用注意力机制在音频特征和头部空间区域之间建立明确的联系，大大提高了视觉质量和嘴唇同步精度。

2.2 高效的神经辐射场

NeRF 技术利用全连接网络来存储三维场景的几何与外观信息，在新视角合成上取得了前所未有的逼真效果，也受到了人像合成领域的广泛关注。面对 NeRF 技术漫长的训练时间和缓慢的推理速度，最近的研究工作开始探索提升 NeRF 效率的方法，特别是结合显式和隐式表达方式来优化静态场景重建。Instant-NGP[9] 使用多哈希表来存储三维场景中的体积特征，显著提升了渲染质量和内存利用率。然而，Instant-NGP 未能解决哈希表长度限制引起的碰撞问题，而是通过隐式的 MLP 进行缓解。EG3D[19] 将一个高维场景分解为 3 个相互正交的低维特征平面，以此来获得更为紧凑而高效的场景表示。这种方法有效降低了模型的复杂度，同时充分保留了场景的详细信息。本文提出了一种基于三特征平面的显式表示方法，用于表征标准躯干场。该方法将复杂的三维几何信息映射到二维特征平面，每个特征平面负责捕获躯干在特定方向上的特征，从而显著降低了计算的复杂度。这种特征提取和融合策略减少了信息的冗余且避免了歧义，同时确保了细节和纹理信息得到充分保留，进而实现了真实且自然的躯干运动。

3 本文方法

3.1 预备知识

NeRF[5] 引入了一个神经辐射场(隐式函数 F)，用于表征静态三维场景。该函数被定义为 $F:(X,d)\rightarrow(\sigma,c)$，其中 $\boldsymbol{X}=(x,y,z)$ 代表三维空间中的点坐标，$d=(\theta,\phi)$ 表示观察方向，函数输出 $c=(r,g,b)$ 和 σ 分别对应该空间点的颜色和体积密度。基于神经辐射场预测的密度和颜色，NeRF 使用体渲染技术从任意指定的相机位置和方向合成新视角图像。具体而言，算法从相机位置 o 向成像平面上的每一像素点发出光线，并对每条光线沿射线方向进行采样，采样点形式为 $x_i=o+t_i d$。对于射线上的每一个采样点，隐式函数 F 被用来计算其对应的密度和颜色，随后采用积分的方式计算该射线对应像素的颜色。

$$\hat{\boldsymbol{C}}(\boldsymbol{r})=\sum_i T_i\alpha_i c_i;\ T_i=\prod_{j<i}(1-\alpha_j) \tag{1}$$

其中，T_i 是投射率，$\alpha_i=1-\exp(-\sigma_i\delta_i)$ 为不透明度，$\delta_i=t_{i+1}-t_i$ 是射线上相邻采样点的间隔步长。借助这种完全可微的渲染方式，NeRF 能够将三维场景中的离散空间点渲染成特定视角下的图像。

3.2 方法概览

在预处理阶段，模型从输入的训练集视频中提取了 3 种不同模态的信号：使用自动语音识别模型[10](Automatic Speech Recognition，ASR)从输入音频序列中提取的音频特征 $\boldsymbol{A}$、应用 3DMM 模型[15]估算的头部姿势 $\boldsymbol{P}$、通过 Dlib 工具从面部结构中提取的眼部运动特征 $\boldsymbol{E}$[10](一个由眼部和面部面积比例乘以 100 而得到的值，用于精确控制眼睛的开闭程度)。鉴于说话人在连续视频帧中的运动特性，我们将完整的说话人像分解为头部、躯干和静态背景。由于头部和躯干具有不同的运动模式，应用相同的刚性变换会导致躯干出现大幅抖动，因此我们采用两个独立的神经辐射场对这两部分进行建模。本文的说话人像合成模型的框架如图 1 所示。

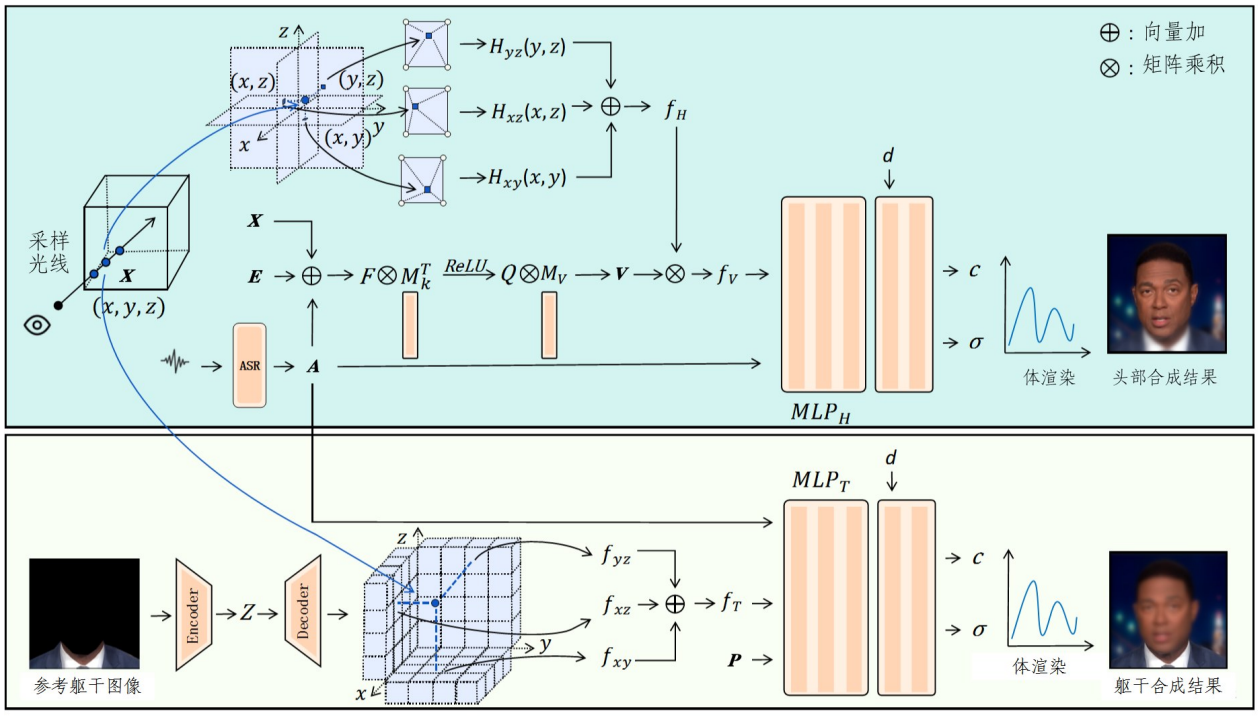

图 1 本文模型的框架

Fig. 1 Framework of the proposed method

对于头部的建模，首先由头部姿势 $\boldsymbol{P}$ 解算对应的相机位姿，并在此基础上计算采样点的空间坐标 $\boldsymbol{X}$。然后通过一个三平面哈希编码器将点坐标编码为对应的体积特征 $\boldsymbol{f}_H$。区域显著性感知模块通过音频特征 $\boldsymbol{A}$ 和眼部运动特征 $\boldsymbol{E}$ 预测显著性向量 $\boldsymbol{V}$，从而对体积特征 $\boldsymbol{f}_H$ 进行通道级加权。结合加权后的体积特征 $\boldsymbol{f}_V$、音频特征 $\boldsymbol{A}$ 和视角信息 d，解码器（MLP_H）预测头部采样点的颜色 c 和密度 σ，并通过体渲染的方式合成说话人的头部。

对于躯干的建模，通过空间特征提取模块的编解码器，从输入的参考躯干图像中提取规范化的躯干场，并将采样点 X 投影到躯干场的 3 个正交平面上。利用双线性插值法获取空间点在这个维度上的体积特征（$\boldsymbol{f}_{xy}$，$\boldsymbol{f}_{xz}$，$\boldsymbol{f}_{yz}$），并计算这些特征的均值 $\boldsymbol{f}_T$。随后将综合体积特征 $\boldsymbol{f}_T$ 与音频 $\boldsymbol{A}$ 和头部姿势 $\boldsymbol{P}$ 串联后输入解码器（MLP_T）来预测躯干采样点的颜色 c 和密度 σ，再利用体渲染合成说话人的躯干。

最后，通过生成的二值掩码将头部和躯干图像覆盖在静态背景上，从而生成真实且自然的说话人像。

3.3 区域显著性感知模块

区域显著性感知模块包含动态显著性预测网络及通道级加权两个主要部分。为了适应多模态输入的变化并动态调整有效面部区域，本文算法采用了一个可训练的显著性预测网络（见图 2）。该网络能够根据不同的输入信号，生成相应的显著性向量。通道级加权部分通过显著性向量对空间区域的体积特征进行通道级加权，从而强化表面点的体积特征，同时相应地弱化了无效区域的体积特征。

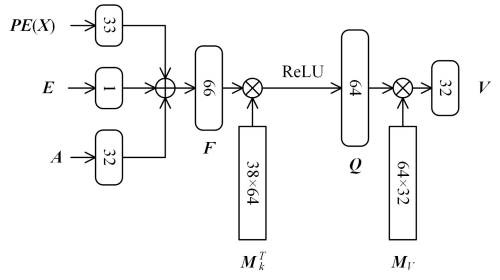

图 2 显著性预测网络结构图

Fig. 2 Saliency prediction network architecture

本文算法通过外部注意力机制实现显著性预测网络。Beyond Self-Attention[20] 通过实验证明两个线性层是有效且轻量级的注意力机制实现方法，因此本文使用两个线性层和一个归一化层来构建动态显著性预测网络，其中两个线性层被视为外部记忆单元，分别记作 $\boldsymbol{M}_k$ 和 $\boldsymbol{M}_v$。在训练过程中，这些记忆单元隐式地学习多模态输入信息与空间点体积特征之间的关系。

$$\begin{aligned}\boldsymbol{Q}&=\mathrm{ReLU}(\boldsymbol{F}\boldsymbol{M}_k^{\mathrm{T}})\\ \boldsymbol{V}&=\boldsymbol{Q}\boldsymbol{M}_v\end{aligned}\tag{2}$$

其中，串联了多模态输入信号的中间向量 $\boldsymbol{F}\in\mathbb{R}^{66}$ 可以理解为音频特征 $\boldsymbol{A}\in\mathbb{R}^{32}$、空间点坐标 $\boldsymbol{X}\in\mathbb{R}^{33}$ 和眼动特征 $\boldsymbol{E}\in\mathbb{R}^{1}$ 的联合表示。显著性预测网络的设计源于对体渲染机制的深入分析，即在动态三维场景对说话人进行建模时，面部区域的有效性很大程度上依赖于这些多模态特征。

随后，使用提取到的显著性向量 $\boldsymbol{V}$ 对体积特征 $\boldsymbol{f}_H$ 进行通道级加权，其中 $\boldsymbol{f}_H$ 是通过三平面哈希编码对采样点 $\boldsymbol{X}$ 进行编码所得。最终输出的特征向量为：

$$\boldsymbol{f}_V=\boldsymbol{V}\odot\boldsymbol{f}_H\tag{3}$$

其中，符号 $\odot$ 代表哈达玛矩阵乘积。通过此操作得到的输出特征向量 $\boldsymbol{f}_V$ 是体积特征 $\boldsymbol{f}_H$ 经显著性向量 $\boldsymbol{V}$ 加权调整后的结果。因此，根据多模态输入信息得到的显著性向量能够对三维空间区域的体积特征进行有效的放大或衰减，从而提升渲染精度。

将音频特征 $\boldsymbol{A}$ 与加权后的体积特征 $\boldsymbol{f}_V$ 串联后，采用一个多层感知机（MLP_H）作为解码器，输出采样点的颜色 c 和密度 σ。基于计算得到的颜色和密度，算法利用体渲染技术来合成与输入音频相对应的说话人头部图像。

$$c,\sigma=MLP_H(\boldsymbol{f}_V,\boldsymbol{A})\tag{4}$$

在训练阶段，对于空间中的有效点，注意力机制使得其对应的显著性向量逐渐增强，以便充分利用体积特征 $\boldsymbol{f}_H$。相反，对于被视为无效的空间点，体积特征将被判定为不具参考价值，相应的显著性向量 $\boldsymbol{V}$ 将趋向于零向量。这样的处理减少了无效特征的干扰，并使得这些点的输出密度和颜色接近零，从而提高了模型的渲染效率和合成质量。

3.4 空间特征提取模块

空间特征提取模块采用改进的 U-Net 卷积神经网络作为编码器-解码器，用于空间特征提取。该结构使算法能够从参考躯干图像中提取出包含体积特征的躯干场，这一躯干场不仅存储了躯干的几何和外观信息，还保留了表面的纹理和细节。得益于躯干拓扑结构在不同姿势下的一致性，所提算法能够在新的姿势中使用存储在躯干场中的特征，重建出与参考图像在视觉上一致的新躯干图像。

图 3 展示了用于提取躯干场的编码器-解码器结构。具体来说，首先使用编码器从参考图像中提取躯干的潜在编码 $\boldsymbol{Z}$，该编码包含了躯干的几何和外观信息。然后，通过一个全连接层将潜在编码 $\boldsymbol{Z}$ 转化为 4096 维的特征向量，随后该向量被重塑为一个 256 通道的特征图。接着，使用包含多个反卷积层的解码器网络对特征图进行上采样，生成一个 96 通道、512×512 分辨率的特征图。此特征图被进一步分解为 3 个 32 通道的特征图，这些特征图被重构为 3 个相互正交的平面（xy，xz，yz），定义为躯干表面场。对于采样点 $\boldsymbol{X}=(x,y,z)$，将其投影至 3 个特征平面上，并采用双线性插值的方法检索相应的特征向量 $\boldsymbol{f}_{xy}$，$\boldsymbol{f}_{xz}$，$\boldsymbol{f}_{yz}$。因此，采样点 $\boldsymbol{X}$ 处的空间特征为：

$$\boldsymbol{f}_T=\frac{\boldsymbol{f}_{xy}+\boldsymbol{f}_{xz}+\boldsymbol{f}_{yz}}{3}\tag{5}$$

其中，$\boldsymbol{f}_T\in\mathbb{R}^{32}$ 为三特征平面所提取特征的均值，代表从三维躯干表面场中提取的空间特征，综合反映了对应空间点的几何和纹理等多维度的语义信息。鉴于音频特征和头部姿势对躯干运动的影响，本文算法将空间特征 $\boldsymbol{f}_T$、音频特征 $\boldsymbol{A}$ 和头部姿势 $\boldsymbol{P}$ 进行串联，进而输入至一个多层感知机（MLP_T）中，以预测空间点的密度和颜色信息。最终，通过体渲染技术计算每个像素的 RGB 颜色值并合成相应的躯干图像。

$$c,\sigma=MLP_T(\boldsymbol{f}_T,\boldsymbol{A},\boldsymbol{P})\tag{6}$$

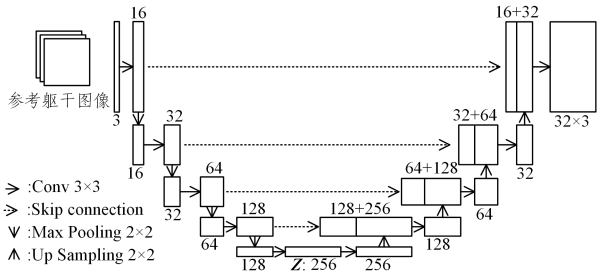

图 3 编码器-解码器网络结构图

Fig. 3 Network architecture of the encoder-decoder

3.5 训练细节

3.5.1 三平面哈希编码

三平面哈希编码将三维空间点坐标 $\boldsymbol{X}=(x,y,z)$ 投影至 3 个二维平面坐标 (x,y),(x,z),(y,z)。这些二维坐标随后被二维哈希编码为对应的体积特征。算法对这些体积特征进行逐元素相加,得到最终表示三维空间点体积特征的 $\boldsymbol{f}_H$:

$$\boldsymbol{f}_H=H_{xy}(x,y)\oplus H_{yz}(y,z)\oplus H_{xz}(x,z) \tag{7}$$

其中,$\oplus$ 表示特征加,该操作将不同平面的哈希特征整合为一个维度为 $L\times K$ 的向量;H_{xy},H_{yz},H_{xz} 分别表示在 xy,xz 和 yz 平面上执行的多级分辨率哈希编码过程,每一编码过程均包含 L 层,每层的特征维度为 K。先前的研究工作[21]通过实验证明,相较于传统的三维哈希编码,三平面哈希编码机制将哈希碰撞限制在二维平面,从而有效地减少了哈希碰撞的发生次数并提高了体积特征的表达能力和内存利用率。

3.5.2 网络结构

图 2 为显著性预测网络的结构图。网络接受音频特征 $\boldsymbol{A}$、眼动特征 $\boldsymbol{E}$ 和编码后的空间点坐标 $\boldsymbol{PE}(x)$ 作为输入,预测 32 维度的显著性向量 $\boldsymbol{V}$。图 2 中圆角矩形中的数字表示特征向量的维度,而方角矩形中的数字表示矩阵的维度,位置编码(PE)的频率设置为 10。位置编码技术将三维空间坐标转换成正弦和余弦函数的高维序列,有效地增强了模型对高频信息的捕捉能力。

图 3 展示了用于提取躯干场的编码器-解码器架构,这是一种基于 U-Net 的网络设计。结构中的方块代表特征图,相邻的数字则表示特征图的通道数。网络接受参考躯干图像作为输入,并通过一系列的 3×3 卷积层逐步提取图像特征。在每个卷积层后,网络通过 ReLU 激活函数和 2×2 的最大池化操作来降低特征的空间维度并增强模型的抽象处理能力。特征通道从 16 逐步增加到 256,使得网络能够逐步捕捉从简单到复杂的图像细节。在编码器到达底层后,网络通过上采样以及跳跃连接技术逐步重建图像的详细特征,从而保留编码阶段捕获的重要特征。最终,网络输出一个具有 96 通道、512×512 分辨率的特征图,该特征图被进一步重构为 3 个相互正交的特征平面,定义为躯干场。这种设计使得网络能够准确捕获躯干的几何和纹理特征,从而提高躯干重建的质量和准确性。

图 4 展示了两个用于预测密度和颜色的神经辐射场。方块中的数字表示线性层宽度,MLP_H 和 MLP_T 分别表示用于建模头部和躯干的神经辐射场。

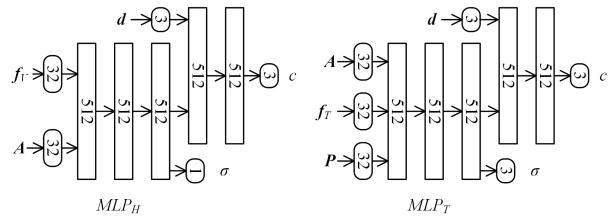

图 4 神经辐射场网络结构图

Fig. 4 Network architecture of the neural radiance field

3.5.3 失函数

本文采取了分阶段的训练策略,包含粗调阶段和微调阶段,以进一步提高图像质量。在这两个训练阶段中采用了不同的损失函数来优化网络参数。在粗调阶段,利用最小化均方误差损失(Mean Square Error,MSE)来训练网络,使其能够准确预测图像 I 中每个像素的颜色 $\hat{C}(i)$。

$$\mathcal{L}_{\text{coarse}}=\sum_{i\in I}\| C(i)-\hat{C}(i)\|_2^2 \tag{8}$$

通过均方误差损失,模型能够迅速收敛至最小误差值。然而 MSE 损失的应用可能会导致高频细节的损失,这限制了算法对图像细节的优化,使得最终合成的图像呈现出相对模糊的视觉效果。为了克服这一限制,在微调训练阶段,本文引入了感知相似度损失 LPIPS 来进行进一步的优化。算法从整幅图像中随机选取一组像素点 p,并利用权重 λ 对 LPIPS 损失进行调整,以便更加精细地增强图像细节。

$$\mathcal{L}_{\text{fine}}=\sum_{i\in I}\| C(i)-\hat{C}(i)\|_2^2+\lambda LPIPS(\hat{p},p) \tag{9}$$

4 实验

4.1 实验设置

4.1.1 数据集

与先前的研究相比,本文算法不需要大规模数据集或长达数十小时的视频,仅需几分钟的单一主体演讲视频就能生成高保真度的说话人像视频。为了深入评价本文算法在实际应用场景中的表现,本文从已有工作[6,16]中筛选并整理了 4 个数据集,它们均源自公开可获取的演讲视频,每个视频长度介于 3~5 min 之间,帧率为 25 fps。除了从 AD-NeRF 获得的奥巴马演讲视频的分辨率为 450×450 外,其他 3 个视频的分辨率均被处理为 512×512。本研究遵循基于 NeRF 的说话人像合成方法的标准预处理流程,对输入视频进行以下 4 个阶段的处理:1)利用 Dlib 和 OpenCV 工具提取面部关键点,以识别人脸上的重要区域,如眼睛、鼻子和眉毛;2)采用语义分析算法对上半身图像进行细分,将其分割为头部、躯干和背景 3 部分;3)应用三维面部模型(3DMM)[15]进行面部跟踪,以提取头部姿势参数,并据此计算相机的位姿;4)使用 ASR 自动语音识别模型,从音轨中提取音频特征。

4.1.2 对比方法与评价指标

本研究将所提算法与近期一些标志性的单图像驱动的说话人像生成模型(PC-AVS[22],DINet[23])进行了比较。此外,还将其与 3 种基于 NeRF 的方法(AD-NeRF[6],RAD-NeRF[10]和 ER-NeRF[21])进行了对比分析。考虑到基于 NeRF 的方法能够合成躯干部分,本文在定性评价中对躯干

的合成结果也进行了展示。

在定量评价中，为了衡量重建结果的图像质量，本文首先使用峰值信噪比[10]（Peak Signal to Noise Ratio，PSNR）来量化失真程度。

$$PSNR=20\cdot\log_{10}\left(\frac{255}{\sqrt{MSE}}\right) \tag{10}$$

其中，255 为像素点颜色的最大数值，MSE 为合成结果 $\boldsymbol{I}$ 和真值 $\boldsymbol{K}$ 间的三通道均方误差。MSE 的定义如下：

$$MSE=\frac{1}{hw}\sum_{i=0}^{h-1}\sum_{j=0}^{w-1}[\boldsymbol{I}(i,j)-\boldsymbol{K}(i,j)]^2 \tag{11}$$

其中，h 和 w 分别表示图像的高度和宽度。

其次，采用感知相似度损失[24]（Learned Perceptual Image Patch Similarity，LPIPS），通过预训练的 VGG 模型从人类感知的角度评价生成图像的质量。

$$D^l(\boldsymbol{I},\boldsymbol{K})=\frac{1}{h_l w_l}\sum_{(h,w)}\parallel\hat{\boldsymbol{I}}_{hw}^l-\hat{\boldsymbol{K}}_{hw}^l\parallel_2^2 \tag{12}$$

其中，感知损失 D^l 表示合成结果 $\boldsymbol{I}$ 和真值 $\boldsymbol{K}$ 在第 l 层特征空间的差异。通过对 L 层的感知损失进行加权求和，可以得到最终的 LPIPS 损失。

$$LPIPS(\boldsymbol{I},\boldsymbol{K})=\sum_l w_l\cdot D^l(\boldsymbol{I},\boldsymbol{K}) \tag{13}$$

此外，本文使用了生成模型的评估指标[25]（Fréchet Inception Distance，FID），通过预训练的 InceptionV3 模型提取图像特征，并通过比较生成图像和真实图像的特征均值和协方差矩阵的分布差异来衡量其相似性。

$$FID=\parallel\mu_i-\mu_k\parallel_2^2+Tr\left(\sum_i+\sum_k-2\left(\sum_i\sum_k\right)^{\frac{1}{2}}\right) \tag{14}$$

其中，μ_i 和 μ_k 是生成图像和真实图像的特征均值，Σ_i 和 Σ_k 是特征协方差矩阵。

再次，通过计算真实图像和生成图像的面部关键点的欧氏距离，能够判断合成结果的表情是否准确。面部关键点距离[26]（Landmark Distance，LMD）的计算式如下：

$$LMD=\frac{1}{N}\sum_{i=1}^{N}\sqrt{(x_i-x_i')^2+(y_i-y_i')^2} \tag{15}$$

其中，(x_i,y_i) 和 (x_i',y_i') 分别是真实图像和生成图像的第 i 个面部关键点坐标。

然后，本文通过 SyncNet[27-28] 置信度（Sync）评估合成的说话人唇部和驱动音频的同步程度。预训练的 SyncNet 模型从每一帧输入视频及其对应音频片段中提取特征向量 $\boldsymbol{V}_t$ 和 $\boldsymbol{A}_t$，然后计算它们之间的余弦相似度，并在整个视频长度 T 内对所有视频帧的相似度得分取平均，得到最终的置信度分数。

$$Sync=\frac{1}{T}\sum_{t=1}^{T}sim(\boldsymbol{V}_t,\boldsymbol{A}_t) \tag{16}$$

最后，本文使用了动作单元误差[29]（Action Unit Error，AUE）来评估生成结果的面部结构是否准确。算法通过预训练的 OpenFace 模型分别检测合成结果和真值的面部动作单元（用于表征面部特定肌肉运动，如眉毛上扬或嘴角拉动等），并计算两者的差值作为动作单元误差。

$$AUE=|AU_I-AU_K| \tag{17}$$

4.1.3 比较设置

本研究通过自驱动头部重建和交叉驱动头部合成两种实验设置评估了所提算法的有效性。在自驱动头部重建实验中，参照先前的工作[6,10,21]，将 4 个数据集分割成训练集和测试集。随后，采用测试集中音频驱动算法合成说话人视频并使用测试集视频来评估合成视频的质量。在此实验设置中，由于拥有合成结果的真值，研究除了使用 Sync 置信度和 AUE 外，还引入了 PSNR，LPIPS，FID 和 LMD 指标来全面评估合成结果的视觉质量。在交叉驱动头部合成实验中，本文在奥巴马数据集上训练了所提算法及其他算法，并从 NVP[14] 和 SynObama[30] 中选取了两段音频片段，分别构成测试集 A 和测试集 B。由于该设置下缺乏合成结果的真值，因此只采用 Sync 置信度和 AUE 来评价唇读准确性和面部动作一致性。

4.1.4 实施细节

本文的代码基于 Pytorch 框架实现，并在单张 NVIDIA RTX3090 GPU 上执行所有实验。对于每个数据集，头部训练分为两个阶段，即 200000 步的粗调阶段和50000步的微调阶段。在粗调阶段，采用均方误差（MSE）损失以加速模型收敛；在微调阶段，则引入了感知相似度损失（LPIPS），专注于改善嘴唇区域的细节。在每次迭代中，模型采样 256 条射线，并在每条射线上生成 16 个采样点进行训练。躯干部分的训练则进行了 100000 次迭代，采样策略与头部相同。模型训练采用 Adam 优化器，设置哈希编码器的学习率为 0.01，其他模块为 0.001。头部和躯干部分的训练时间分别为 2.5 h 和 5 h。此训练策略旨在平衡训练效率与生成说话人像的视觉质量，确保模型能够在保持高效合成的同时，优化结果的细节和真实感。

4.2 定量评价

在上述两种设置下的实验结果分别如表 1 和表 2 所列，其中最佳结果以粗体表示。表 1 中的实验结果通过计算测试集中的 350 张图像的定量评估指标并取其平均值得到。此外，本文还评估了各算法在两个非训练集音频驱动下的表现，其中每个音频的长度为 30 s，相应生成了 750 帧说话人像。表 2 列出了这些图像的平均动作单元误差（AUE）和嘴唇同步精度（Sync）指标，用于评估面部动作与音频同步的精确度。

表 1 列出了自驱动头部重建设置下各种方法的实验结果。得益于区域显著性感知模块，本文算法可以有效地减少无效空间区域对合成结果的干扰，从而生成更高质量的说话人像。结果显示，本文算法在多个关键性能指标上显著优于现有技术。具体来说，本文算法在峰值信噪比（PSNR）、感知相似度损失（LPIPS）和 FID 方面均获得了最高分。这不仅证明了该算法在保持面部细节和纹理方面的优势，还反映了其在保持视觉一致性方面的能力。此外，本文算法在 LMD 和动作单元误差上也展现了卓越的表现，进一步证明了其在复现精确面部结构方面的能力。值得注意的是，DINet 和 PC-AVS 在唇读准确性（Sync）上取得了高于本文算法的得分，这主要归因于它们在训练过程中直接将 Sync 指标纳入损失函数。这种极端的优化策略使得其 Sync 指标甚至高于训练用的真实视频，显然这是不合理的。作为代价，这种过度优化策略使得其在其他关键性能指标上的表现不尽人意。为了使每一帧都尽可能匹配音频，这些方法往往会牺牲自然的过度动作，从而产生视觉上的抖动与模糊，这就降低了整体视

频的真实感。这一点在随后的定性评价中也得到了体现。在综合评估这种策略后，我们决定不采用此方法，而是追求更全面的质量提升，不是单一指标的极端优化，这也更符合实际应用场景中用户对合成视频的质量要求。最后，本文还比较了各方法的推理速度和训练时间，其中 PC-AVS 和 DINet 使用了预训练的图像生成模型，因此没有固定的训练时间（表 1 中使用斜线表示）。结果表明，本文算法不仅实现了实时推理，而且在训练时间和推理速度上也仅次于当前最先进的方法 ER-NeRF。这体现了本文算法在效率和实用性方面的显著优势。

表 1 自驱动下头部重建结果的定量比较

Table1 Quantitative comparison for self-driven head reconstruction results

Method	*PSNR* ↑	*LPIPS* ↓	*FID* ↓	*LMD* ↓	*Sync* ↑	*AUE* ↓	*Time* ↓ /h	*FPS* ↑
GT	—	—	—	—	7.63	—	—	—
PC-AVS	18.25	0.24	101.97	4.81	**8.39**	3.14	—	32
DINet	30.42	0.03	25.42	3.68	7.76	3.35	—	32
AD-NeRF	30.75	0.09	18.60	3.34	5.25	5.20	18	0.13
RAD-NeRF	33.26	0.03	12.20	2.81	5.05	2.10	5	32
ER-NeRF	33.10	0.03	10.42	2.74	5.70	1.62	**2**	**34**
Ours	**36.63**	**0.02**	**10.34**	**2.71**	6.60	**1.39**	2.5	33

表 2 交叉驱动下头部合成结果的定量比较

Table 2 Quantitative comparison for cross-driven head synthesis results

Method	测试集 A		测试集 B	
	Sync ↑	*AUE* ↓	*Sync* ↑	*AUE* ↓
GT	6.55	—	7.04	—
PC-AVS	**6.78**	1.15	7.15	1.51
DINet	6.65	1.09	**7.24**	1.75
AD-NeRF	4.04	1.13	5.36	1.53
RAD-NeRF	6.02	1.15	6.76	1.30
ER-NeRF	6.24	1.13	6.93	1.38
Ours	6.37	**1.07**	6.95	**1.20**

在交叉驱动的实验设置下，本研究进一步对比了不同的说话人像合成算法的性能，特别是在同步性和泛化能力方面。表 2 详细记录了各算法在此设置下的评价结果。与自驱动的实验结果类似，PC-AVS 和 DINet 算法在 Sync 置信度指标上显著高于用于训练的真实视频，这种优化策略以牺牲其他关键性能指标为代价，如图像质量和嘴唇同步精度。进一步的分析显示，与当前代表性的方法相比，本文算法在动作单元误差（AUE）分数上表现最佳，同时在基于神经辐射场（NeRF）的方法中实现了最高的唇读准确性（Sync）。这些结果验证了本文算法在特定任务上的优越性，并突显了其在不同数据和交叉场景下的出色泛化能力。

综上所述，本文算法在自驱动和交叉驱动实验设置下均展现出了卓越的性能。该算法在关键的图像质量评价指标（PSNR，LPIPS，FID，LMD）上取得了最高分数，不仅证明了其在图像质量和面部结构准确性上的优势，也反映了其精确的细节捕捉能力。此外，本文算法在嘴唇同步精度上也表现出了显著的优势，这一点在 Sync 和 AUE 指标的优异表现上得到了体现。

4.3 定性评价

图 5 展示了在自驱动头部重建下，各算法在奥巴马数据集上的测试成果。图 6 则呈现了基于 NeRF 的方法在 4 个不同主体上的合成效果，包括对特定区域的放大，以便于更精细地对比细节表现。图 7 展示了交叉驱动测试视频与代表性方法合成结果的比较分析。

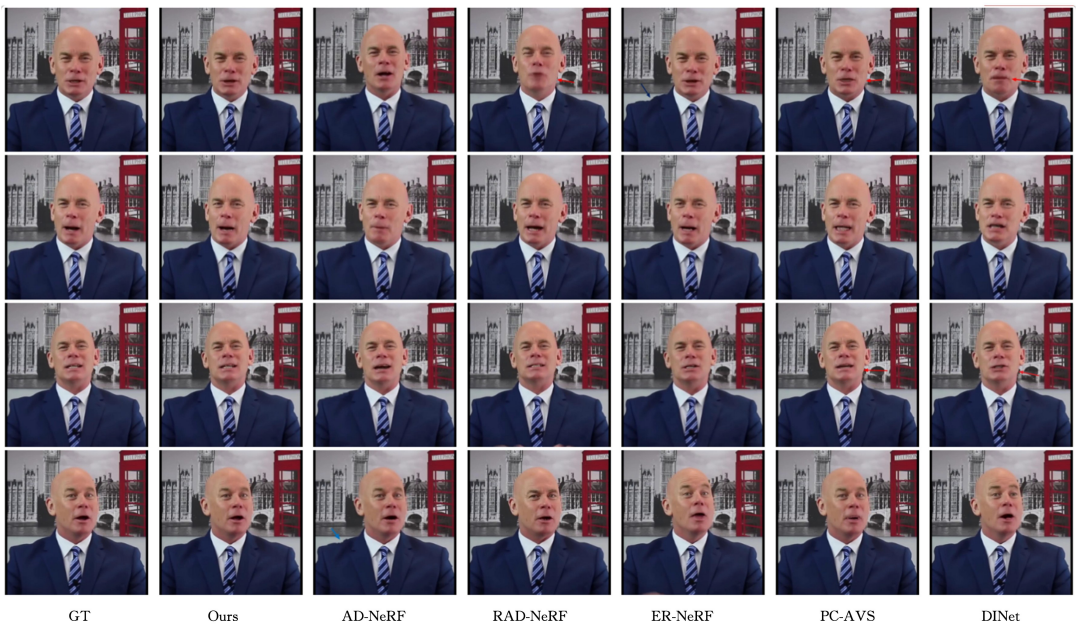

图 5 自驱动下代表性方法头部重建实验结果

Fig. 5 Generated results for self-driven head reconstruction of representative methods

结果表明，尽管 PC-AVS 和 DINet 在唇读准确性（Sync）上得分较高，但它们生成的结果在视觉上与真实图像之间存在显著差异。具体来说，PC-AVS 在嘴部区域的合成结果上出现明显模糊，而 DINet 未能准确保留说话人的身份特征，影响了合成结果的真实感。AD-NeRF 在合成结果中出现了较为严重的抖动现象，这点从其头部与躯干的分离的结果中可以看出。图 6 展示了基于 NeRF 的方法在 4 个主体上的头部与躯干的合成结果，并对特定细节放大，以评估说话人像的视觉质量。RAD-NeRF 的合成结果在唇部出现了明显异常，而 ER-NeRF 在躯干部分的细节上表现得不尽人意，且其嘴唇结构与真实情况存在一定偏差。结果显示，本文算法在头部合成质量上实现了显著的提升，同时在躯干稳定性和细节复现方面也超越了其他基于 NeRF 的方法，展现出了更加出色的性能。图 7 展示了交叉驱动用的测试视频和各方法在不同单词下的合成结果。相较于其他的代表性方法，本文算法能够更加稳定且准确地合成音频同步的嘴唇动作，尤其是在嘴唇有较大幅度开合时（例如发音“Be”）。

为了深入评估本文算法在实际应用场景中的表现，特别是在视觉质量和真实性方面，我们组织了一项用户研究。这项研究邀请了 20 位不同背景的参与者，以确保评价的客观与多样性。他们被要求对本文算法和其他代表性算法合成的 30 段视频进行评价。本文在交叉驱动设置下合成了以上的视频片段，以确保评价结果符合现实应用场景的需求和挑战。评价采用平均意见得分（Meaning Opinion Score，MOS）协议，要求参与者基于 3 个维度对合成视频进行评价：音频-唇部同步准确性、视频真实性以及图像质量。每项评价的得分范围设定为 1～5 分，分数越高代表效果越好。表 3 汇总了各算法在这些评价指标上的平均得分。结果显示，本文算法在这 3 个评价维度上均取得了最佳表现，进一步验证了所提算法在生成高保真说话人像方面的卓越表现。

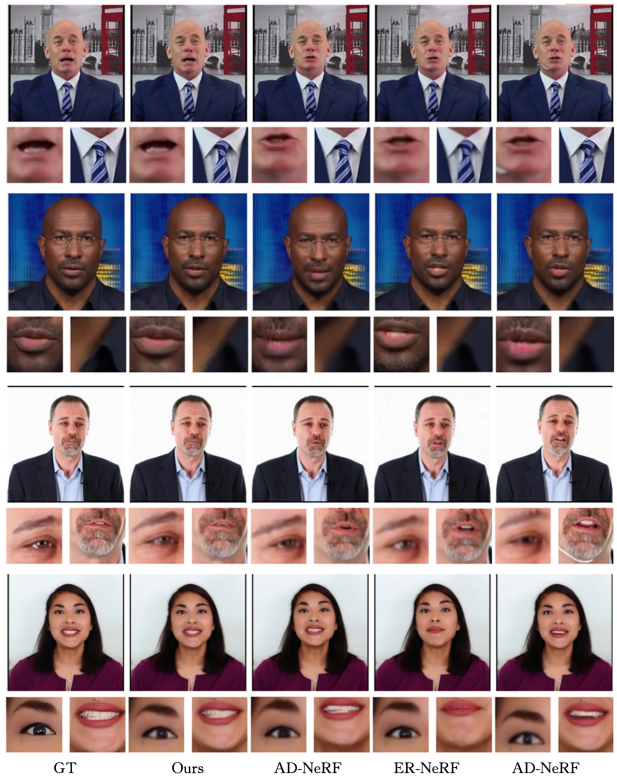

图 6 自驱动下基于 NeRF 的方法的头部重建实验结果

Fig. 6 Generated results for self-driven head reconstruction of NeRF-based methods

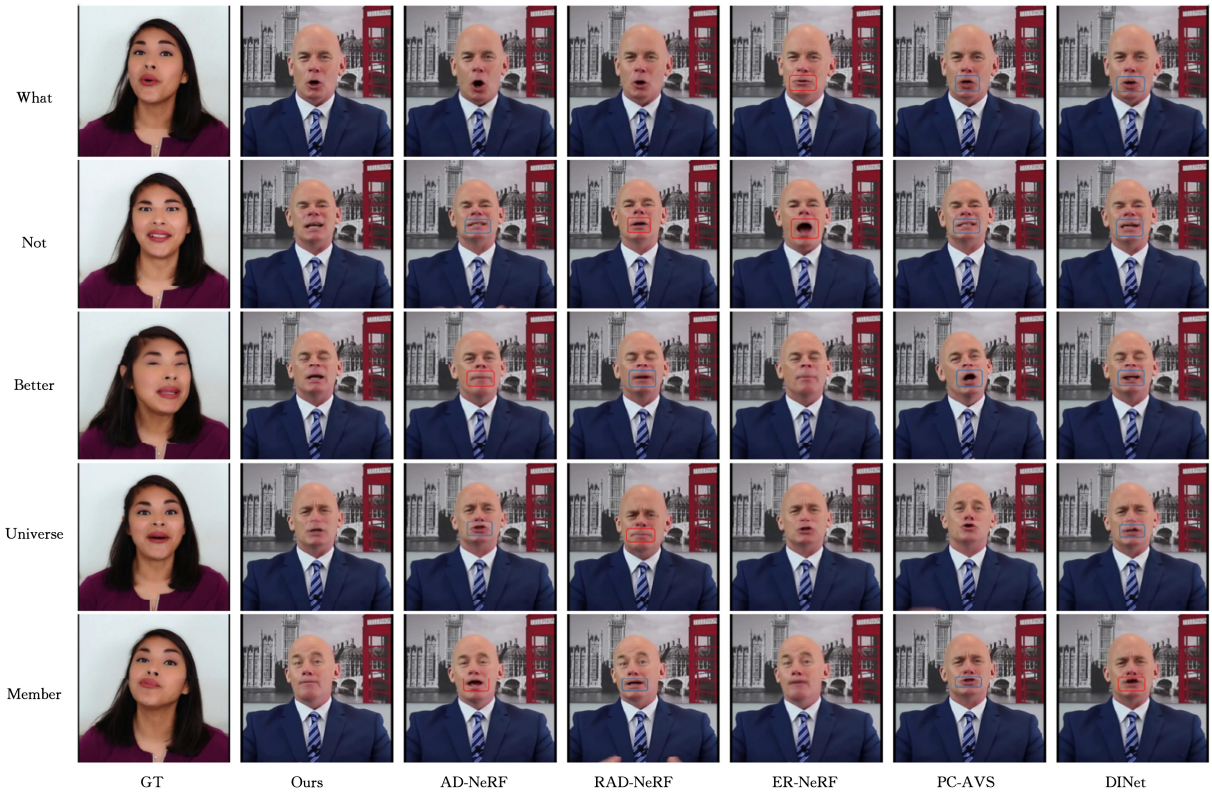

图 7 交叉驱动下合成视频的质量比较

Fig. 7 Cross-driven quality comparison

表 3 用户研究

Table 3 User study

方法	音唇同步准确性↑	视频真实度↑	图像质量↑
PC-AVS	3.47	2.79	2.75
DINet	2.56	2.65	2.67
AD-NeRF	2.63	2.60	2.03
RAD-NeRF	3.02	2.86	3.22
ER-NeRF	3.61	3.35	3.23
Ours	**3.63**	**4.63**	**3.74**

4.4 消融实验

本节通过消融实验评估了两项关键模块的有效性，实验在自驱动头部重建设置进行，使用从奥巴马数据集中分割出的训练集和测试集进行训练和评估。为了全面评估各模块的贡献，分别合成了说话人像的头部和躯干，并对合成的整体人像进行了评价。实验结果如表 4 所列。表 4 列出了本文提出的完整算法、移除区域显著性感知模块、移除空间特征提取模块，以及同时移除两个关键模块后合成的人像质量。可以清楚地看到，在移除两个关键模块之后，合成结果的视觉质量(PSNR，LPIPS) 和唇读准确性(Sync)显著下降，这证明了两个模块在提高人像合成质量方面的有效性。进一步地可以看出，区域显著性感知模块在建立音频与面部体积特征之间的约束关系方面起到的显著作用，该模块的移除对音频-唇部同步精度(Sync) 造成了较大影响。此外，尽管空间特征提取模块在提升躯干合成结果的视觉效果方面有显著效果，但其对唇读准确性的影响有限。

表 4 消融实验

Table 4 Ablation experiment

	PSNR↑	*LPIPS*↓	*Sync*↑
本文算法	**28.71**	**0.03**	**6.66**
无区域显著性感知	26.42	0.04	5.53
无空间特征提取	25.96	0.05	6.49
基准模型	24.44	0.06	5.42

结束语 本文介绍了一种新型说话人像合成方法，该方法的核心创新点在于两个关键模块：1)区域显著性感知模块，它通过预测的显著性向量对空间点的体积特征进行动态调整，从而精细化地构建头部模型，以提高面部合成的精度和效率；2)空间特征提取模块，它通过高效的三特征平面表征躯干场，并通过单独的神经辐射场建模躯干，有效减少了头部运动对躯干合成的干扰，实现了自然且稳定的躯干运动。相较于现有技术，本文算法在图像质量与合成效率上均有显著提升。通过精心设计的实验和广泛的用户评价，验证了所提方法在合成说话人像方面的显著优势。

本文算法尽管取得了不错的进展，但在合成新目标人像时，仍需针对特定个体的短视频进行定制化训练，这突显了提高算法泛化能力在未来研究工作中的必要性。此外，目前的音频特征提取过程依赖于英语自动语音识别模型，这限制了算法在处理非英语音频时的唇部同步准确性，暴露了跨语言应用中的局限性。针对这些挑战，我们计划在未来的研究工作中深入探讨并寻求有效的解决策略。

参 考 文 献

[1] CHUNG J S, JAMALUDIN A, ZISSERMAN A. You said that? [J]. arXiv: 1705.02966, 2017.

[2] CRESWELL A, WHITE T, DUMOULIN V, et al. Generative adversarial networks: An overview[J]. IEEE signal processing magazine, 2018, 35(1): 53-65.

[3] WANG Q Q, ZHANG J L. Face Pose and Expression Correction Based on 3D Morphable Model[J]. Computer Science, 2019, 46(6): 263-269.

[4] TANG Y X, WANG B J. Research Progress of Face Editing Based on Deep Generative Model[J]. Computer Science, 2022, 49(2): 51-61.

[5] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99-106.

[6] GUO Y, CHEN K, LIANG S, et al. Ad-nerf: Audio driven neural radiance fields for talking head synthesis[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition. Montreal: IEEE, 2021: 5784-5794.

[7] XIE Z F, ZHENG J H, WANG J, et al. Speech-Driven Facial Reenactment Guided by Structured Latent Codes in NeRF[J]. Journal of Computer-Aided Design and Graphics, 2023, 41(3): 1003-1015.

[8] ZHENG B W, DONG J W, WU L T, et al. A Method and System for Generating Virtual Anchors Based on Neural Radiance Fields and Hidden Attributes: CN-202311094348. 7[P]. 2023-12-05.

[9] MULLER T, EVANS A, SCHIED C, et al. Instant neural graphics primitives with a multiresolution hash encoding[J]. ACM Transactions on Graphics(ToG), 2022, 41(4): 1-15.

[10] TANG J, WANG K, ZHOU H, et al. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition[J]. arXiv: 2211.12368, 2022.

[11] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]// Proceedings of the Medical Image Computing and Computer Assisted Intervention. Munich: MICCAI, 2015: 234-241.

[12] GU K, ZHOU Y, HUANG T. Flnet: Landmark driven fetching and learning network for faithful talking facial animation synthesis[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 10861-10868.

[13] ZHANG Z, LI L, DING Y, et al. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 3661-3670.

[14] THIES J, ELGHARIB M, TEWARI A, et al. Neural voice puppetry: Audio-driven facial reenactment[C]// Proceedings of the European Conference on Computer Vision. ECCV, 2020: 716-731.

[15] BLANZ V, VETTER T. A morphable model for the synthesis of 3D faces[C]// Proceedings of the Seminal 26th Annual Conference on Computer Graphics and Interactive Techniques. New York: ACM, 1999: 187-194.

[16] LIU X, XU Y, WU Q, et al. Semantic-aware implicit neural audio-driven video portrait generation[C]// Proceedings of the European Conference on Computer Vision. Switzerland: ECCV,

2022：106-125.

[17] SHEN S，LI W，ZHU Z，et al. Learning dynamic facial radiance fields for few-shot talking headsynthesis[C]// Proceedings of the European Conference on Computer Vision. Switzerland：ECCV，2022：666-682.

[18] YAO S，ZHONG R Z，YAN Y，et al. DFA-NeRF：Personalized talking head generation via disentangled face attributes neural rendering[J]. arXiv：2201. 00791，2022.

[19] CHAN E R，LIN C Z，CHAN M A，et al. Efficient geometry-aware 3D generative adversarial networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans：IEEE，2022：16123-16133.

[20] GUO M H，LIU Z N，MU T J，et al. Beyond self-attention：External attention using two linear layers for visual tasks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2022，45(5)：5436-5447.

[21] LI J，ZHANG J，BAI X，et al. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris：IEEE，2023：7568-7578.

[22] ZHOU H，SUN Y，WU W，et al. Pose-controllable talking face generation by implicitly modularized audio-visual representation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville：IEEE，2021：4176-4186.

[23] ZHANG Z，HU Z，DENG W，et al. DINet：Deformation inpainting network for realistic face visually dubbing on high resolution video[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Washington D. C：AAAI，2023：3543-3551.

[24] ZHANG R，ISOLA P，EFROS A，et al. The unreasonable effectiveness of deep features as a perceptual metric[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE，2018：586-595.

[25] HEUSEL M，RAMSAUER H，UNTERTHINER T，et al. Gans trained by a two time-scale update rule converge to a local hash equilibrium[J]. Advances in Neural Information Processing Systems，2017，30(4)：6626-6637.

[26] CHEN L，LI Z，MADDOX R K，et al. Lip movements generation at a glance[C]// Proceedings of the European Conference on Computer Vision. Salt Lake City：ECCV，2018：520-535.

[27] GUAN J，ZHANG Z，ZHOU H，et al. StyleSync：High-fidelity generalized and personalized lip sync in style-based generator[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver：IEEE，2023：1505-1515.

[28] CHUNG J S，ZISSERMAN A. Lip reading in the wild[C]// Proceedings of the Computer Vision Asian Conference on Computer Vision. Waikoloa：IEEE，2017：87-103.

[29] BALTRUSAITIS T，ROBINSON P，MORENCY L P. Openface：An open source facial behavior analysis toolkit[C]// Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision(WACV). Waikoloa：IEEE，2016：1-10.

[30] SUWAJANAKORN S，SEITZ S M，KEMELMACHER S I. Synthesizing Obama：Learning lips sync from audio[J]. ACM Transactions on Graphics(TOG)，2017，36(4)：1-13.

WANG Xingbo，born in 1975，Ph.D，lecturer. His main research interests include robot control and target tracking algorithm.

XIE Jiucheng，born in 1992，Ph.D，lecturer. His main research interests include computer vision and artificial intelligence.

（责任编辑：何杨）