音频驱动的情感强度可控面部动画生成模型

阎文博[1]
YAN Wenbo

摘 要 音频驱动的面部动画生成技术是计算机视觉领域的新兴方向，旨在利用音频驱动源将静态肖像转化为动态视频。这项技术可以降低视频制作时间与成本，并创造出独一无二的虚拟人形象。虽然现有研究在动画的唇音同步和面部表情生成上有一定进展，但并未关注面部表情的强度问题，生成的面部动画标签受驱动因子所限，表情较为单一。因此，提出了一种音频驱动的情感强度可控面部动画生成模型，该模型是由 3 个子网络组成的级联结构非端到端算法，旨在生成唇音对齐且情感强度可控的面部动画。在传统技术的基础之上，利用情感强度标签的映射，增强情感源肖像的驱动能力，从而控制生成 3 种情感强度不同的面部表情。定性和定量实验结果表明， 所提方法在情感表达与情感强度控制上，具有一定的可行性与优越性。

关键词 计算机视觉；面部动画生成技术；面部表情生成

doi：10.3969/j.issn.1672-9528.2025.08.038

0 引言

基于音频驱动的面部动画生成技术是计算机视觉领域新兴的科研方向，旨在通过音频驱动源（或视频驱动源）的作用，将静态的目标肖像转化为动态的、唇音同步的视频 [1]。在该面部动画生成技术的加持下，普通用户可以节省视频录制与制作的时间；媒体公司一定程度上节省影视制作成本，甚至可以创作独特的虚拟人形象。如何精准地生成可信度高、灵活性强且生动形象的面部动画视频，是一个极具科研价值和经济意义的研究方向。

近几年，随着图像到图像转换技术 [2-3] 的发展，面部动画生成技术以图形为出发点的传统方法 [4-5]，向泛化能力更强、灵活性更高以音频为驱动源的方法转变。而高质量且具有情感标签视听数据库 [6] 的提出，让科研人员有机会探索动态视频中面部表情的生成与控制。

现有情感可控的面部动画生成技术可以根据算法范式分为两类，即具有完整性输入输出的端到端算法范式 [7-8]，和被解耦为多个中间步骤的非端到端算法范式 [9-10]。对于端到端算法范式，Fang 等人 [7] 通过音频驱动源直接生成动画的面部表情，而 Liang 等人 [8] 选择额外添加驱动因子，以控制肖像表情的变化。然而，端到端算法本身对超参和网络架构具有较高的敏感性，对于未见过的数据表现较差。另一方面，对于具有级联结构的非端到端算法，Ji 等人 [9] 和 Agarwal 等人 [10] 通过引入面部运动关键点 [11] 作为模型的中间表示，辅助动画的生成与面部表情的控制。然而，上述方法虽然可以生成具有情感的面部动画，但表情的强度受驱动因子等因素的限制，情感可控性较低，很难在面部表情需求过大的场景上使用。

针对上述问题，本文提出基于音频驱动的情感强度可控的面部动画生成模型。该模型共包括 4 个输入，即无面部表情的目标肖像、用于驱动情感的情感肖像、用于驱动唇部动作的音频以及控制并协调情感强度的情感强度标签，输出则是唇音同步且情感强度可控的面部动画。本文模型设计成级联结构的非端到端算法，使用结构化的面部特征坐标作为中间表示，模型共包含唇音特征坐标预测网络、面部特征坐标预测网络和视频生成网络 3 个子网络。其中，两个预测网络作为上游任务，是通过音频序列去驱动预处理后的面部肖像，从而生成对应的特征坐标序列。本文方法通过定性实验和定量实验充分展现其有效性和优越性。本文的主要内容及创新点总结如下：

（1）提出了一种音频驱动的情感强度可控面部动画生成技术，在额外添加的情感肖像和情感强度标签的驱动下，生成唇音同步、3 种情感强度可控的面部动画。

（2）核心创新点是面部特征坐标预测网络，通过情感强度标签辅助情感肖像驱动源，控制生成 3 种情感强度不同的面部表情，在增加情感多样性的同时，提升了面部动画的真实性。

（3）在传统级联结构的非端到端算法基础之上，引入解耦的设计思想，将唇部运动的控制和面部表情的生成又分为两个并联的子网络。该方法避免了多个目标间的互相干扰，保证了生成动画的质量。

1. 中国能源建设集团山西省电力勘测设计院有限公司
山西太原 030000

1 相关工作

1.1 基于音频驱动的面部动画生成技术

基于音频驱动的面部动画生成技术利用音频信号（或视频信号）的运动信息，驱动目标肖像的外观信息生成面部动画[1]。随着近几年短视频创造需求[12]的增加和高质量视听数据库[13]的提出，该方向得到研究机构与科研人员的关注，发展也尤为迅速。

早期的研究大多通过基于图像的方法，对目标单一肖像唇部区域进行建模，具有数据需求量大与泛化能力薄弱等问题。随着深度学习的发展，通过音频驱动和生成对抗网络（generative adversarial networks, GAN）的方式成为主流。Chung 等人[14]利用卷积神经网络（convolutional neural network, CNN）学习目标肖像和音频驱动源的联合嵌入，并利用该嵌入生成相应的面部动画。此外，Chen 等人[15]通过引入结构化的面部特征坐标，提出了具有级联结构的非端到端算法，在保证唇音同步的前提下，大幅提升了模型的泛化能力与灵活性。上述方法有一个共同问题，均没有考虑面部动画的表情控制，致使动画的面部表情被输入的目标肖像所局限，真实性有待提高。

1.2 静态肖像面部表情的控制

面部表情是人类最自然、有效、普遍的情绪表达方式之一。在计算机视觉和图像生成领域，对面部表情控制的探索始于静态肖像。例如，Zhang 等人[16]提出基于三重感知损失函数的非端到端算法，通过面部特征坐标控制表情的变化。Liu 等人[17]则是在此基础上，增加面部旋转模块和面部表情增强生成器，实现肖像姿态和表情更细腻度的控制。而随着动态短视频需求的不断增加，上述方法已无法满足任务需求，如何生成情感可控的面部动画，已成为一个备受瞩目的前沿挑战。

1.3 基于音频驱动面部动画生成技术的表情控制

随着情感标签视听数据库的丰富，近几年面部动画生成技术，从唇音同步的探索向面部表情的控制上转变。现有的方法可根据是否有中间表示划分为两类，即具有完整性的端到端算法和级联结构的非端到端算法。

端到端的面部动画生成技术可以从原始输入的肖像与驱动源中学习到与任务相关的特征，且模型不会产生额外的中间表示。具体而言，Fang 等人[7]将提取的音频嵌入直接与目标肖像融合，并在身份与情感两对判别器和分类器的作用下，生成情感可控的面部动画。最近，Liang 等人[8]提出一种细粒度的面部动画生成技术，该技术需输入 3 个驱动源（两个视频和一个肖像），以控制面部动画的唇部运动、面部表情和头部姿态方向。然而，端到端算法网络架构的敏感性较强，对输入要求较高，Fang 等人[7]生成的视频存有模糊与伪影现象。Liang 等人[8]则是肖像和背景处理能力薄弱，不得不用单一纯色场景进行替换。

另一方面，级联结构的非端到端算法通过中间表示连接不同驱动源与目标肖像，生成的面部动画在可解释性与灵活性均更加优秀。Ji 等人[9]和 Agarwal 等人[10]将肖像降维成 10 个一组的面部运动关键点，并利用掩码技术对面部区域进行加强，最后通过生成网络得到面部动画。上述方法均只注重生成面部表情，并未考虑面部表情的强度问题。针对此问题，本文提出具有级联结构的情感强度可控的面部动画生成模型，以通过情感强度标签的映射，增强情感源肖像的驱动能力，生成唇音同步、情感强度可控的面部动画。

2 网络结构

针对现有面部动画生成模型面部表情控制能力薄弱、情感可控性较低的问题，本文提出基于音频驱动的情感强度可控面部动画生成模型。该网络是采用级联和解耦设计思想的非端到端算法，将唇音同步任务和面部表情生成任务分离的同时，通过额外输入的情感肖像和情感强度标签，共同控制动画的面部表情。

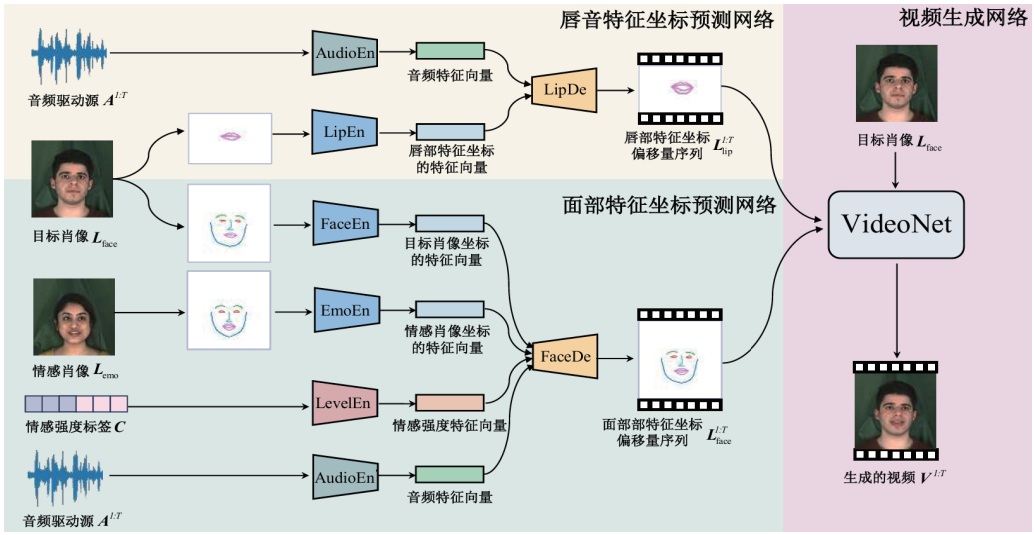

图 1 模型框架图

图 1 是本文模型的框架图，模型输入为两幅面部肖像图片、一段音频和一个情感强度标签，输出是一个唇音对齐且情感强度可控的面部肖像视频。其中，情感强度标签包含两部分，即额外肖像的情感强度和预计生成的情感强度。模型由 3 个子网络模型组成，即唇音特征坐标预测网络、面部特征坐标预测网络和视频生成网络。具体公式分别为：

$$\boldsymbol{L}_{\text{lip}}^{1:T}=\text{LipNet}(\boldsymbol{L}_{\text{lip}},\boldsymbol{A}^{1:T}) \tag{1}$$

$$\boldsymbol{L}_{\text{face}}^{1:T}=\text{FaceNet}(\boldsymbol{L}_{\text{face}},\boldsymbol{L}_{\text{emo}},\boldsymbol{A}^{1:T},\boldsymbol{C}) \tag{2}$$

$$\boldsymbol{L}_{\text{new}}^{1:T}=\boldsymbol{L}_{\text{face}}+\boldsymbol{L}_{\text{lip}}^{1:T}+\boldsymbol{L}_{\text{face}}^{1:T} \tag{3}$$

$$\boldsymbol{V}^{1:T}=\text{VideoNet}(\boldsymbol{L}_{\text{new}}^{1:T},\boldsymbol{I}_{\text{face}}) \tag{4}$$

式中：LipNet 表示唇音特征坐标预测网络；FaceNet 表示面部特征坐标预测网络；VideoNet 表示视频生成网络；$\boldsymbol{L}_{\text{lip}}$ 表示目标肖像的唇部特征坐标；$\boldsymbol{A}^{1:T}$表示音频驱动源；$\boldsymbol{L}_{\text{lip}}^{1:T}$表示预测的唇部特征坐标偏移量序列；$\boldsymbol{L}_{\text{face}}$ 表示目标肖像的面部特征坐标；$\boldsymbol{L}_{\text{emo}}$ 表示额外肖像的面部特征坐标；$\boldsymbol{C}$ 表示情感强度标签；$\boldsymbol{L}_{\text{face}}^{1:T}$表示预测的面部特征坐标偏移量序列；$\boldsymbol{L}_{\text{new}}^{1:T}$表示修正后目标肖像的面部特征坐标序列；$\boldsymbol{I}_{\text{face}}$ 表示目标肖像的原始图片；$\boldsymbol{V}^{1:T}$表示最终生成的视频；T 表示音频的长度。

此外，因 3 个子网络模型都是以面部特征坐标为设计蓝图，本文将使用预训练的面部特征关键点识别算法获取面部特征坐标。再者，本文将使用梅尔倒谱系数（mel frequency cepstral coefficient, MFCC）作为音频的表示，以更好地适应长短时记忆网络（long short-term memory, LSTM），并提升模型整体的性能和稳定性。本文将在以下小节中对所提出的子网络模型进行详细描述。

2.1 唇音特征坐标预测网络

网络 LipNet 是编码器－解码器结构，包含一个唇部特征坐标编码器 LipEn、一个音频编码器 AudioEn 和一个唇部特征坐标解码器 LipDe。网络输入是目标肖像的唇部特征坐标 $\boldsymbol{L}_{\text{lip}}\in 20\times 3$，并使用音频作为驱动源$\boldsymbol{A}^{1:T}$，以预测出与该音频对齐的唇部特征坐标偏移量序列$\boldsymbol{L}_{\text{lip}}^{1:T}\in 20\times 3\times T$。

具体而言，唇部特征坐标编码器 LipEn 是一个四层的多层感知机（multi-layer perceptron, MLP），其通过多层的非线性变换和全连接结构，以学习并捕捉唇部特征坐标的非线性关系和隐含特征表示。该编码器输入维度为 20×3，两个隐藏层输入维度分别为 128 和 256，输出数据维度为 256 的唇部特征坐标的隐含特征向量。此外，输入层和隐藏层后，均会使用 BatchNorm 函数加速模型收敛，使用 LeakyReLU 激活函数引入非线性因素增加模型的表达能力。本文其余 MLP 的输入层和隐藏层同样会使用 BatchNorm 函数和 LeakyReLU 激活函数，后文将不做过多赘述。

音频编码器 AudioEn 是一个 LSTM，其通过门控单元控制信息的流动和内容的更替，以捕捉音频长时间的依赖性关系和长序列隐含特征。该编码器输入维度为 (T, 40)，并经过 3 个维度为 256 的隐藏层，得到维度为 (T, 256) 的音频特征向量。

唇部特征坐标解码器 LipDe 是一个五层的 MLP，其负责捕捉唇部特征坐标隐含特征向量和音频特征向量间的映射关系，以预测出长度为 T 的唇部特征坐标偏移量序列$\boldsymbol{L}_{\text{lip}}^{1:T}$。先逐帧将唇部特征坐标的隐含特征与音频的隐含特征线性融合，得到维度为 (T, 512) 的总隐含特征。之后将其逐帧送入解码器中，再经过三层输入维度为 256、128、64 的隐藏层，并通过输出层得到维度为 (T, 60) 的坐标偏移量$\boldsymbol{L}_{\text{lip}}^{1:T}$。上述内容可用公式表示为：

$$\boldsymbol{L}_{\text{lip}}^{1:T}=\text{LipDe}(\text{LipEn}(\boldsymbol{L}_{\text{lip}}),\text{AudioEn}(\boldsymbol{A}^{1:T})) \tag{5}$$

此外，为了训练该网络的参数，本文使用了两个损失函数，即唇部坐标欧几里得距离损失函数，唇部坐标拉普拉斯距离损失函数。其一是面部动画生成常用的损失函数，以使预测坐标和参考坐标之间的距离最小化。其二是为更好地捕捉预测坐标和参考坐标之间的相似性，以保留更多唇部变化的细节。损失函数具体用公式表示为：

$$\begin{aligned}\text{Loss}_{\text{lip}}=&\sum_{t=1}^{T}\sum_{i=1}^{N}\left\|(\boldsymbol{L}_{\text{lip}}^{t,i}+\boldsymbol{L}_{\text{lip}}^{i})-\overline{\boldsymbol{L}}_{\text{lip}}^{t,i}\right\|_2^2+\\&\lambda_1\sum_{t=1}^{T}\sum_{i=1}^{N}\left\|\mathcal{L}(\boldsymbol{L}_{\text{lip}}^{t,i}+\boldsymbol{L}_{\text{lip}}^{i})-\mathcal{L}(\overline{\boldsymbol{L}}_{\text{lip}}^{t,i})\right\|_2^2\end{aligned} \tag{6}$$

式中：Loss_{lip} 表示唇音特征坐标预测网络的损失函数；t 表示对应帧数；i 表示坐标的索引；T 表示音频的长度；N 表示唇部坐标的总索引数，其值为 60；$\boldsymbol{L}_{\text{lip}}^{t,i}$表示预测的 t 帧唇部特征坐标第 i 个索引的位置偏移量；$\boldsymbol{L}_{\text{lip}}^{i}$表示人的唇部特征坐标第 i 个索引的初始位置；$\overline{\boldsymbol{L}}_{\text{lip}}^{t,i}$表示参考的 t 帧唇部特征坐标第 i 个索引的位置；λ_1 表示唇部坐标拉普拉斯距离损失函数的加权值，这里 λ_1=1；$\mathcal{L}$ 表示拉普拉斯距离。

2.2 面部特征坐标预测网络

该网络输入目标肖像的面部特征坐标$\boldsymbol{L}_{\text{face}}\in 68\times 3$，通过使用音频作为直接驱动源$\boldsymbol{A}^{1:T}$，使用额外肖像的面部特征坐标$\boldsymbol{L}_{\text{emo}}\in 68\times 3$作为情感驱动源，使用情感强度标签 $\boldsymbol{C}$ 控制面部情感强度，以预测出与该音频对齐、具有额外肖像情感且情感强度可控的面部特征坐标偏移量序列$\boldsymbol{L}_{\text{face}}^{1:T}\in 68\times 3\times T$。该网络共包含 4 个编码器和 1 个解码器，编码器分别是面部特征坐标编码器 FaceEn、音频编码器 AudioEn'、情感编码器 EmoEn 和情感强度编码器 LevelEn，解码器是面部特征坐标解码器 FaceDe。

面部特征坐标编码器 FaceEn 和情感编码器 EmoEn 是两个结构相同的六层 MLP，作用均为学习面部特征坐标的隐含特征表示，但 FaceEn 负责提取目标肖像的肖像特征，EmoEn 负责提取额外肖像的面部情感特征。两个编码器输入维度为 204，4 个隐藏层输入维度均为 256，输出数据维度为 256 的身份特征向量或情感特征向量。此外，该网络的音频编码器 AudioEn' 与唇音特征坐标预测网络的音频编码器 AudioEn 相同，均为相同网络架构和时间窗口长度的 LSTM。

面部特征坐标解码器 FaceDe 同为一个五层的 MLP，负责将情感特征向量和情感强度特征向量映射到身份特征向量上，并在音频特征向量的驱动下，生成对应时间步长 T 的面部特征坐标偏移量序列$\boldsymbol{L}_{\text{face}}^{1:T}$。在音频特征向量的基础上，每帧 t 上与身份特征向量、情感特征向量和情感强度特征向量

进行线性融合，得到步长为 T，维度为 1 024 的融合向量。之后将其逐步输入至解码器中，再经过三层输入维度为 512、256、256 的隐藏层，并通过输出层得到维度为 204 的面部坐标偏移量。逐帧拼接后得到长度为 T 的面部偏移量序列 $\boldsymbol{L}_{\text{face}}^{1:T}$。上述内容可用公式表示：

$$\begin{aligned}\boldsymbol{L}_{\text{face}}^{1:T} = &\text{FaceDe}(\text{FaceEn}(\boldsymbol{L}_{\text{face}}),\text{EmoEn}(\boldsymbol{L}_{\text{emo}}),\\ &\text{AudioEn}'(\boldsymbol{A}^{1:T}),\text{LevelEn}(\boldsymbol{C}_1,\boldsymbol{C}_2))\end{aligned} \tag{7}$$

对于该网络，除了捕捉特征坐标的位置外，还需要关注情感表达和情感强度的真实性。为此，采用生成对抗网络的设计思想，添加了判别器，用来判别预测的面部特征坐标的真实性。模型结构为六层的 MLP，输入层为 204，四层隐藏层均为 256，输出层为 1。

与唇音特征坐标预测网络一样，为了训练网络的参数，仍会使用欧几里得距离和拉普拉斯距离作为损失函数，只是这次将特征坐标索引范围由唇部扩大至面部，其余均保持一致。

2.3 视频生成网络

预测的唇部特征坐标偏移量序列 $\boldsymbol{L}_{\text{lip}}^{1:T}$ 和面部特征坐标偏移量序列 $\boldsymbol{L}_{\text{face}}^{1:T}$ 需对目标肖像初始的面部特征坐标 $\boldsymbol{L}_{\text{face}}$ 进行偏移，以此得到唇音对齐、面部情感强度可控的面部特征坐标序列。最终的面部特征坐标需要与目标肖像的原始图片 $\boldsymbol{I}_{\text{face}}$ 需要经过预处理后，一起送入视频生成网络。预处理过程是使用预定义的彩色线条，将序列数据的面部特征坐标按区域连接，并转换为三通道的 RGB 图片。在此基础上与同为三通道的目标肖像进行通道相加，得到长度为 T，维度 (256, 256, 6) 的图片序列，并输入至视频生成网络。

最终得到分辨率为 256 px×256 px 的唇音对齐、面部情感强度可控的目标肖像视频。

3 实验与分析

3.1 实验设置

本文网络模型共包含 3 个子网络，因 3 个网络间数据的传递需要统一，故所用到的音频波形的采样率均为 16 kHz，视频的均以 30 帧 /s 进行切分，图像大小均为 256 px×256 px。

为了更好地获得特征坐标的预测结果，两个预测网络使用了在室内录制的视听数据库 MEAD。该数据库视频较为规范且清晰度较高，具有 60 名配音演员、8 种情感、3 种情感强度、13 个公有语料句式、17 个情感独占语料句式和 7 种头部姿态。两个预测网络使用了正面姿态的所有数据作为子集，并以 60%、20% 和 20% 用于训练、测试和验证。而视频生成网络为了提升视频生成的鲁棒性，选择了在室外采集的视听数据库 VoxCeleb2，本网络共用到的 1 981 个身份源。

本文框架由 PyTorch1.8.0 和 CUDA11.1 实现的，3 个模块训练所用 GPU 为 4 块 Nvidia Tesla V100S，其单块显存为 32 GB，训练时长共约 40 h。

3.2 定性实验

对于面部动画生成技术，定性实验可以更直观地展现模型的特点与优劣。本文定性实验对比对象分别为真实数据和同样级联结构的音频转换和视觉生成网络（audio transformation and visual generation network, ATVGnet）与 MakeItTalk，上述方法只需要输入目标消息和音频驱动源即可，而本文还需额外输入情感肖像驱动源和情感强度标签。

从图 2 的定性实验展示图可以直观地看出，本文模型可以生成 3 种情感强度不同的面部动画。相较于 ATVGnet 与 MakeItTalk 只关注了唇部变化，本文可以通过情感强度标签的映射，增强了情感肖像的偏移能力，提升了视频的多样性，扩宽了实际使用的应用场景。然而，本文方法依赖于情感标签的视听数据库 MEAD，情感强度标签需与情感肖像驱动源的情感等级相符，才可达到最好的生成效果。

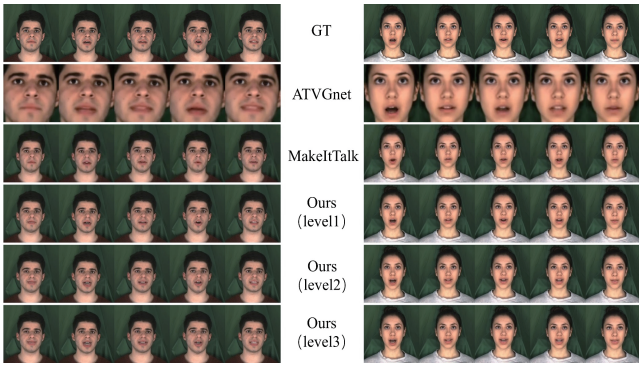

图 2 定性实验展示图

3.3 定量实验

除 ATVGnet 和 MakeItTalk 外，本文定量实验还选取了情感视频肖像模型（emotional video Portraits, EVP）。实验所在视听数据库使用的是 MEAD，评价指标共七项，即用来评估图像质量的峰值信噪比（peak signal-to-noise ratio, PSNR）和结构相似性（structure similarity Index measure, SSIM）；用来评估唇音同步的唇音同步置信度得分 [18]（synchronisation network, SyncNet），唇部特征坐标距离（lip landmarks distances, L-LD）和唇部特征坐标速度距离差异（lip landmark velocity difference, L-LVD）；用来评估面部变化的面部特征坐标距离（face landmarks distances, F-LD）和面部特征坐标速度距离差异（face landmark velocity Difference, F-LVD）。其中，PSNR、SSIM 和 SyncNet 分数越高效果越好，其余则是分数越低效果越好。

本文方法可生成 3 种情感强度不同的面部动画，具体数值如定量实验表 1 所示。可以看出，本文方法（level1）在 7 个指标下均优于对比对象。其中评价指标 F-LD 和 F-LVD 提升最为明显，相较于 EVP 提升了 4.32% 和 5.13%。但随着情感强度的增强，7 个指标均出现了不同程度的下降，这是因为部分面部表情重塑过度所致。

表 1 定量实验

方法	SSIM↑	PSNR↑	SyncNet↑	L-LD↓	L-LVD↓	F-LD↓	F-LVD↓
AVTG	0.60	28.58	2.22	2.57	1.84	3.82	1.71
MakeItTalk	0.69	28.92	2.20	2.80	1.87	3.46	1.68
EVP	0.71	29.53	—	2.45	1.78	3.01	1.56
本文方法（level1）	0.72	30.14	2.23	2.42	1.72	2.88	1.48
本文方法（level2）	0.71	29.82	2.21	2.48	1.79	2.92	1.51
本文方法（level3）	0.69	28.47	2.18	2.63	1.93	3.07	1.57

4 结语

针对现有面部动画生成模型在面部表情强度控制的薄弱，本文提出一个级联结构非端到端算法，即音频驱动的情感强度可控面部动画生成模型。利用额外输入的情感强度标签，通过其映射增强情感源肖像的驱动能力，控制生成三种情感强度不同的唇音同步的面部动画。该方法增加了情感强度的控制能力，提升了视频的多样性，扩宽了实际使用的应用场景。然而，本文方法随着情感强度的增强，部分帧易出现面部表情重塑过度而引发的画面扭曲现象，算法稳定性仍有待提升。

参考文献：

[1] CHEN L L, CUI G F, KOU Z Y, et al. What comprises a good talking-head video generation? a survey and benchmark[EB/OL].(2020-05-07)[2024-05-01].https://doi.org/10.48550/arXiv.2005.03201.

[2] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//2017 IEEE International Conference on Computer Vision (ICCV) .Piscataway:IEEE,2017:2242-2251.

[3] LIU M Y, BREUEL T, KAUTZ J. Unsupervised image-to-image translation networks[C]//NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems.Piscataway:IEEE,2017:700-708.

[4] GARRIDO P, VALGAERTS L, SARMADI H, et al. VDub: modifying face video of actors for plausible visual alignment to a dubbed audio track[J].Computer graphics forum, 2025, 34(2):193-204.

[5] 王涵 , 夏时洪 . 视频驱动的语义表情基动画方法 [J]. 计算机辅助设计与图形学学报 , 2015, 27(5): 873-882.

[6] CAO H W, COOPER D G, KEUTMANN M K,et al. CREMA-D: crowd-sourced emotional multimodal actors dataset [J]. IEEE transactions on affective computing, 2014, 5(4): 377-390.

[7] FANG Z, LIU Z, LIU T T, et al. Facial expression GAN for voice-driven face generation [J]. The visual computer, 2022, 38: 1151–1164.

[8] LIANG B R, PAN Y, GUO Z Z, et al. Expressive talking head generation with granular audio-visual control[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE,2022:3377-3386.

[9] JI X Y, ZHOU H, WANG K S Y, et al. EAMM: one-shot emotional talking face via audio-based emotion-aware motion model[C]// SIGGRAPH '22: ACM SIGGRAPH 2022 Conference Proceedings.NewYork:ACM,2022:1-10.

[10] AGARWAL M, MUKHOPADHYAY R, NAMBOODIRI V, et al. Audio-visual face reenactment[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).Piscataway:IEEE,2023:5167-5176.

[11] SIAROHIN A, LATHUILIÈRE S, TULYAKOV S, et al. First order motion model for image animation[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems.NewYork:ACM,2019:7137-7147.

[12] 宋一飞 , 张炜 , 陈智能 , 等 . 数字说话人视频生成综述 [J]. 计算机辅助设计与图形学学报 ,2023,35(10):1457-1468.

[13] CHUNG J S, NAGRANI A, ZISSERMAN A. Voxceleb2: deep speaker recognition [EB/OL]. (2018-06-27)[2025-01-25].https://doi.org/10.48550/arXiv.1806.05622.

[14] CHUNG J S, JAMALUDIN A, ZISSERMAN A. You said that? [EB/OL]. (2017-07-18)[2024-11-12].https://doi.org/10.48550/arXiv.1705.02966.

[15] CHEN L L, MADDOX R K, DUAN Z Y, et al. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway:IEEE, 2019: 7824-7833.

[16] ZHANG J N, ZENG X F, WANG M M, et al. FreeNet: multi-identity face reenactment[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 5325-5334.

[17] LIU J, CHEN P, LIANG T, et al. Li-Net: large-pose identity-preserving face reenactment network[C]//2021 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE, 2021:1-6.

[18] CHUNG J S, ZISSERMAN A. Out of time: automated lip sync in the wild[C]//Computer Vision-ACCV 2016. Berlin:Springer,2017:251–263.

【作者简介】

阎文博（1982—），女，山西太原人，硕士，高级工程师，研究方向：管理信息化、大数据、数据分析。

（收稿日期：2025-04-09 修回日期：2025-08-05）