

基于草图引导的少样本说话人视频生成算法研究

魏清扬, 徐树公

(上海大学 通信与信息工程学院, 上海 200444)

摘要: 说话人视频生成需要对面部纹理和驱动语音进行精准联合建模; 为实现该目标, 对语义引导的纹理特征形变进行了研究, 提出一种基于草图引导的少样本说话人视频生成框架, 采用双阶段生成技术进行模态对齐; 在第一阶段使用真实先验关键信息进行语音到目标关键点的生成, 第二阶段将关键点转化为草图作为中间表征与参考图片进行语义对齐; 草图的引入有效地解决了语音与图像的模态不匹配问题; 通过实验测试, 算法在公开数据集 HDTF 和 MEAD 上的 FID 指标达到了 15.676 和 8.618; 经上述结果验证, 提出的算法可通过中间表征有效建模目标音频驱动下的面部纹理, 达到与最先进算法相当的生成效果。

关键词: 高保真生成; 说话人视频生成; 关键点生成; 多模态学习; 音唇同步

Research on Few-Shot Speaker Video Generation Algorithm Guided by Sketches

WEI Qingyang, XU Shugong

(School of Communication & Information Engineering, Shanghai University, Shanghai 200444, China)

Abstract: Speaker video generation requires precise joint modeling of facial texture and driven audio; to achieve this goal, research on semantic-guided texture feature deformation has been conducted, a sketch-guided few-shot speaker video generation framework is proposed, dual-stage generation technique is employed for modality alignment. In the first stage, the information on the real prior facial landmarks is used to generate from the audio to the target facial landmarks, and in the second stage, facial landmarks are transformed into sketches as intermediate representations for semantic alignment with reference images. The introduction of sketches effectively addresses the modality mismatch between audio and images; Through experimental testing, the algorithm achieves the FID scores of 15.676 and 8.618 on the public HDTF and MEAD datasets, respectively. The proposed algorithm effectively models facial texture under the drive of target audio through intermediate representations, achieving a generation performance comparable to state-of-the-art algorithms.

Keywords: high-fidelity generation; talking face generation; landmark generation; multi modal learning; lip synchronization

0 引言

随着人工智能生成内容 (AIGC) 领域技术的快速发展以及计算硬件性能的显著提升, 基于神经网络的说话人视频生成算法取得了突破性进展, 能够生成更为真实且具有高保真度的人脸视频。这些算法在不少样本场景下表现出色, 因为它们通常只需要较少的数据输入并展示出较高的泛化能力, 使得它们在科研和工业应用中受到了广泛的关注。少样本说话人视频生成技术主要依赖于从少量的样本中学习并生成高质量的视频内容, 目的是通过让一组参考帧进行形变来模仿源视频的头部姿态, 同时确保唇形与输入的驱动音频保持一致。

当前学者们已经提出了多种引导视频生成的方法。例如, Wav2Lip^[1]模型采用复杂的编码器-解码器架构, 并在解码器部分引入了一种融合机制来整合视频帧和音频片段

的特征, 从而有效实现纹理的形变和视频内容的生成。PC-AVS^[2]算法通过隐式模块化音频-视觉表示, 实现对不同说话人的姿态控制。EAMM^[3]算法基于无监督学习的运动表示, 通过情绪控制模块生成可以表达不同情绪的说话人视频。此外, DINet^[4]网络利用视频帧特征与音频特征的相互引导, 实现在特征层面的纹理变换, 以提升生成视频的真实感和细节丰富度。在这些视频生成方法中, 针对音频和视频特征的引导大多集中在上半张脸的特征和音频内容生成方面, 这会使网络难以成功学习全脸信息, 导致上下脸衔接处的不匹配现象, 导致生成视频的下半部分与音频内容之间存在不连贯或不自然的联系, 降低了视频的整体质量和真实感。

因此, 一些方法尝试采用中间表征的策略, 这些中间表征能为视频生成提供必要的人脸结构信息, 使得模型关注视频中特定位置或区域的变化和重要性, 更好地控制视

收稿日期: 2024-04-28; 修回日期: 2024-05-09。

基金项目: 国家自然科学基金(61871262)

作者简介: 魏清扬(2000-), 女, 硕士研究生。

通讯作者: 徐树公(1969-), 男, 博士, 教授。

引用格式: 魏清扬, 徐树公. 基于草图引导的少样本说话人视频生成算法研究[J]. 计算机测量与控制, 2024, 32(10): 236-242, 249.

频内容的演变过程, 进一步辅助生成更加准确和细致的视频。例如, MakeIttalk^[5]使用自注意力机制和 LSTM^[6]模型来预测音频中人脸关键点的位移, 从而引导说话过程中面部表情的动态变化, 这种方法有效地模拟了说话者自然的面部动画。又如 Apb2Face^[7]使用线性结构从音频特征、姿势和眨眼预测关键点, 一定程度上提高了视频生成的自然度和真实感。IP_LAP^[8]算法利用关键点信息作为中间特征, 从音频中生成具有身份保留特性的说话人面部视频。然而, 直接引入关键点作为中间表征也有其局限性, 引入关键点之后, 如果只是简单地连接驱动信息而不进行各模态的空间对齐, 将会限制从参考图像中提取有意义特征的能力。

针对上述问题, 本文提出双阶段说话人视频生成算法, 该算法由关键点生成阶段和视频帧生成阶段组成。本文引入的关键点生成阶段, 通过有限的先验关键点信息来预测目标关键点, 为后续视频帧的生成提供更多中间引导特征; 引入的草图转化和对齐模块, 能够强化对说话者面部轮廓生成的约束。这两个模块的设计能够实现形变网络中特征纹理的细粒度对齐。

1 算法设计

本文所提基于中间特征引导的双阶段说话人视频生成算法, 旨在通过一段驱动音频和一段原始视频生成与驱动音频一致的说话人视频。DINet 是一个少样本人脸视频生成的算法, 该算法引入的基于自适应注意力转换 (AdaAT) 算子^[9], 该算子能够模拟复杂的空间变形, 实现未对齐图像的高保真变形。在该模块的基础上, 本文提出了双阶段生成算法, 第一阶段为音频驱动的关键点生成阶段, 第二阶段为音频和关键点特征共同驱动的视频生成阶段, 下文首先对算法整体进行介绍, 再对两个阶段涉及的细节进行更深入地讨论。

1.1 整体流程

本文提出的双阶段生成框架的详细流程如图 1 所示。模型的输入包括一系列原始视频帧及相应的驱动音频。在第一阶段中, 利用 Mediapipe^[10]工具对原始视频进行逐帧分析, 以提取面部的关键点, 其中特别将上半部分的面部关键点作为姿态先验, 同时随机选取五帧全脸的关键点作为参考。此外, 本研究采用 DeepSpeech^[11]模型来提取音频中的特征。这些关键点信息与音频特征共同构成了第一阶段的输入, 随后通过关键点生成网络来产生目标帧的关键点数据。进入第二阶段, 系统对视频进行逐帧分析处理, 选取上半部分的面部图像作为目标帧的上半面部图像。借助于第一阶段提取的音频特征以及新生成的目标帧关键点, 与随机选取的五帧全脸参考图像共同指导目标视频的空间形变处理, 最终生成所需的目标视频。

1.2 关键点生成阶段

在关键点生成阶段, 整体网络结构如图 2 所示, 本文参考 IP_LAP^[8]的关键点生成结构并做针对性改进, 目标是在给定姿态先验关键点 $\{I_p \in R^{2 \times n_p}\}_{t=1}^T$ 和驱动音频 $\{a'_d \in$

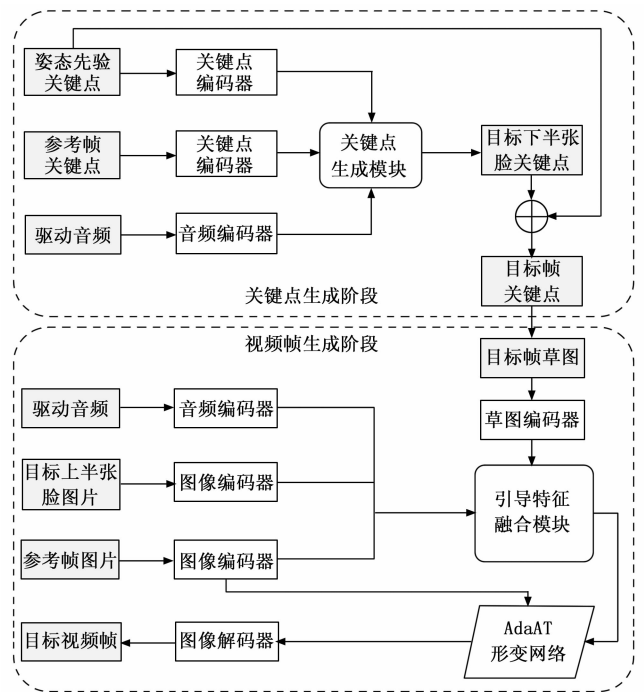


图 1 双阶段生成框架流程图

$R^2\}_{t=1}^T$ 的情况下, 结合参考帧关键点 $\{I_r \in R^{2 \times n_r}\}_{t=1}^N$ 提供的说话人身份特征, 生成目标帧的下半张脸关键点 $\{\tilde{I}_d \in R^{2 \times n_d}\}_{t=1}^T$ 。其中, T 为每次生成的连续帧数量, 实验中该数值为 5; N 为参考关键点的帧数量, 本文中每生成一帧目标关键点, N 的取值都为 5; n_r 、 n_p 、 n_d 分别为每帧参考帧关键点的数量、每帧姿态先验关键点的数量、每帧目标下半张脸关键点的数量, 在本文中, 数值分别为 131、74、57。对于每一帧, 本文都基于 DeepSpeech 预训练模型提取对应的音频特征 a'_d , 将其转化为尺寸 1×29 的向量。

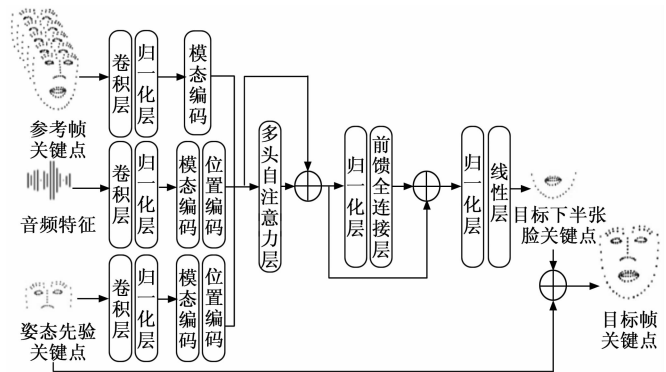


图 2 关键点生成网络

首先, 本文使用由卷积层和归一化层组成的编码器模块 E_p 和 E_r 对姿态先验关键点和参考帧关键点进行编码处理, 以提取姿态嵌入 p_i 和参考嵌入 r_i 。同时, 音频数据也经过卷积编码模块 E_a 进行编码, 得到对应的音频嵌入 a_i 。该过程如公式 (1) ~ (3) 所示:

$$p_i = E_p(I_p), i = 1, 2, \dots, T \quad (1)$$

$$r_i = E_r(I_r), i = 1, 2, \dots, N \quad (2)$$

$$a_t = E_a(a'_t), t = 1, 2, \dots, T \quad (3)$$

其中: $p_t, r_t, a_t \in R^d$, d 为这些嵌入向量的维度, 实验中设置为 512。

然后, 对姿态、参考、音频嵌入分别引入 3 个可学习向量 $e_p^{cyp}, e_r^{cyp}, e_a^{cyp} \in R^d$, 以区分输入的嵌入向量类型。此外, 本文对姿态嵌入和音频嵌入引入可学习向量 $e_t^{pos} \in R^d$, 以表示姿态关键点和驱动音频的第 t 帧时间位置编码, 该向量的计算遵循正弦位置编码^[12]。这些编码变量被添加到姿态、参考和音频嵌入中, 具体如公式 (4) (5) (6) 所示:

$$\bar{p}_t = p_t + e_p^{cyp} + e_t^{pos}, t = 1, 2, \dots, N \quad (4)$$

$$\bar{r}_i = r_i + e_r^{cyp}, i = 1, 2, \dots, N \quad (5)$$

$$\bar{a}_t = a_t + e_a^{cyp} + e_t^{pos}, t = 1, 2, \dots, N \quad (6)$$

3 种输入特征 (姿态先验关键点、驱动音频、参考帧关键点) 被融合后, 本文使用自注意力模块来捕获 3 种类型的嵌入之间的类内和类间关系。初始令牌 z^0 通过连接 $\{\bar{p}_t\}_{t=1}^T, \{\bar{r}_i\}_{i=1}^N, \{\bar{a}_t\}_{t=1}^T$ 形成。先经过多头自注意力层 (MSA), 以细致地捕获 3 种类型的复杂关系。融合后的特征与多头自注意力层的输出进行残差连接, 以确保信息的有效传递, 避免深层网络中梯度消失或梯度爆炸的问题, 此外, 归一化层 (LN) 也被应用, 以保持特征的稳定性和均衡, 确保模型更好地学习到特征的分布。具体如公式 (7) 和公式 (8) 所示:

$$\bar{z}^l = MSA[LN(z^l - 1)] + z^l, l = 1, \dots, L \quad (7)$$

$$\bar{z}^l = MLP[LN(\bar{z}^l - 1)] + \bar{z}^l, l = 1, \dots, L \quad (8)$$

其中: $z^l \in R^{(N+2T) \times d}$ 表示第 l 层 Transformer 结构的输出, 本文一共用了 $L=6$ 层 Transformer 结构。

最后, 特征被输入到由多层感知机 (MLP) 组成的前馈全连接层, 以加强对融合后特征的学习, 并通过残差连接确保深层次的非线性学习。这个过程有助于模型更全面地理解 3 种特征的融合后所带来的信息, 并提高模型的建模能力和泛化能力。通过这一系列处理步骤, 融合后的特征得到了充分的提取和学习, 为模型预测目标帧的下半张脸关键点提供了有效的信息支持和基础, 具体如公式 (9) 所示:

$$\hat{l}_d^t = MLP(z_{t+N+T}^l), t = 1, 2, \dots, T \quad (9)$$

将预测出的目标帧下半张脸关键点和姿态先验关键点进行结合, 即可得到姿态一致且唇部运动符合音频发声的目标帧全脸关键点 $\{\hat{I}_w \in R^{2 \times n_w}\}_{t=1}^T$, 其中 $n_w=131$ 为每一帧全脸关键点的数量。

1.3 视频帧生成阶段

在本阶段中, 本文使用了一个中间表征引导的策略, 整体网络如图 3 所示, 目标是结合参考帧图像 $\{I_r^t \in R^{3 \times h \times w}\}_{t=1}^N$ 提供的说话人身份特征, 在给定目标上半张脸图像 $\{I_s^t \in R^{3 \times h \times w}\}_{t=1}^T$ 、驱动音频 $\{A_d^t \in R^1\}_{t=1}^T$ 和章节 1.2 生成的目标帧关键点 $\{\hat{I}_w \in R^{2 \times n_w}\}_{t=1}^T$ 的情况下, 生成目标帧图像 $\{\hat{I}_{gen} \in R^{3 \times h \times w}\}_{t=1}^T$ 。其中, 图像的尺寸为 $h \times w$, 本文中数值均为 128; T 为每次生成的连续帧数量, 本文中数值为 5; N 为参考帧图像的数量, 本文中每生成一帧目标图像, N 的

取值都为 5。对于每一帧的生成, 本文采用 DeepSpeech^[13] 预训练模型提取对应的音频特征, 将其转化为 1×29 的维度。此外, 本文在图像平面将 1.2 生成的目标帧关键点坐标绘制为目标草图 $\{\hat{I}_{skt} \in R^{3 \times h \times w}\}_{t=1}^T$, 这不仅能够更直观地应用关键点来表示面部的基本结构, 而且能够使上下半脸之间的过渡更自然, 并且实现关键点模态与图像模态的初步对齐, 提升最终生成结果的一致性和真实感。

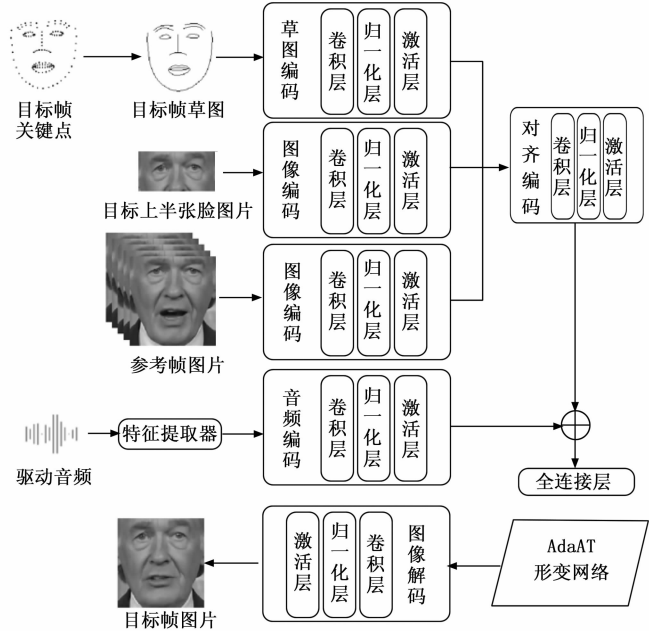


图 3 视频帧生成网络

本阶段分别对输入信息进行编码, 每个编码器都由多组卷积层、归一化层和激活层组成的下采样网络。首先使用音频编码器从 A_d 中提取音频特征 $F_{audio} \in R^{128}$; 然后将目标草图 \hat{I}_{skt} 输入到草图编码器中, 得到草图特征 $F_{skt} \in R^{128}$ 。最后, 将 I_s 和 I_r 输入到两个独立的图像特征编码器中, 分别提取源特征 $F_s \in R^{256 \times h/4 \times w/4}$ 和参考特征 $F_r \in R^{256 \times h/4 \times w/4}$ 。接下来, 对草图特征 F_{skt} 、参考帧图像特征 F_r 和目标上半张脸特征 F_s 通过对齐编码器进行模态内对齐, 得到图像维度上的对齐特征 $F_{align} \in R^{128}$ 。最后将 F_{align} 和 F_{audio} 融合后的特征用以引导 F_r 空间变形为目标帧的特征 F_d 。

F_{align} 和 F_{audio} 融合后的特征首先通过全连接层计算出 AdaAT 网络放射变换需要的旋转、平移和尺度系数, 最后将这些系数 F_r 和一起送入 AdaAT 网络中, 生成形变后的目标帧特征 F_d , 形变网络具体涉及过程如公式 (10) 所示:

$$\begin{bmatrix} \hat{x}_c \\ \hat{y}_c \end{bmatrix} \begin{bmatrix} s^c \cos(\theta^c) & s^c [-(\sin\theta^c)] \\ s^c \sin(\theta^c) & s^c \cos(\theta^c) \end{bmatrix} \begin{bmatrix} t_x^c \\ t_y^c \end{bmatrix} = \begin{bmatrix} x_c \\ y_c \\ 1 \end{bmatrix} \quad (10)$$

其中: x_c, y_c 和 \hat{x}_c, \hat{y}_c 分别表示仿射变换前后的像素坐标。 c 表示 F_r 中的第 $c \in [1, 256]$ 个通道。

得到目标帧的特征 F_r 之后, 使用同样由多组卷积层、归一化层和激活层组成的上采样解码器, 将特征解码为图像, 输出生成的视频帧 $\{I_o \in R^{3 \times h \times w}\}_{t=1}^T$ 。

1.4 损失函数

在第一阶段, 所提框架使用 L_1 损失和正则化损失训练模型。

首先, 运用 L_1 损失来约束预测的目标帧关键点, 使其接近真实目标帧关键点, 计算如公式 (11):

$$L_1 = \frac{1}{T} \sum_{t=1}^T (\| \hat{l}_d^t - l_d^t \|_1), t = 1, 2, \dots, T \quad (11)$$

其中: l_d^t 表示目标帧下半张脸的真实关键点。

此外, 本文还在时序上进行正则化约束, 以提高生成关键点在前后帧间的平滑程度, 计算如公式 (12) 所示:

$$L_c = \frac{1}{T-1} \sum_{t=1}^{T-1} \| (\hat{l}_d^{t+1} - \hat{l}_d^t) - (l_d^{t+1} - l_d^t) \|_2, \quad t = 1, 2, \dots, T \quad (12)$$

综上所述, 关键点生成阶段的总体损失函数为公式 (13):

$$L = L_1 + L_c \quad (13)$$

在第二阶段, 所提框架使用 GAN 损失^[13]、感知损失^[14]和唇形同步损失来训练模型^[1]。

首先, 使用 GAN 损失来提高生成图像 I_o 的真实度, 使用鉴别器将一系列的下采样特征提取模块和卷积层将输入图像转换为一个可以表示其真伪概率的标量值, 从而实现对生成图像的真实性的评估。GAN 损失包括生成器损失 L_G 和鉴别器损失 L_D , 具体过程如公式 (14) (15) 所示:

$$L_G = E[D(I_o) - 1]^2 \quad (14)$$

$$L_D = \frac{1}{2} E[D(I_{\text{real}}) - 1]^2 + \frac{1}{2} E[D(I_o) - 0]^2 \quad (15)$$

其中: G 表示第二阶段网络, D 表示鉴别器, I_{real} 表示与生成图像 I_o 匹配的真实图像。

此外, 本阶段在单帧和 5 个连续帧上使用 GAN 损失。总体的 GAN 损失为公式 (16):

$$L_{\text{GAN}} = L_G + L_D \quad (16)$$

同时, 本阶段对真实图像和生成图像使用感知损失, 以进一步提高生成真实度。具体来说, 将生成视频帧 I_o 与真实视频帧 I_{real} 分别下采样到 $\hat{I}_o, \hat{I}_{\text{real}} \in R^{3 \times k/2 \times w/2}$, 然后将图像对 $\{\hat{I}_o, \hat{I}_{\text{real}}\}$ 和 $\{I_o, I_{\text{real}}\}$ 输入到预训练的 VGG-19^[15] 网络中计算感知损失。具体计算如公式 (17) 所示:

$$L_p = \sum_{i=1}^N \frac{\|V_i(I_o) - V_i(I_r)\|_1 + \|V_i(\hat{I}_o) - V_i(\hat{I}_{\text{real}})\|_1}{2NW_i H_i C_i} \quad (17)$$

其中: $V_i(\cdot)$ 表示第 i 层 VGG-19 网络, $W_i H_i C_i$ 是第 i 层 VGG-19 网络的特征尺寸大小。

最后, 为了约束唇部变化与驱动音频的一致性, 和 Syncnet^[13] 网络相似, 本文引入唇形同步专家鉴别器, 具体计算如公式 (18) 所示:

$$L_{\text{sync}} = E(\text{syncnet}(A_d, I_o) - 1)^2 \quad (18)$$

综上, 视频帧生成阶段所用的总体损失函数为公式 (19):

$$L = L_{\text{GAN}} + \lambda_p L_p + \lambda_{\text{sync}} L_{\text{sync}} \quad (19)$$

其中: λ_p 和 λ_{sync} 分别为感知损失 L_p 和唇形同步损失

L_{sync} 的权重, 在本文设置 0.5。

2 实验与分析

2.1 数据集

为了验证所提算法的有效性, 本文选择两个公开视听数据集 HDTF^[16] 和 MEAD^[17] 进行训练和测试, 在这两个数据集上能与现有算法进行全面且公平的对比。HDTF 数据集包含超过 300 个不同身份的人的谈话视频, 视频质量为 720 P 或 1 080 P, 总时长达到 16 小时。为达到实验目的, 本研究从中随机选取了 220 个各为一分钟长的视频片段, 合成一个总长为 220 分钟的视频库。其中包含了 200 分钟的视频用于训练集, 以及 20 分钟的视频用作测试集。MEAD 数据集是一个实验室环境下收集的面部表情语料库, 包括 60 名说话者在 3 种不同的情绪强度下的视频, 涵盖 8 种不同的情绪。本研究主要关注无表情的、正面的说话视频, 因此筛选出了 650 个具有中性表情的视频作为训练集, 以及 150 个视频作为测试集。

HDTF 和 MEAD 数据集均提供了视听数据, 但可能存在一些低质量或短时长的视频, 这可能会对网络训练产生干扰。因此在选用数据时, 本文首先用 Syncnet 网络检查了视频和对应音频的同步性, 选取长度在 3 秒以上, 且视听延迟度小于 3 帧的视频进行训练, 且为了统一训练和推理标准, 本文将所有视频的帧率转为 25 帧/秒。接着, 本文通过 Mediapipe^[10] 工具对每帧进行人脸检测和面部关键点提取, 最终将视频帧裁剪为 128×128 尺寸的人脸图片, 并且保存每张图片对应的 131 个人脸关键点坐标。此外, 本文预先用 Deepspeech^[11] 工具提取出每帧视频对应的音频特征。最后, 将裁剪得到的人脸图片、保存的关键点坐标和提取出的音频特征一起用于所提生成框架的训练和测试。

2.2 实施细节

实验分为训练和测试两个阶段。训练阶段, 对于关键点生成和视频帧生成网络, 均使用 Adam^[18] 优化器进行模型参数优化。关键点生成模型学习率 0.000 1, 训练对应批处理尺寸为 64; 视频帧生成模型中生成网络和鉴别网络的学习率均为 0.000 02, 对应批处理尺寸为 32。测试阶段, 本文在两个数据集中分别选取测试集进行测试, 并确保选用的视频和视频中的人物身份没有出现在训练数据中, 同时对测试数据使用与训练数据相同的处理方法。在推理时, 目标人物的初始视频作为参考视频, 根据输入的驱动音频来合成对应的说话视频。针对所提的双阶段说话人视频生成算法, 本文选用峰值信噪比 (PSNR)、结构相似性指数 (SSIM)^[19]、弗雷歌特初始距离 (FID)^[20]、学习型感知图像块相似度 (LPIPS)^[21] 4 个客观指标评估视觉真实度; 选用唇形同步误差距离 (LSE-D)^[1] 和唇形同步误差一致性 (LSE-C) 两个客观指标评估唇形同步效果。除此之外, 平均意见分数 (MOS) 被选用, 以进行输出视频在视觉真实度、唇形同步一致性质量上的主观评估。

2.3 可视化结果与对比分析

本节首先对所提双阶段说话人视频生成算法进行可视

化输出, 结果如图 4 所示。图中从上到下呈现的顺序是真实视频帧、第一阶段生成的关键点、草图以及第二阶段生成的视频帧。通过该图可以观察到, 本文提出的算法在双阶段生成过程中能够达到预期效果。它不仅能够根据原始图片和驱动音频生成目标帧的关键点, 准确地反映目标帧的头部姿态和嘴部动作, 还能够通过这些目标帧关键点连成的草图生成目标视频, 且生成视频的真实性和唇形同步准确性均能达到理想水平。

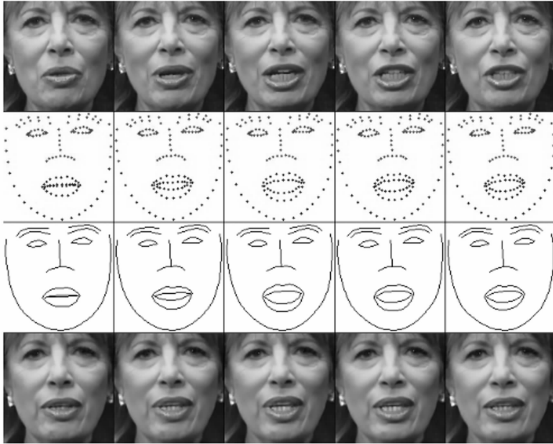


图 4 双阶段生成算法可视化结果

从上到下依次是真实视频帧、第一阶段生成结果、草图、第二阶段生成结果。

此外, 本文将所提双阶段说话人视频生成算法与其他 5 个先进算法进行可视化对比, 包括 Wav2Lip^[1]、PC_AVS^[2]、EAMM^[3]、IP_LAP^[8]、DINet^[4]。定性对比实验结果如图 5 和图 6 所示, 由此可得不同模型的生成效果差异。

第一行为连续 5 张真实视频帧图片, 后续每行依次为 Wav2Lip、PC_AVS、EAMM、IP_LAP、DINet、本文算法的生成效果。

第一行为连续 5 张真实视频帧图片, 后续每行依次为 Wav2Lip、PC_AVS、EAMM、IP_LAP、DINet、本文算法的生成效果。

在对两个数据集的生成效果进行深入分析后, 可以发现相同的生成规律和问题。具体来看, 与第一行展示的真实视频帧相比, 第二行中 Wav2Lip 模型生成的视频帧分辨率较低, 并且在人物姿态变化时, 容易出现视觉伪影。第三行的 PC_AVS 模型在生成的姿态方面显得尤为僵硬, 缺乏自然流畅性。EAMM 模型在第四行中展示, 该模型生成的视频帧中头部动作和表情都较为僵硬, 需要更精细的姿态控制来实现更加自然的生成效果。而 IP_LAP 模型, 如第五行所示, 其生成的图像常常存在模糊问题, 且会在下巴区域产生伪影, 同时其唇形同步的效果也未能达到理想状态 (参见图 5 和图 6 的矩形区域)。第六行 DINet 模型的生成结果, 除了唇形同步不准确之外, 还可能会在原始真实视频与编辑区域之间产生显著的伪影, 并伴随着肤色差异的问题 (参见图 6 的矩形区域)。

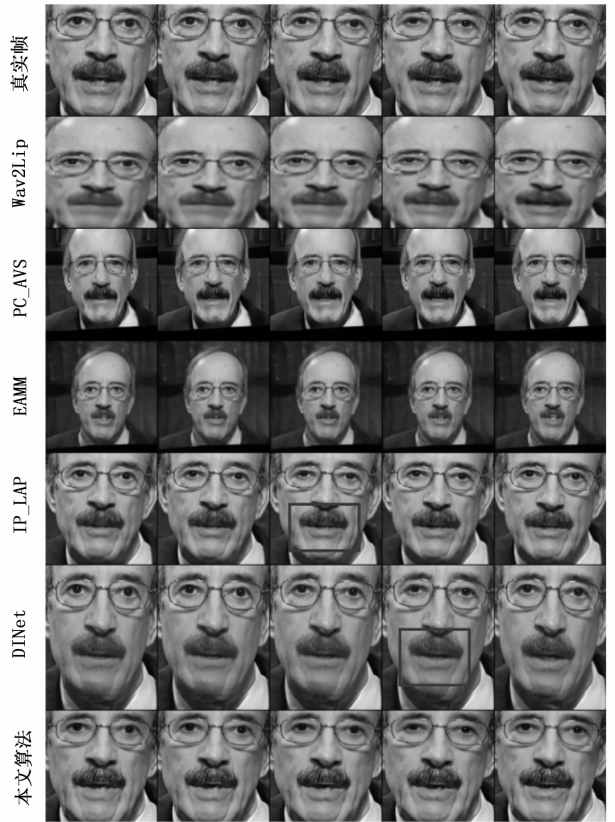


图 5 HDTF 数据集定性对比图

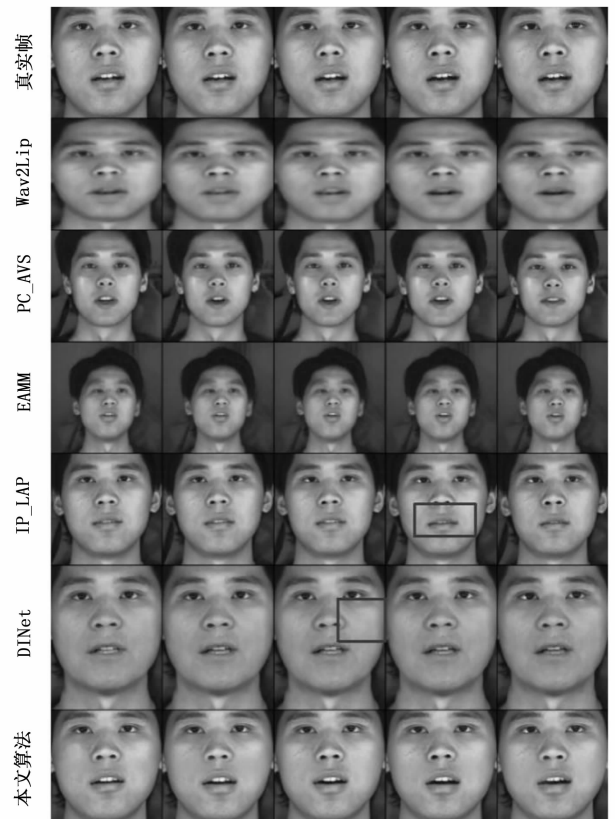


图 6 MEAD 数据集定性对比图

通过综合分析, 本研究提出的双阶段生成模型在提升图像生成质量方面表现出色。该模型不仅在视觉清晰度上有所提升, 更在细节处理上展现了独特的优势。相较于其他模型, 双阶段生成模型的生成效果与真实视频帧的接近度更高, 尤其在唇形同步的准确性和面部表情和姿态的自然性上, 显示出明显的改进。

2.4 定量对比实验结果与分析

本文将双阶段说话人视频生成算法与其他 5 个先进算法进行了定量对比。为了确保公平比较, 所有说话人视频生成算法都使用已经公开的数据集模型进行对比。对于那些没有针对本文数据集进行训练的算法, 本文均严格遵循各自论文中所述的实验设置和数据处理方法进行复现, 并使用相同的测试视频进行推理。具体来说, Wav2Lip、IP_LAP 以及本文提出的模型生成了 128×128 像素大小的面部视频帧, 然后将这些视频帧与音频结合起来, 组合成了大小为 $T \times 128 \times 128$ 的目标视频, 其中 T 为视频的帧数。相应地, DINet 也遵循了类似的生成步骤, 但其生成的面部图像尺寸为 416×320 。另一方面, PC_AVS 和 EAMM 模型首先对输入进行了对齐, 然后沿用了类似的步骤, 生成了大小为 $T \times 256 \times 256$ 的说话人视频。

首先从视觉真实性的角度对说话人视频生成算法进行客观评估, 表 1 和表 2 分别展示了在 HDTF 和 MEAD 数据集上的评估结果。表中加粗的数字表示最佳模型的指标, 而添加了下划线的数字则表示排名第二模型的指标。

表 1 HDTF 数据集视频真实度评估结果

算法名称	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	FID \downarrow
Wav2Lip	29.213	0.943	0.071 1	25.972
PC_AVS	23.450	0.640	0.123 8	37.412
EAMM	16.546	0.398	0.363 8	85.810
IP_LAP	<u>30.254</u>	0.924	0.029 3	19.852
DINet	30.142	<u>0.931</u>	<u>0.027 9</u>	<u>17.936</u>
本文算法	30.547	0.940	0.019 3	15.676

表 2 MEAD 数据集视频真实度评估结果

算法名称	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	FID \downarrow
Wav2Lip	28.729	0.886	0.045 2	24.151
PC_AVS	25.548	0.758	0.108 3	72.411
EAMM	17.72	0.356	0.242 0	108.69
IP_LAP	30.486	<u>0.915</u>	0.034 3	19.457
DINet	<u>30.967</u>	0.909	<u>0.025 2</u>	<u>9.222</u>
本文算法	31.185	0.926	0.018 7	8.618

从表中可以看出, 在 HDTF 和 MEAD 两个数据集上, 本文提出的算法在视觉真实性方面表现都更加出色。尤其是在 LPIPS 指标上, 相较于排名第二的 DINet 算法, 本文提出的算法在两个数据集上分别提升了 30.82% 和 25.79%; 而在 FID 指标上, 相较于 DINet 算法, 本文提出的算法分别提升了 12.60% 和 6.55%; 其他指标也获得一定程度的提升。

实验中还对说话人视频生成算法进行唇形同步质量的客观评估, 结果在表 3 中得到了详细展示, 分别对应 HDTF 和 MEAD 数据集。特别值得一提的是 Wav2Lip 算法, 该算法专注于提升唇形同步的准确性, 得到了学术界的广泛认可, 而忽略对整体真实度的优化。通过对比分析可得, 相比于除 Wav2Lip 外的其他算法, 在 HDTF 和 MEAD 两个数据集上, 本文提出的算法在唇形同步质量方面均取得了显著的改进。在 LSE-D 指标上, 与排名第二的 DINet 算法相比, 本文提出的算法在 HDTF 和 MEAD 两个数据集上分别实现了 2.32% 和 1.72% 的性能提升。此外, 在 LSE-C 指标上, 本文提出的算法同样表现出色, 比 DINet 算法在 HDTF 数据集上提升了 2.19%, 比 IP_LAP 算法在 MEAD 数据集上提升了 2.59%。这些结果充分证明了本文算法在提升说话人视频生成质量方面的有效性和优越性。

表 3 唇形同步准确性评估结果

算法名称	HDTF		MEAD	
	LSE-D \downarrow	LSE-C \uparrow	LSE-D \downarrow	LSE-C \uparrow
Wav2Lip	6.908	8.443	7.092	7.795
PC_AVS	8.389	6.819	8.546	6.109
EAMM	9.241	4.918	9.196	4.842
IP_LAP	7.883	7.063	8.903	6.527
DINet	7.801	7.091	8.773	6.130
本文算法	<u>7.620</u>	<u>7.213</u>	<u>8.581</u>	<u>6.696</u>

接着, 对生成视频的视觉真实性和唇形同步准确性进行主观对比。为了保证测试结果的稳定, 本文在不同时间对不同的受访者进行测试, 共有来自 30 名受访者参与, 对上述 6 种方法进行打分。本打分系统从 HDTF 和 MEAD 两个数据集中各选了 5 个视频进行评估, 每个指标的范围从 1 到 5, 数值越高则生成效果越好, 真实视频默认为 5 分。表 4 列出了所有受访者主观性评估的平均意见分数, 本文所提方法获得了最高的评分, 比 IP_LAP 方法在视觉真实性上有 4.2% 的性能提升, 比 DINet 方法在唇形同步准确性上提高了 9.39%。由于 Wav2Lip 在视觉真实性方面的不足, 该方法的唇形同步观感也不佳。

表 4 主观评测结果

算法名称	视觉真实性	唇形同步准确性
Wav2Lip	5.000	5.000
PC_AVS	2.029	2.630
EAMM	2.179	2.395
IP_LAP	1.810	2.889
DINet	<u>3.039</u>	3.041
本文算法	2.991	<u>3.227</u>

上述 3 个实验结果表明, 无论是客观指标还是主观观感, 本文所提说话人视频生成算法生成的视频在真实性上均具有显著的效果提升, 且指标在性能提升上具有一致性。相比于其他工作, 本文所提算法在视觉真实度和唇形同步准确性上取得了卓越的平衡, 能够给观众带来更真实、更

细腻的视觉感受。

2.5 中间表征类型消融实验

本文使用 HDTF 数据集验证不同中间表征类型对所提算法的影响,设置了一组关键点引导和草图引导的对比实验。具体来说,本文在前文实验的基础上,直接人脸将关键点作为向量输入,省略关键点转换为草图的步骤,探究不同中间表征对齐模块对生成结果的作用力度。首先对实验结果进行定性评估,对应实验结果如图 7 所示,由图中矩形区域可以看出,使用关键点引导时,上下脸会出现明显分层现象,且脸部和唇周轮廓不够明显,得益于草图对全脸轮廓的强调,生成的视频帧中面部纹理的衔接显得更为自然,同时上下脸的整体协调性也得到了显著提升。



图 7 消融实验对比图

随后,对消融实验结果进行定量评估,对应视觉真实度和唇形同步准确性结果分别如表 5 和表 6 所示,分析可得,当中间表征由关键点转换为草图后,在视觉真实度和唇形同步准确性上总体都取得了较为明显的效果提升,其中 FID 指标提升了 11.35%,本实验结果跟定性实验存在相同规律。

表 5 消融实验视觉真实度对比评估

表征类型	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
关键点	30.497	0.929	0.024 3	9.721
草图	31.185	0.926	0.018 7	8.618

表 6 消融实验唇形同步准确性对比评估

表征类型	LSE-D ↓	LSE-C ↑
关键点	7.924	6.851
草图	7.620	7.213

3 结束语

本文介绍了一种基于草图特征引导的双阶段说话人视频生成框架。该框架首先通过第一阶段预测目标人脸关键点,然后将这些关键点作为中间表征,引导第二阶段视频帧图像的生成,从而加强全脸轮廓的完整性与连贯性。实验结果显示,与其他同类算法相比,本研究所提出的算法在 HDTF 和 MEAD 两个不同数据集上分别实现了 12.60% 和 6.55% 的视觉真实性提升。同时,它也实现了视觉真实性与

唇形同步准确性之间的有效平衡,显著优化了视频的整体观感。此外,经过实验验证,草图引导和对齐模块的运用在全脸轮廓生成中起到了关键作用,进一步提升了生成图像的自然度和细腻度。

参考文献:

- [1] PRAJWAL K R, MUKHOPADHYAY R, NAMBOODIRI V P, et al. A lip sync expert is all you need for speech to lip generation in the wild[C]//Proceedings of the 28th ACM International Conference on Multimedia,2020: 484 - 492.
- [2] 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,2021: 4176 - 4186.
- [3] JI X, ZHOU H, WANG K, et al. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model[C]//ACM SIGGRAPH 2022 Conference Proceedings,2022: 1 - 10.
- [4] 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, 2023, 37(3): 3543 - 3551.
- [5] ZHOU Y, HAN X, SHECHTMAN E, et al. Makeltalk: speaker-aware talking-head animation [J]. ACM Transactions On Graphics (TOG), 2020, 39(6): 1 - 15.
- [6] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735 - 1780.
- [7] ZHANG J, LIU L, XUE Z, et al. Apb2face: Audio-guided face reenactment with auxiliary pose and blink signals [C]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 4402 - 4406.
- [8] ZHONG W, FANG C, CAI Y, et al. Identity-preserving talking face generation with landmark and appearance priors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2023: 9729 - 9738.
- [9] ZHANG Z, DING Y. Adaptive affine transformation: a simple and effective operation for spatial misaligned image generation [C]//Proceedings of the 30th ACM International Conference on Multimedia,2022: 1167 - 1176.
- [10] LUGARESI C, TANG J, NASH H, et al. Mediapipe: a framework for building perception pipelines[J]. Arxiv Preprint Arxiv:1906.08172, 2019.
- [11] HANNUN A, CASE C, CASPER J, et al. Deep speech: scaling up end-to-end speech recognition[J]. Arxiv Preprint Arxiv:1412.5567, 2014.
- [12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017:30.
- [13] MAO X, LI Q, HAORAN X, et al. Least squares generative adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 2794 - 2802.

(下转第 249 页)