具有可控情绪的表现力语音驱动面部动画

Yutong Chen[a]、*Junhong Zhao* [*b] 和 *Wei-Qiang Zhang* [*c]

[a]清华大学，chenyt19@tsinghua.org.cn[b]惠灵顿维多利亚大学，新西兰，junhong.jennifer@gmail.com[c]清华大学，wqzhang@tsinghua.edu.cn。

arXiv:2301.02008v2 [cs.CV] 4 Jan 2024

摘要

生成高度逼真的面部动画需求旺盛，但这仍然是一项艰巨的任务。现有的语音驱动面部动画方法能够生成令人满意的嘴部运动和唇部同步效果，但在表现强烈情绪和灵活控制情绪方面仍有所不足。本文提出一种新颖的深度学习方法，可根据语音生成富有表现力的面部动画，并呈现丰富多样的面部表情，同时支持对情感类型和强度进行控制。我们提出一种情感控制器模块，用于学习情感变化（例如类型和强度）与相应面部表情参数之间的关系。该模块支持情感可控面部动画，用户可根据需要连续调整目标表情。定性和定量评估结果表明，我们的方法生成的动画具有丰富的面部情绪表现力，同时保持了准确的唇部运动，优于其他最先进的方法。

索引词— 3D面部动画，3D虚拟形象，说话人头部，情感，富有表现力的面部动画

1. 引言

面部动画是一个不断发展的研究领域，已广泛应用于教育、虚拟现实（VR）和数字娱乐等领域。商业产品通常要求动画角色具有很高的可信度和表现力，以满足用户对沉浸式体验的需求。创作如此逼真的面部动画是一项巨大挑战。语音驱动面部动画是这一领域的关键组成部分之一，受到了广泛关注。我们的工作聚焦于3D语音驱动面部动画。与制作二维图像动画相比，调控3D模型可以直接应用于大多数3D应用场景，例如3D游戏和视觉后期效果。人们已尝试探索音频与3D面部运动之间的依赖关系，其中大多数工作聚焦于下半脸和唇部运动及其与语音的同步 [1, 2, 3, 4, 5, 6]。尽管这些方法能够生成可信的基本面部动作，但在合成面部表情方面仍远不能令人满意，尤其是在呈现情感方面。

仅根据音轨实现情感动画是一项具有挑战性的任务。尽管同一个人的声音产生与唇部运动之间存在确定的对应关系，但它们与不同情感类别及强度下的表情之间的关系却高度模糊。不同用户表达同一种情感的方式各不相同。这种固有的模糊性使神经网络难以根据音频输入处理情感变化，即使提供长上下文信息也不例外。Karras 等人和 Pham 等人 [7, 8] 尝试从语音中提取情感特征，并将其隐式嵌入神经网络以实现情感合成。然而，由于训练数据有限，他们的方法不可避免地会遭遇过度平滑的回归问题，往往导致表现力不足。

我们提出了一种实现情感可控面部动画的新方法。不同于从音频中还原唇部动作，我们的方法增强了情感表现力，并支持调节情感效果的强度，以满足动画师的需求。我们提出了一个情感控制器模块，其中包含一个情感预测器和一个情感增强网络，用于明确建模情感变化与相应面部表情参数之间的关系。我们使用基于图像的情感识别生成情感信息，作为先验引导训练过程。在推理阶段，指定的情感条件将作用于语音驱动面部动画，以增强并定制情感表现。通过明确地建模情感，我们的动画系统便于动画师使用，可根据给定的情感类型和强度值进行情绪控制。我们的方法在合成可控情感面部动画方面展现出良好效果，同时保持了高精度的唇形同步，

* Junhong Zhao 和 Wei-Qiang Zhang 为共同通讯作者。本研究得到中国国家自然科学基金项目（编号：62276153）的资助。

并优于当前最先进方法。实现代码可在 *https://github.com/on1262/facialanimation*获取

2. 相关工作

尽管许多工作专注于从图像或视频生成面部动画 [9, 10, 11, 12] ，或生成语音驱动的二维说话头像 [13, 14, 15, 16, 17], ，但我们将研究重点放在语音驱动的三维面部动画上，主要致力于改进情绪表达合成。我们回顾了最相关的基于深度学习的方法。

语音驱动的3D模型。早期的语音驱动3D面部动画方法基于音素标注和视素模型混合 [1, 18]。VisemeNet [1] 利用长短期记忆网络进行音素分组和面部标志点预测，并基于预测结果回归视素参数，以实现唇部动画。VOCA [5] 尝试将身份相关信息编码到动画中，并合成多种说话风格。MeshTalk [6] 提出了一种将音频相关信息与音频无关信息解耦的方法，从而在实现精准唇部运动的同时，为上半脸生成更自然的动态效果。FaceFormer [2]提出了一种基于变换器的自回归模型，用于编码长期音频上下文并合成更理想的唇部运动。Taylor 等人 [3]提出了一种滑动窗口回归方法，根据音素标签预测参考下半脸的主动外观模型（AAM）参数。Liu 等人 [19]考虑了几何表示的影响，并利用几何表示生成具有泛化能力的说话人无关面部动画。尽管该领域已取得显著进展，但上述大多数工作都专注于唇部运动的准确性，而未考虑表现力，这限制了生成结果的真实感。

条件情感合成。 近年来，一些面部动画方法 [15, 7, 8, 20] 致力于解决情感合成中的若干问题，以生成更加生动的面部动画。Pham 等人使用长短期记忆网络，在 [8]（并在 [21] 中改进为卷积神经网络-循环神经网络）中建模声学特征与面部表情之间的长程上下文关系，从而使生成的动画具备情感感知能力。Karras 等人提出的方法 [7] 从音频中提取潜在情感表示，而不识别情感类别。这两种方法都以隐式方式表示情感，因此无法为情感强度控制提供有意义的引导。Ji 等人 [15] 将语音分解为情感和内容成分，以生成情感控制的二维说话头。Chun 等人 [20]提出了一种情感引导方法，通过情感识别引导增强情感表达混合形状，然后将其与口部表情混合形状融合。然而，该技术需要人工为每种情感准备模板，且生成的表情往往缺乏上半脸动态。

3. 方法

3.1. 网络架构设计

图 1 展示了我们的处理流程。我们的目标是增强语音驱动面部动画中的情感表现力，并让用户能够控制情感变化。首先，我们设计了一个神经网络，利用局部信息和长程上下文信息，根据输入音频估计由 FLAME 参数表示的面部运动 [22] （见补充材料）。该问题可表述如下:

$$(\vec{\phi_t}, \vec{\theta_t}) = F(S(\vec{x}_{t\pm\Delta t}) \to \vec{\psi_t}; W(\vec{x}_{t\pm\Delta t}) \to \vec{\omega_t}) \quad (1)$$

给定时间 t 的音频片段 x_t 及其相邻帧，首先分别提取 wav2vec2.0$W(\cdot)$ 学习到的局部内容特征（内容向量 $\vec{\omega_t}$）和 Transformer编码器 $S(\cdot)$ 学习到的 $\vec{}$ 全局风格特征（风格向量 ψ_t）。然后将二者拼接起来，以预测 FLAME 参数。音频与 FLAME 参数之间的映射由 Auido2FLAME 模块 $F(\cdot)$ 学习，该模块是一个多层卷积神经网络。预测得到的 FLAME$\vec{}$ 参数，包括表情参数 ϕ_t 和姿态参数 $\vec{}\theta_t$，与给定身份的形状参数相结合，转换为三维网格作为输出。此阶段将重点关注唇形同步，并确保嘴部运动的精确性。

我们引入了一个情感控制模块，其中包括一个基于双向长短期记忆网络的情感预测器，后接一个嵌入层，用于生成情感相关潜在特征（情感特征向量）；还包括一个基于卷积神经网络的情感增强网络，用于基于情感特征增强 FLAME 参数的表现力。情感增强过程可表示为:

$$E(\vec{\psi}_{t\pm\Delta t}, \vec{\omega}_{t\pm\Delta t}, \gamma_{u,t}) : (\vec{\phi_t}, \vec{\theta_t}) \to (\vec{\phi'_t}, \vec{\theta'_t}) \quad (2)$$

其中， $E(\cdot)$表示情感控制模块。利用音频特征和用户定制的情感条件$\gamma_{u,\ t}$，该模块将 Audio2FLAME 模型预测的（$\vec{\phi_t}$, $\vec{\theta_t}$）映射为情感增强的面部参数（ϕ'_t, θ'_t）。我们在处理流程中以残差方式引入情感增强网络，从而能够专门优化情感相关表情，同时保留内容相关表情。通过这种方式，用户可以在帧级明确设计情感强度和类别，并调节动画输出。此外，我们还尝试了其他长短期记忆网络变体，例如门控循环单元（GRU），但发现其性能与双向长短期记忆网络模型相比没有显著差异。

3.2. 情感控制模块

实现情感模拟完全可控的核心挑战在于，使模型不仅能够适应情感状态的变化，

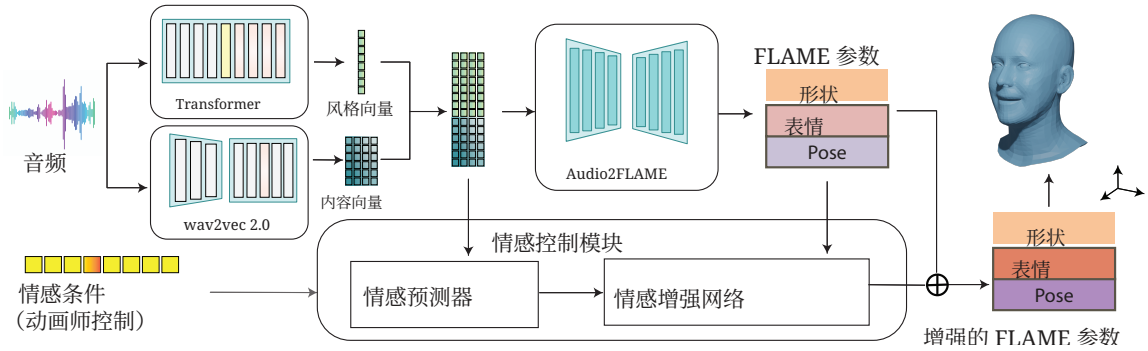


图 1：我们的处理流程概览。

还能够适应情感强度的变化，从而实现直接的强度调节。我们提出的训练与推理流程如图 2 所示。

情感预测与控制。在训练阶段，为了让网络感知输入中的情感变化，我们利用基于图像的情感识别模型获取帧级情感信息，并将其作为情绪先验，以促进模型训练。在我们的实验中使用了 Weng 等人提出的 DAN 模型 [23]，也可以用其他类似模型替代。由于音频中的情感与上下文密切相关，我们认为从二维视觉图像中解码出的情感特征比从音频中提取的情感特征更可靠、信息量更丰富；后者可以从单帧中提取。

然而，DAN 输出的情感分类概率无法表示情感的强弱。相反，我们发现，情感识别网络最终 Softmax 层之前的情绪逻辑值是一个七维向量，对应快乐、愤怒等七种情感，并且与感知到的情感强度高度一致。因此，我们将其用作模型训练的情绪先验，并将其与用户的情感控制相结合。关于情绪逻辑值与情感强度之间线性关系的简化概念性证明，详见补充材料。在实验中，我们发现情绪逻辑值有助于情感合成，并且能够很好地配合情感类别和强度的调整。情感增强网络能够适应 DAN 模块情绪先验的预测误差，并据此生成通用的情绪表达。

在推理阶段，通过双向长短期记忆网络从音频中提取情绪先验，并根据定制的情感类别和强度进行调整。双向长短期记忆网络通过最大化基于音频和基于视频的情感先验之间的互信息进行训练，并将基于视频的情感先验作为伪真值。用户给出的情感条件将转换为取值范围为 0 到 1 的独热向量（$\gamma_{u,\ t}$），并添加到去中心化的基于音频的情感先验（片段级均值归一化（γ_a））中，从而得到最终的情绪先验 $\gamma_t = \gamma_{u,\ t} + (\gamma_{a,\ t} - \gamma_a)$。

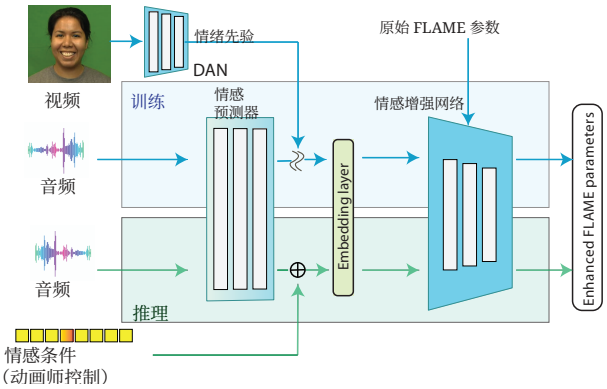


图 2：我们提出的情感控制模块的训练/推理处理流程。

情感增强网络。输入情感增强网络之前，情绪先验会通过嵌入层转换为情感特征向量，以驱动三维面部表情增强。嵌入层是一个可学习的二维矩阵，会在训练过程中进行优化。将 7 维先验向量与学习得到的嵌入矩阵相乘，可将其转换为大小为 128 的情感特征向量，然后输入由卷积神经网络模块构成的情感增强网络，实现情感引导的面部表情增强。为了在唇部同步和情感表现力之间取得适当平衡，我们以残差方式添加了情感增强网络（见公式 3）。从 Audio2FLAME 提取的原始面部参数预计具有较好的唇部同步效果。原始面部参数与真实值（$\vec{\phi}_{gt,\ t}$；$\vec{\theta}_{gt,\ t}$）之间的差异主要由二者不同的情绪表达造成，因此更适合通过增强网络 $A(\cdot)$进行学习。

$$A(\vec{\phi}_t;\vec{\theta}_t) + (\vec{\phi}_t;\vec{\theta}_t) \cong (\vec{\phi}_{gt,t};\vec{\theta}_{gt,t}) \tag{3}$$

3.3. 损失设计

我们使用以下损失函数训练模型：

$$L = w_1 L_{vx} + w_2 L_{lm} \tag{4}$$

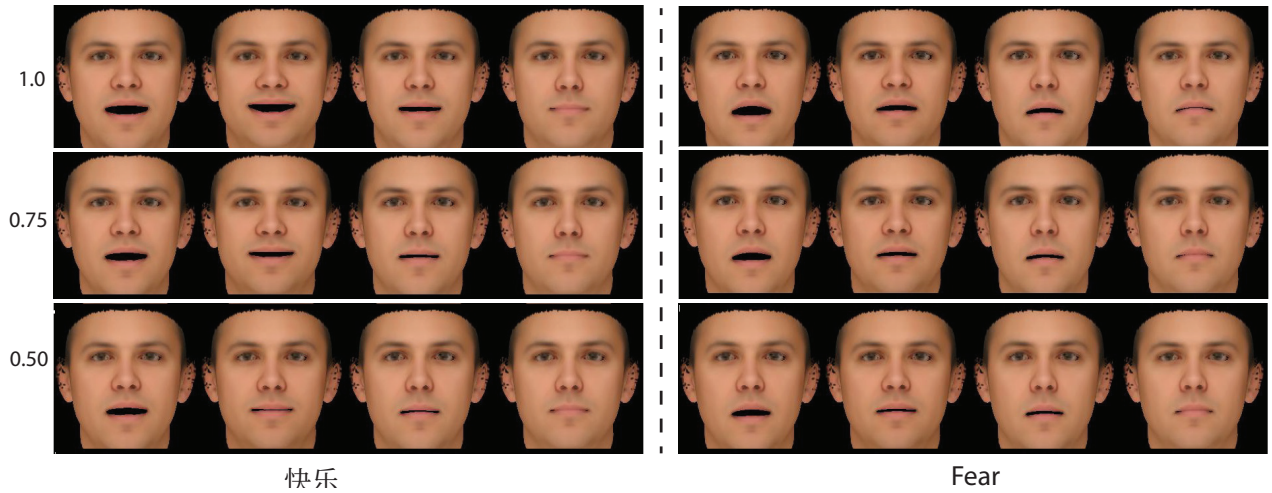


图 3：同一段语音在不同情感类别和自定义强度下的情绪动画结果。为便于可视化，纹理取自真实视频。左：快乐的动画帧；右：恐惧的动画帧。从上到下，强度分别为 1.0、0.75 和 0.5。

其中， L_{vx} 是顶点位置损失， L_{lm} 是嘴型损失， w 是权重。

顶点位置损失： 根据预测的 FLAME 参数生成三维网格，并计算转换后的顶点与真实顶点之间的 L1 差异，记为 L_{vx}。对网格应用顶点掩码，以覆盖前脸区域并排除耳朵和眼睛。

嘴型损失： 为确保唇部同步，我们选取最上方（v_t）、最下方（v_b）、最左侧（v_l）和最右侧（v_r）的顶点位置，并计算嘴部高度 $V = |v_t - v_b|$ 和嘴部宽度 $H = |v_l - v_r|$，将其作为形状描述。真实值与预测嘴型的高度和宽度之间的 L1 距离用作嘴型损失。

$$L_{lm} = d_1 * ||H_p - H_g||_1 + d_2 * ||V_p - V_g||_1. \quad (5)$$

我们在实验中设置了 $d_1 = 1/0.0476$ 和 $d_2 = 1/0.017$。

4. 实验

数据设置： 带有丰富情绪多样性的人工标注三维数据集十分有限。我们的模型使用包含情感丰富的语音、视频或三维模型配对数据的三维和二维数据集混合进行训练，包括 VOCASET [5](三维) 和 CREMAD[24](二维)。对于二维 CREMAD 数据集，我们使用 EMOCA [25] 方法从二维图像重建三维人脸模型（参见补充材料中的 EMOCA 结果），并提取 FLAME 参数作为真实值。

4.1. 情绪动画控制

以用户控制为条件来实现三维情绪面部动画，是我们方法的主要贡献之一。图— 3 展示了可控面部动画结果，分别呈现强度为 0.5、0.75 和 1.0 的快乐与恐惧情绪。所有示例均由相同的音频输入驱动。结果表明，在相同音频输入的驱动下，我们的模型能够生成多样化的情绪面部动画效果，体现用户对情感的定制。生成的面部表情同时受到用户调整和音频输入的影响。

具体而言，在相同的音频输入下，恐惧情感会使嘴部上扬，而快乐情感则会使嘴部下垂。情感幅度的连续变化也能带来令人满意的表情变化。例如，在恐惧情感动画中，随着情感强度从 0.5 增至 1.0，嘴巴张得更大，嘴角也向下移动。快乐情感也有类似变化：强度为 1.0 时，嘴角和脸颊上扬幅度更大；强度降至 0.5 时，则会变为中性闭嘴状态。借助我们可控情感的方法，用户无需准备任何其他依赖项，只需指定所需的强度值，即可在任意关键帧编辑动画；与 [20]等传统方法相比，这种方式更加直接高效。

4.2. 对比

我们将本方法与最先进的语音驱动面部动画方法进行比较，以展示本方法在情感合成和唇部同步方面的表现。我们选取 Pham 等人 [21] 和 Chun 等人 [20] 的研究进行情感合成对比，并选取 VOCA [5]和 FaceFormer [2] 进行唇部同步对比（分别记为 [Pham17], [Chun21], [VOCA19]和 [Face- Former22] ）。

情感合成。 图 4 展示了一个愤怒情绪下的面部动画示例。我们观察到， [VOCA19]和 [FaceFormer22]的唇部运动很准确，但表情中性——

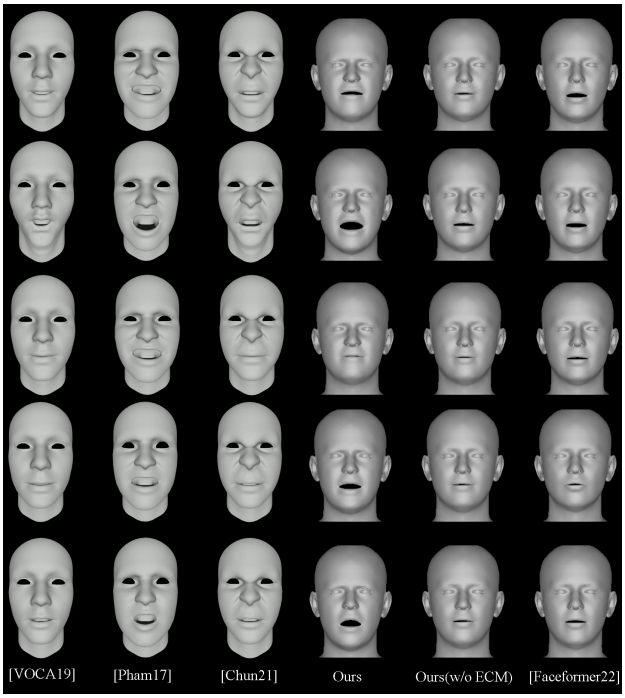


图 4：愤怒动画的定性最先进方法对比。句子“狗正坐在门边”并未出现在我们的训练数据集中。我们选取了该句子开头（上）、中间（中）和结尾（下）阶段的不同动画帧。**更多可视化结果请参见提交的视频**。

——而我们的方法能够在不牺牲唇形同步和口语内容可理解性的前提下，额外加入一层情感变化。在我们的结果中，眉毛下垂，出现脸颊收缩，且嘴部张合程度会随情绪高峰与低谷而变化。 [Pham17] 能够呈现一些情绪表达，但其唇部运动不如我们准确，上半脸的动态也相当有限（几乎静止）。 [Chun21]提供了精准的唇部同步情绪表达，但由于情感模板是手动设定并统一应用于所有帧，生成的动画在说话过程中缺乏情绪动态，尤其是上半脸区域（参见起始帧和结束帧）。相比之下，我们的方法在动画过程中呈现出更丰富的情绪起伏，看起来像真人的表情一样自然。此外，值得注意的是，我们的方法能够将情感从一种身份迁移到另一种身份，这得益于 FLAME 方法对形状与表情的解耦，而 [Chun21]和 [Pham17] 的方法并未实现这一点。情绪一致性的定量评估见补充材料。

唇部同步。对于语音驱动的动画而言，准确的唇部运动对于传达内容信息至关重要。为了公平比较唇形同步效果，我们使用 Audio2FLAME 输出的 FLAME 参数，与其他已有工作进行定量比较。测试数据与 FaceFormer [2],所用数据相同，其中包含 VOCASET 中两位受试者的数据，每位受试者各有 20 个句子样本。

为避免不同方法输出的网格之间引入额外的对齐误差，我们在计算指标之前进行线性三维变换，使它们具有可比性。对齐后，我们在唇部周围均匀选取了 24 个关键点顶点，并计算其与真实值之间的距离，作为唇部同步的度量。序列 i 第 j 帧中关键点顶点 k 的唇部运动误差计算如下：

$$D_{i,j,k} = \sqrt{(x_k - \hat{x_k}) + (y_k - \hat{y_k}) + (z_k - \hat{z_k})} \qquad (6)$$

整体平均距离和最大距离被用作两项指标。

表 1 显示，与其他所有方法相比，我们的方法取得了最佳结果，表明所提出的处理流程能够生成令人满意的唇形同步，从而增强情感表现。正如预期，不经训练步骤、通过随机初始化创建的 RandInit 模型准确率最低。该模型可作为基准，用于展示不同训练方法带来的提升。

表 1：**唇部运动误差（mm）对比**。

	RandInit	[VOCA19]	[FaceFormer22]	Ours
均值↓	2.31	1.94	1.97	1.92
Max↓	3.87	3.41	3.33	3.24

4.3. 消融研究

我们对损失项、网络组件和训练数据集进行消融，以考察它们各自的贡献。表2总结了所提出的模型在不使用顶点位置损失、不使用嘴型损失，以及不使用从Transformer编码器提取的风格向量时所产生的影响。我们的发现表明， L_{vx} 和 L_{lm} 损失都是模型整体损失函数不可或缺的组成部分，省略任一项都会导致性能下降。移除其中任何一项都会使误差指标上升。风格向量还有助于网络从音频输入中捕捉长上下文特征，并改善唇部同步。

我们还尝试了不同的训练数据集，以探究适当训练数据的益处。仅使用3D VOCASET数据集会降低唇部运动的准确性。我们还考察了LRS2 [26] 数据集。我们选取了BBC电视节目中的3000个人像视频片段用于训练，这些片段包含不同的主体、背景噪声和环境。尽管结果显示，在实验数据集上唇部运动准确性没有提升，但我们观察到，该数据集提高了模型在真实场景测试中的泛化能力。

我们还对比了有无情感控制模块的网络结构，进行了定性分析。图 4

表2：通过唇部运动误差衡量的消融研究。

	平均值（毫米）↓	最大值（毫米）↓
Ours	**1.92**	**3.24**
不含 L_{vx} loss	2.01	3.39
不含 L_{lm} 损失	2.02	3.34
不含风格向量	2.03	3.38
仅使用 VOCASET	1.97	3.30
使用 LRS2	1.99	3.34

展示了基于 Auido2FLAME 的 FLAME 参数（记为 Ours（不含 ECM））以及情感控制模块（记为 Ours）生成的结果。可以看出，与不使用该模块生成的中性动画相比，情感控制模块能够提升动画的表现力，实现更强烈的情感表达。

5. 结论与未来工作

我们提出了一种基于深度学习的新方法，用于生成可控的语音驱动情感面部动画。我们设计了一个情感控制器模块，以增强情感表达能力，并支持动画师自定义情绪强度和类别。我们采用基于图像的情感识别来生成情绪先验，从而促进显式情感学习。未来的工作可探索如何突破极端情感生成的边界，同时实现精准的唇形同步；还可通过提升基于视频的情感识别的时间性能，改进动画生成。

6. 参考文献

[1] Yang Zhou、Zhan Xu、Chris Landreth、Evangelos Kalogerakis、Subhransu Maji 和 Karan Singh，《Visemenet：由音频驱动、以动画师为中心的语音动画》，*ACM*图形学汇刊（*TOG*），第 37 卷，第 4 期，第 1–10 页，2018 年。[2] Yingruo Fan、Zhaojiang Lin、Jun Saito、Wenping Wang 和 Taku Komura，《Faceformer：基于 Transformer 的语音驱动三维面部动画》，载于 计算机视觉与模式识别会议，第 18770–18780 页。[3] Sarah Taylor、Taehwan Kim、Yisong Yue、Moshe Mahler、James Krahe、Anastasio Garcia Rodriguez、Jessica Hodgins 和 Iain Matthews，《一种通用语音动画的深度学习方法》，*ACM*图形学汇刊（*TOG*），第 36 卷，第 4 期，第 1–11 页，2017 年。[4] Guanzhong Tian、Yi Yuan 和 Yong Liu，《Audio2face：利用基于注意力的双向 LSTM 网络，从单段音频生成语音/面部动画》，载于*2019* 年 *IEEE* 多媒体与博览会研讨会（*ICMEW*）。IEEE，2019 年，第 366–371 页。[5] Daniel Cudeiro、Timo Bolkart、Cassidy Laidlaw、Anurag Ranjan 和 Michael J Black，《三维说话风格的捕捉、学习与合成》，载于计算机视觉与模式识别会议，2019 年，第 10101–10111 页。[6] Alexander Richard、Michael Zollhöfer、Yandong Wen、Fernando De la Torre 和 Yaser Sheikh，《MeshTalk：通过跨模态解耦实现由语音驱动的三维面部动画》，载于国际计算机视觉大会，2021 年，第 1173–1182 页。[7] Tero Karras、Timo Aila、Samuli Laine、Antti Herva 和 Jaakko Lehtinen，《通过对姿态和情感进行联合端到端学习实现音频驱动的面部动画》，*ACM*图形学汇刊（*TOG*），第 36 卷，第 4 期，第 1–12 页，2017 年。[8] Hai X Pham、Samuel Cheung 和 Vladimir Pavlovic，《具有隐式情感感知能力的语音驱动三维面部动画：一种深度学习方法》，载于*IEEE*计算机视觉与模式识别会议论文集（研讨会），2017 年，第 80–88 页。

[9] Hyeongwoo Kim、Pablo Garrido、Ayush Tewari、Weipeng Xu、Justus Thies、Matthias Niessner、Patrick P´erez、Christian Richardt、Michael Zollh¨ofer 和 Christian Theobalt，“深度视频肖像”，*ACM*图形学汇刊 *(TOG)*，第 37 卷，第 4 期，第 1–14 页，2018 年。[10] Chen Cao、Hongzhi Wu、Yanlin Weng、Tianjia Shao 和 Kun Zhou，“基于图像的动态头像实时面部动画”，*ACM*图形学汇刊，第 35 卷，第 4 期，2016 年。[11] Shih-En Wei、Jason Saragih、Tomas Simon、Adam W Harley、Stephen Lombardi、Michal Perdoch、Alexander Hypes、Dawei Wang、Hernan Badino 和 Yaser Sheikh，“通过多视图图像转换实现虚拟现实面部动画”，*ACM*图形学汇刊 *(TOG)*，第 38 卷，第 4 期，第 1–16 页，2019 年。[12] Justus Thies、Michael Zollhofer、Marc Stamminger、Christian Theobalt 和 Matthias Nießner，“Face2face：RGB视频的实时面部捕捉与重演”，载于*IEEE*计算机视觉与模式识别会议论文集，2016 年，第 2387–2395 页。[13] Xin Wen、Miao Wang、Christian Richardt、Ze-Yin Chen 和 Shi-Min Hu，“逼真的音频驱动视频肖像”，*IEEE*可视化与计算机图形学汇刊，第 26 卷，第 12 期，第 3457–3466 页，2020 年。[14] Justus Thies、Mohamed Elgharib、Ayush Tewari、Christian Theobalt 和 Matthias Nießner，“神经语音木偶：音频驱动的面部重演”，载于欧洲计算机视觉会议。施普林格，2020 年，第 716–731 页。[15] Xinya Ji、Hang Zhou、Kaisiyuan Wang、Wayne Wu、Chen Change Loy、Xun Cao 和 Feng Xu，“音频驱动的情感视频肖像”，载于计算机视觉与模式识别会议，2021 年，第 14080–14089 页。[16] Ran Yi、Zipeng Ye、Juyong Zhang、Hujun Bao 和 Yong-Jin Liu，“基于学习的个性化头部姿态音频驱动说话人脸视频生成”，*arXiv*预印本 *arXiv:2002.10137*，2020 年。[17] Supasorn Suwajanakorn、Steven M Seitz 和 Ira Kemelmacher-Shlizerman，“合成奥巴马：从音频学习唇形同步”，*ACM*图形学汇刊 *(ToG)*，第 36 卷，第 4 期，第 1–13 页，2017 年。[18] Pif Edwards、Chris Landreth、Eugene Fiume 和 Karan Singh，“J ali：面向动画师的富有表现力的唇形同步视素模型”，*ACM*图形学汇刊 *(TOG)*，第 35 卷，第 4 期，第 1–11 页，2016 年。[19] Jingying Liu、Binyuan Hui、Kun Li、Yunke Liu、Yu-Kun Lai、Yuxiang Zhang、Yebin Liu 和 Jingyu Yang，“几何引导的密集透视网络，用于语音驱动面部动画”，*IEEE*可视化与计算机图形学汇刊，2021 年。[20] Sewhan Chun、Daegeun Choe、Shindong Kang、Shounan An、Young-bak Jo 和 Insoo Oh，“情感引导的语音驱动面部动画”，载于亚洲计算机图形学国际会议 *2021* 海报，第 1–2 页，2021 年。[21] Hai X Pham、Yuting Wang 和 Vladimir Pavlovic，“从语音原始波形端到端学习 3D面部动画”，*arXiv*预印本 *arXiv:1710.00920*，2017 年。[22] Tianye Li、Timo Bolkart、Michael J Black、Hao Li 和 Javier Romero，“从 4D扫描中学习面部形状和表情模型”，*ACM*图形学汇刊，第 36 卷，第 6 期，第 194–1 页，2017 年。[23] Zhengyao Wen 、 Wenzhong Lin 、 Tao Wang 和 Ge Xu，“分散你的注意力：用于面部表情识别的多头交叉注意力网络”，*arXiv*预印本 *arXiv:2109.07270*，2021 年。[24] Houwei Cao、David G Cooper、Michael K Keutmann、Ruben C Gur、Ani Nenkova 和 Ragini Verma，“CREMA-D：众包情感多模态演员数据集”，*IEEE*情感计算汇刊，第 5 卷，第 4 期，第 377–390 页，2014 年。[25] Radek Danˇeˇcek、Michael J Black 和 Timo Bolkart，“EMOCA：情感驱动的单目人脸捕捉与动画”，载于计算机视觉与模式识别会议，2022 年，第 20311–20322 页。[26] Triantafyllos Afouras、Joon Son Chung、Andrew Senior、Oriol Vinyals 和 Andrew Zisserman，“深度视听语音识别”，*IEEE*模式分析与机器智能汇刊，2018 年。

7. 补充材料

7.1. FLAME

FLAME 是一种可微分的统计三维头部模型，能够将表情与形状解耦。该模型包含$\vec{}$ $\vec{}$身份形状参数 β、姿态参数 θ和面部表$\vec{}$情参数 ϕ。设定所有参数后，FLAME 会生成一个由一组顶点（$n_v = 5023$）和面片（$n_v = 9976$）组成的网格。在 FLAME 中，通过学习同一人的中性面部与有表情面部之间的差异，将形状与表情解耦。一些先进的人脸重建方法，包括 DECA 及其后续模型 EMOCA，使用 FLAME 作为人脸模型，并从单目图像中恢复细节丰富且富有表现力的三维面部表情。我们的工作也使用 FLAME 作为 3D面部模型。

7.2. 情绪逻辑值作为训练中的情绪先验

在这项工作中，我们使用人脸识别网络中 softmax 之前一层生成的情绪逻辑值，作为模型训练的情绪先验。这里，我们给出一个简化的概念性证明，说明情绪逻辑值与不同情绪状态的感知情绪强度之差呈线性关系。以二分类（情绪状态 A 和 B）为例，我们将 a_k 和 b_k 定义为样本 S_k 的情感 A 和 B 的强度，而 $\eta^{a,}_{k}{}^{b} = a_k - b_k$ 表示二者之差。样本 S_k 属于情绪 A 的可能性为 $P(S_k = A|\eta^{a,}_{k}{}^{b})$。我们假设所有类别均匀分布且连续可微。那么，各类别的先验概率相同，即 $P(S_k = A) = P(S_k = B)$。因此，根据贝叶斯定理，我们得到

$$P(S_k=A|\eta_k^{a,b})=\frac{P(\eta_k^{a,b}|S_k=A)}{P(\eta_k^{a,b}|S_k=A)+P(\eta_k^{a,b}|S_k=B)}. \quad (7)$$

我们定义给定 $S_k = A$时（a_k， b_k）的联合分布为 $f_{a\hat{b}}$（η），给定 $S_k = B$时（a_k， b_k）的联合分布为 $f_{\hat{a}b}$（η）。并假设它们均服从正态分布 N（μ， σ^2），则

$$P(\eta_k^{a,b}|S_k=A)=f_{a\hat{b}}(\eta_k)=\int_{C_1} f_{a\hat{b}}(a_k,b_k)ds \quad (8)$$

其中 C_1 是 $a_k - b_k = \eta_k$。在假设分布对称的情况下，我们得到：

$$P(\eta_k^{a,b}|S_k=B)=\int_{C_1} f_{\hat{a}b}(a_k,b_k)ds=\int_{C_2} f_{a\hat{b}}(a_k,b_k)ds \quad (9)$$

其中 C_2 是 $b_k - a_k = \eta_k$，可转换为 $a_k - b_k = -\eta_k$。因此：

$$P(\eta_k^{a,b}|S_k=B)=f_{a\hat{b}}(-\eta_k). \quad (10)$$

考虑到情感识别网络中基于情绪逻辑值 Z_{ak} 和 Z_{bk} 的 softmax 投影为 $P(S_k = A|\eta^{a,}_{k}{}^{b}) = \frac{exp(Z_{ak})}{\sum_j exp(Z_{bk})}$，情绪逻辑值的偏差为：

$$\Delta_{z,k}=ln(\frac{P(S_k=A|\eta_k^{a,b})}{P(S_k=B|\eta_k^{a,b})})=\ln(\frac{P(\eta_k^{a,b}|S_k=A)}{P(\eta_k^{a,b}|S_k=B)}). \quad (11)$$

根据公式 8、9 和 10，我们推导得出：

$$\Delta_{z,k}=\ln(\frac{f_{a\hat{b}}(\eta_k)}{f_{a\hat{b}}(-\eta_k)})=\frac{(\eta_k+\mu)^2-(\eta_k-\mu)^2}{2\sigma^2}=\frac{2\mu}{\sigma^2}\eta_k \quad (12)$$

这表明，情绪逻辑值$\Delta_{z,\ k}$ 的偏差与感知情感强度 η_k的偏差呈线性关系。类似地，推导过程也可推广至多类别分类任务。

7.3. 实现细节

我们使用 Adam优化器端到端训练模型，每十个训练轮次将学习率从 $1e^{-4}$ 衰减至 $1e^{-5}$ 。为了在处理流程中有效地对情感进行建模，我们在帧级提取情绪先验，以捕捉情感随时间变化的动态。然而，我们的基于图像的预测方法可能会导致动画序列中出现闪烁。为了解决这一问题，我们采用标签平滑过程，对相邻帧的情绪先验进行低通滤波，以减轻时间不一致性。我们还对预测的 FLAME 参数进行优化，以确保动画人脸的双侧对称性。

音频被转换为 16000Hz，并使用窗口长度为 100ms、步长为 33ms 的滑动窗口切分为多个片段。每个音频片段及其左右相邻片段（$\Delta_t = 1$）均作为网络的音频输入。视频被降采样至 30 fps，并进行人像裁剪，以适应三维人脸模型重建。Wav2vec2 和 EMOCA 使用原论文提供的模型提取音频特征并重建三维人脸模型。在训练期间，两者均保持冻结。

在提取全局风格向量时，我们使用一个 Transformer 编码器，其中包含四个 Transformer 编码器块，每个块由一个自注意力层和一个加法与归一化层组成。最终注意力层最后一个标记输出的 64 维向量用作风格向量。Audio2FLAME 是一个三层卷积神经网络，其输入通道、输出通道、卷积核大小、步长和各层激活函数的配置分别为 (256, 128, 1, 1, ReLU)、(128, 128, 1, 3, ReLU) 和 (128, 56, 1, 1)，其中卷积核大小为一维。

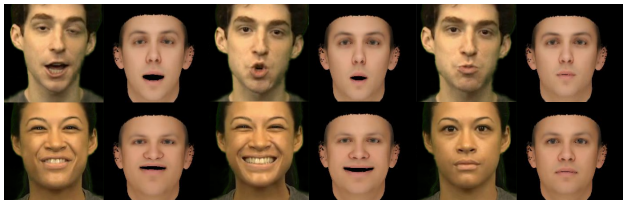


图5：EMOCA 面部几何重建的结果。

情感预测器由一个隐藏维度为 128 的双向 LSTM 层构成。预测得到的情感先验将使用最大最小归一化重新缩放，所用最小值和最大值取自对训练视频数据集中提取的情绪逻辑值进行统计所得的结果。情感增强网络与 Audio2FLAME 结构相同，但在第一个卷积层之前应用了层归一化。将 128 维情感特征向量与 Audio2FLAME 输出的 56 维原始 FLAME 参数拼接后作为输入（总计 184 维），网络将输出 56 维增强的 FLAME 参数。

在训练阶段，PyTorch 和 NumPy 随机化所用的种子设为 1000。更多详情请参阅我们发布的代码。

7.4. 情感合成详情

我们使用 DAN 情绪识别模型评估采用我们的方法生成的表情所呈现的情感。我们选用了包含 1487 个片段的 CREMAD 测试数据集。我们模型的输出会渲染为二维图像，然后使用 DAN 预测这些图像中的情感类别。我们从每段测试视频中随机选取十帧。每段视频的最终预测结果，是所选十帧中各情感类别预测结果的众数。混淆矩阵见下表3。我们的评估表明，识别模型正确识别出了生成图像中呈现的大多数情感类别，这说明我们的结果具有情感区分能力。我们还观察到，结果中的某些类别存在较高程度的模糊性，例如悲伤与中性、恐惧与愤怒之间。

情感	NEU	HAP	ANG	SAD	DIS	FEA
NEU	**188/217**	0/217	0/217	29/217	0/217	0/217
HAP	1/254	**252/254**	0/254	0/254	0/254	0/254
ANG	4/254	0/254	**166/254**	2/254	17/254	**65/254**
SAD	**66/254**	0/254	12/254	**176/254**	0/254	0/254
DIS	5/254	0/254	12/254	3/254	**194/254**	40/254
FEA	4/254	0/254	0/254	13/254	0/254	**237/254**

表3：DAN 对我们生成的动画进行识别的混淆矩阵。

7.5. EMOCA 结果

图5展示了 EMOCA 面部几何重建的结果。