记忆增强对比学习用于说话头生成

Jianrong Wang[1] *Yaxin Zhao*[2] *Li Liu*[3⋆] *Hongkai Fan*[1] *Tianyi Xu*[1] *Qi Li*[4] *Sen Li*[1]

[1]天津大学智能与计算学院，中国天津[2]天津大学天津国际工程师学院，中国天津[3]香港科技大学（广州），中国广州[4]天津大学电气与信息工程学院，中国天津

arXiv:2302.13469v1 [cs.MM] 27 Feb 2023

摘要

给定一张参考面部图像和一段语音作为输入，说话人头部生成旨在合成逼真的说话头视频。然而，生成嘴唇同步且头部运动自然的视频颇具挑战。同一段语音片段可能产生多种不同的嘴唇和头部运动，也就是说，它们之间不存在一一映射关系。为解决这一问题，我们提出了一种基于记忆增强自监督对比学习的语音特征提取器（SFE），通过引入记忆模块来存储多种不同的语音映射结果。此外，我们将混合密度网络（MDN）引入关键点回归任务，以生成多个预测面部关键点。大量定性和定量实验表明，我们的面部动画质量显著优于最先进水平（SOTA）。代码已发布于 https://github.com/Yaxinzhao97/MACL.git。

索引词— 说话人头部生成、对比学习、记忆库、混合密度网络

1. 引言

说话人头部生成对于电影制作 [1], 音视频语音生成 [2] [3], 电脑游戏 [4], 等领域至关重要。头部姿态在增强人们对生成视频真实性的感知方面发挥着重要作用 [5]。

据报道，文献 [6] [7] [8] 直接将语音映射为说话人头部视频。 [9] [10] 首次建立了从语音到高层表示（即面部关键点）的映射，然后根据这些关键点生成说话人头部视频。与人脸像素图像相比，面部关键点相对稀疏，因此我们可以更加专注于捕捉说话者的头部运动。

目前，监督学习是说话人头部生成中语音特征提取的主流方法。监督学习方法需要投入大量精力标注数据，而且无法保证数据标签的准确性，因此语音特征提取容易受到噪声数据的影响。为此，我们提出使用自监督[11]方法提取语音特征。随后，将学习到的语音特征应用于下游任务进行微调，从而获得与下游任务更加一致的语音特征。

另一个挑战是，语音到嘴唇和头部运动的映射并非一一对应。例如，说“嗨”时，不同的人张嘴程度可能有所不同。同样，头部运动也可能不同。我们提出分两个阶段处理这种不确定性。在语音特征提取阶段，我们将记忆模块引入自监督对比学习[12]。语音映射为嘴唇和头部运动时产生的多个结果会被分配到记忆模块中，使特征提取器能够专注于语音特征提取。对于不确定性问题，使用单一模型总会导致次优预测结果 [13]。因此，在面部关键点回归阶段，我们引入混合密度网络来生成多个预测关键点。

我们的工作贡献可概括如下：

1. 我们提出了一种基于记忆增强自监督对比学习的语音特征提取网络，用于提取语音特征。2. 我们将混合密度网络引入面部关键点回归任务，以生成多个预测的面部关键点，从而提升头部运动生成的自然度。3. 多说话者场景下的实验结果表明，在关键点生成和韵律性头部运动方面，我们的模型优于最先进的方法。

2. 方法

如图 1 所示，我们的模型由三个关键组件构成。第一个是语音特征提取器 (SFE)，用于从语音信号中提取内容和身份信息。第二个是混合密度网络 (MDN)，它以语音特征和参考人脸的关键点特征为输入，预测多种可能的面部关键点，并最终选取预测结果最佳的一种作为输出。最后一部分是图像到图像 网络，它

⋆通讯作者：avrillliu@hkust-gz.edu.cn

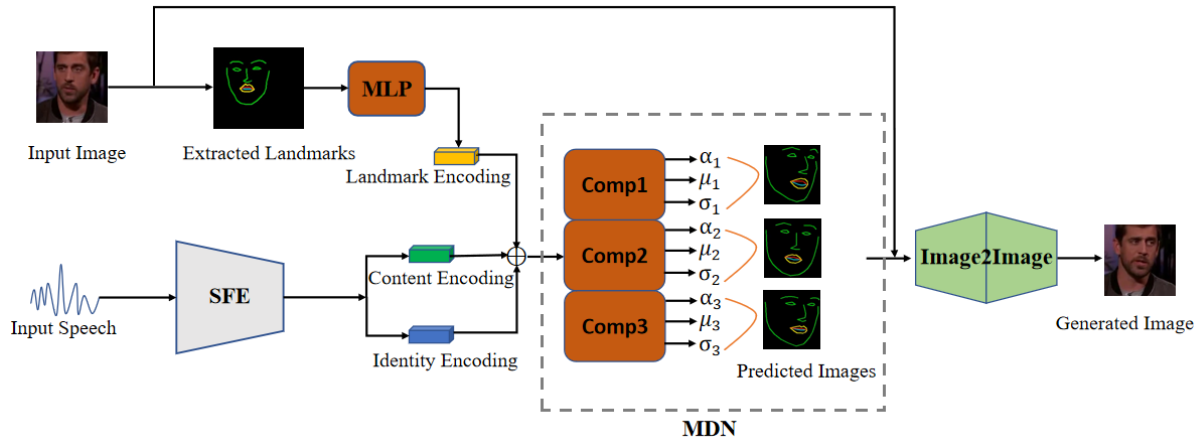


图 1。所提出模型的概览。首先，我们通过语音特征提取器从语音信号中提取内容特征和身份特征。然后，将提取的特征和参考人脸关键点特征输入混合密度网络，以预测面部关键点。每个组件都是一个简单的多层感知机。最后，通过图像到图像 网络将面部关键点转换为逼真的人脸图像。

将面部关键点转换为逼真的人脸图像。我们将在以下小节中介绍算法的各个模块。

2.1. 基于记忆增强对比学习的语音特征提取器

如图 2 所示，所提出的语音特征提取器主要包含三个部分：音频编码器、图像编码器和记忆模块。

记忆模块。为了解决语音到嘴唇和头部运动的一对多映射问题，我们将记忆模块 [14]引入自监督对比学习中，以存储多个映射结果。

具体而言，内容记忆模块记为 $M^c = [m_1^c, m_2^c, ..., m_k^c]^T \in R^{k \times C}$，其中 k 表示记忆模块的大小，C 表示每个记忆槽的维度。每个记忆槽的权重由内容概率分布函数 ϕ_c计算得到，随后计算预测的语音内容特征 $\hat{y}_t^c$。公式如下：

$$p_t^c = Softmax(\phi_c(h_t)), \tag{1}$$

$$\hat{y}_t^c = \sum_{i=1}^{k} p_{i,t}^c \cdot m_i^c = p_t^c M^c, \tag{2}$$

其中，$p_{i,t}^c \in R^k$ 表示第 i个内容记忆槽对时间步t的内容特征表示的贡献。h_t 表示 Bi-GRU 在时间步 t提取的语音特征。

语音身份特征 $\hat{y}_t^s$和内容特征的计算方法相似，但 h_t 通过身份权重分布函数 ϕ_s获得身份权重向量，然后与身份记忆模块 M^s进行乘积运算。

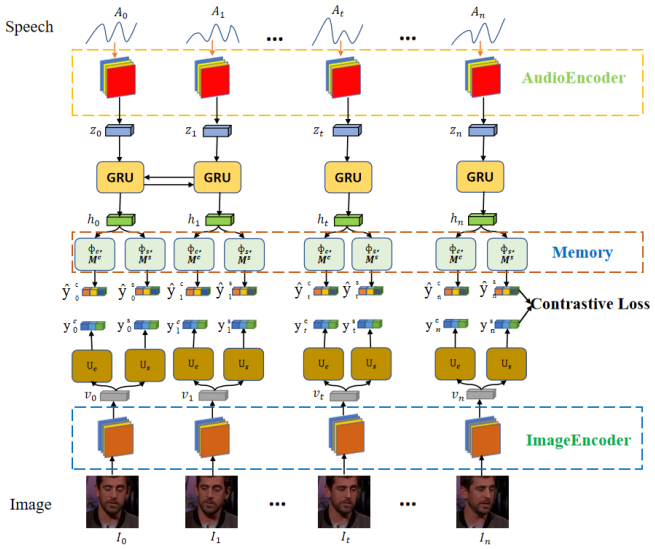


图 2。预训练 SFE 的整体架构。分别从语音和图像中提取内容特征和身份特征，并计算相应的对比损失。最后，取上半部分的音频模块作为音频特征提取器。

对比损失函数。通过计算语音特征与图像特征之间的距离，同步的语音-图像对的距离小于不同步的语音-图像对。身份对比损失函数和内容对比损失函数的形式相似。一组正样本对之间的对比损失函数如下：

$$L_c = -\frac{1}{T}\sum_{t=1}^{T} log\frac{S(\hat{y}_t^c, y_t^c)}{\sum_{k=1}^{T} S(\hat{y}_t^c, y_k^c)}, \quad (3)$$

$$S(\hat{y}_t^c, y_t^c) = exp(w \cdot cos(\hat{y}_t^c, y_t^c) + b), \quad (4)$$

其中，$\hat{y}_t^c$是语音的内容特征表示，y_t^c是图像序列的内容特征表示，t 是帧数，w 和 b是可学习参数。

2.2. 基于混合密度网络的面部关键点回归

大多数现有工作通过最小化单一高斯分布的负对数似然，即均方误差，来生成单个面部关键点。然而，从语音到头部运动和唇部运动的映射具有显著的不确定性，因此以往工作在提升面部关键点预测精度方面遇到了瓶颈。对于不确定性预测任务，混合密度网络是一种强大的工具。本文提出通过最小化混合高斯分布的负对数似然来估计多个面部关键点。MDN 模型的输出是一组混合系数 α 和高斯核参数（即均值μ 和方差 σ）。高斯核的均值即为生成的面部关键点，而混合系数和方差则表示每个生成的预测面部关键点的不确定性。

不同于以往方法，我们并不直接生成最终的目标面部关键点，而是首先生成对齐后的目标面部关键点 p_{align}、旋转矩阵 R_r和偏移矩阵 R_t，然后使用逆仿射变换获得最终的目标面部关键点，即从 p_{align} 中减去 R_t ，再将结果与R_r相乘。这减轻了面部关键点中身份特征的预测压力。

模型推理。 我们可以使用各分量输出的混合结果，或得分最高的分量输出，来生成面部关键点参数。我们尝试了这两种情况，发现使用最大分量能取得稍好的结果。

2.3. 图像到图像

最后，我们输入目标面部关键点和参考面部图像，生成最终的目标面部图像。我们使用图像到图像转换模块in [10]并对其进行微调。

3. 实验与分析

3.1. 实现细节

我们在 VoxCeleb2 [15] 上训练语音特征提取模型和面部关键点回归模型。该数据集包含来自 6,000 多位名人的 100 多万个话语，收集自 YouTube 上约 15 万个视频。该数据集的性别分布较为均衡（男性占 61%）。我们的网络采用 PyTorch 实现。训练期间，我们使用 Adam优化器，所有模型均在单块 NVIDIA Tesla V100 上训练和测试。对于图像流，我们首先提取视频中的所有图像，然后提取关键点。对于语音数据，采样率为 16 kHz。我们以 20 ms 的窗口大小和 10 ms 的帧移提取滤波器组特征。

3.2. 与最先进方法比较

我们提取 VoxCeleb1[16]测试集数据的身份嵌入，并直接使用自监督嵌入进行评估（不进行任何微调）。我们将结果与基线模型 [17]中发表的结果进行比较。如表 1 所示，就等错误率评估指标而言，我们的结果比 [18] 中的最佳结果低 1.05%，这表明我们的方法能够有效提取语音识别特征向量。

表 1。VoxCeleb1 测试集上的说话人验证结果。↓均值越低越好。

方法	等错误率（%）↓
仅 IL [17]	23.15
IL + CL [17]	22.59
IL +解耦损失 [17]	22.09
Ours	**21.04**

我们将所提出的模型与两种最先进方法进行比较 [9] [10]。定量结果如表2所示。我们使用LMD [18]（地标距离）来衡量唇部声音是否同步，并使用RD（旋转距离）来衡量生成的人脸视频是否具有自然的头部运动。为进一步评估不同方法生成图像的质量，我们比较了结构相似性指数 [19]和峰值信噪比 [20]。尽管MakitTalk在结构相似性指数和峰值信噪比方面优于所提出的方法，但其生成速度（帧每秒）低于所提出的方法。请注意，MakeitTalk和我们的方法生成的图像分辨率为 256 × 256，而ATVGNet生成的图像分辨率为 128 × 128。

表2。VoxCeleb2测试集上的评估结果。↑表示越高越好。

方法	LMD↓	旋转距离↓	结构相似性指数↑	峰值信噪比↑	帧每秒↑
ATVGNet [9]	2.12	0.21	0.81	28.14	**34.53**
MakeitTalk [10]	7.85	0.08	**0.83**	**29.77**	22
Ours	**1.83**	**0.07**	**0.83**	28.91	25.30

如图3所示，ATVGNet 仅专注于生成面部图像，而未生成自然的头部运动。虽然 MakeitTalk 能够生成细微的头部运动，但其模型生成的嘴型并不准确。与这些方法相比，我们的方法能够生成更加自然且与唇部同步的面部动画。

我们将各个组件生成的结果可视化，以研究每个混合组件学到的特征，如图4所示。各组件之间的主要差异在于它们学到的与说话人身份相关的特征，

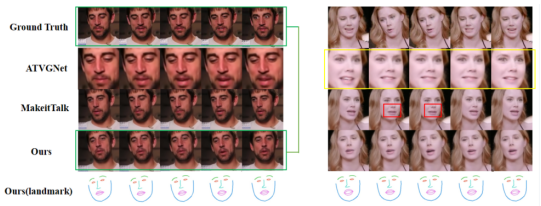


图3。ATVGNet 生成的定性结果 [9],MakeitTalk [10] 和我们的方法。ATVGNet 生成的面部动画没有头部运动（如黄色方框所示），MakeitTalk 生成的唇形不准确（如红色方框所示）。我们的结果呈现出更明显的头部运动趋势（如绿色方框所示）。

包括面部轮廓形状和头部运动。 $Comp_1$学到了纤细狭窄的面部轮廓， $Comp_2$ 学到了宽阔丰满的面部轮廓， $Comp_3$ 学到了随机的面部轮廓。其中，混合概率最大的组件 α_m 生成的面部轮廓最接近真实说话人。

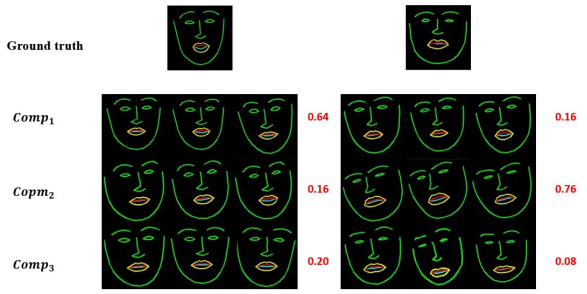


图4。混合密度网络展示了各混合组件的结果。右侧的红色数字是对应于各混合组件的混合系数。

3.3. 消融研究

对于自监督对比学习，我们通过消融实验研究完整模型中记忆模块的作用。对于面部关键点回归模型，我们分别通过消融实验研究完整模型中两个分量的作用：混合成分的数量和面部特征点回归方法。

记忆模块。如表3所示，通过比较不使用记忆模块的 Ours (wo)、仅使用单个记忆模块的 Ours (w)，以及使用身份记忆模块和内容记忆模块的 Ours (cs) 的实验结果，可以看出记忆模块对特征提取质量有显著影响。

混合成分数量。从表4可以看出，对于 LMD，当混合成分数量M 大于 3 时，其值基本不会随着 M 的增加而降低。对于旋转距离，当混合成分数量 M 大于 5 时，其值基本不会降低。

表3. 说话人验证中记忆模块消融实验结果对比。

方法	等错误率(%)↓	LMD↓	RD↓
我们的方法（不含）	22.12	0.14	0.08
我们的方法（含）	21.46	0.13	0.07
我们的方法（cs）	**21.04**	**0.12**	**0.07**

这表明每个分量主要学习与说话人身份相关的信息，而更多的分量可以对更多不同的说话人进行建模。

表4。混合组件数量对 LMD 的影响。

M	1	2	3	5	8
LMD↓	0.14	0.13	**0.12**	**0.12**	**0.12**
RD↓	0.08	0.08	0.07	**0.06**	**0.06**

面部特征点回归方法。在图5中，对于右侧具有较大头部运动的源视频，我们的方法（f_{tt}）能够生成更加自然的头部运动，而我们的方法（f_a）只能生成轻微的头部运动。尽管我们的方法生成的头部运动与真实值并不完全一致，但动态趋势相似。

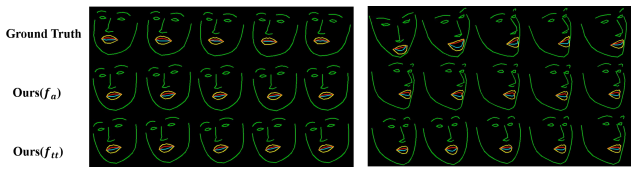


图5。不同面部关键点回归方法结果示意图。 f_a 表示直接回归面部关键点， f_{tt} 表示首先生成与标准人脸对齐的面部关键点、旋转矩阵和偏移矩阵，然后通过逆仿射变换获得面部关键点。

4. 结论

通过利用自监督对比学习、记忆模块和混合密度网络，所提出的方法能够合成具有自然头部运动的说话人头部视频。未来的工作将重点关注从语音中提取情感嵌入，并将情感嵌入添加到说话人头部生成中。

5. 致谢

本研究得到国家自然科学基金（编号：61977049）、国家自然科学基金（编号：62101351）以及广东省基础与应用基础研究基金（编号：2020A1515110376）的资助。

6. 参考文献

[1]Hyeongwoo Kim、Mohamed Elgharib、Michael Zollh¨ofer、Hans-Peter Seidel、Thabo Beeler、Christian Richardt 和 Christian Theobalt，《保留神经风格的视觉配音》，*ACM*图形学汇刊，第38卷，第1–13页，2019年。

[2] Li Liu、Gang Feng、Denis Beautemps 和 Xiao-Ping Zhang，《利用手部先行模型实现自动连续提示语音识别中的重新同步与多模态融合》，*IEEE* 多媒体学报，第 23 卷，第 292–305 页，2020年。

[3]Li Liu、Thomas Hueber、Gang Feng 和 Denis Beautemps，《使用串联 cnn-hmm 方法对连续提示语音进行视觉识别》，载于*Inter-speech*，2018年，第 2643–2647 页。

[4] Kyaw Kyaw Htike，《关于面向说话头模型的数据驱动学习的综述》，智能系统技术与应用国际期刊，第 16 卷，第 169–190 页，2017年。

[5] Lele Chen、Guofeng Cui、Celong Liu、Zhong Li、Ziyi Kou、Yi Xu 和 Chenliang Xu，《具有节律性头部运动的说话头生成》，载于 欧洲计算机视觉会议（*ECCV*）论文集，2020年，第 35–51 页。

[6] Lele Chen、Zhiheng Li、Ross K. Maddox、Zhiyao Duan 和 Chenliang Xu，《一瞥唇部动作生成》，载于欧洲计算机视觉会议（*ECCV*）论文集，2018年，第 538–553 页。

[7] Yang Song、JingwenZhu、Dawei Li、Xiaolong Wang 和 Hairong Qi，《通过条件循环对抗网络生成说话人脸》，*arXiv* 预印本：*1804.04786*，2018年。

[8] Konstantinos Vougioukas、Stavros Petridis 和 Maja Pantic，《利用生成对抗网络实现逼真的语音驱动面部动画》，《国际计算机视觉杂志》，第 128 卷，第 1398–1413 页，2020年。

[9] Lele Chen、Ross K. Maddox、Zhiyao Duan 和 Chenliang Xu，《结合动态逐像素损失的分层跨模态说话人脸生成》，载于*IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*）论文集，2019年，第 7832–7841 页。

[10]Yang Zhou、Xintong Han、Eli Shechtman、Jose Echevarria、Evangelos Kalogerakis 和 Dingzeyu Li，《Makelttalk：说话人感知说话头部动画》，*ACM*图形学汇刊，第39卷，第1–15页，2020年。

[11] Jianrong Wang、Ge Zhang、Zhenyu Wu、XueweiLi 和 Li Liu，《通过视频中的隐式线索进行自监督深度估计》，载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2021年，第2485–2489页。[12] Yixiong Chen、Chunhui Zhang、Li Liu、Cheng Feng、Changfeng Dong、Yongfang Luo 和 Xiang Wan，《Uscl：通过视频对比表征学习预训练深度超声图像诊断模型》，载于医学图像计算与计算机辅助干预（*MICCAI*），2021年，第627–637页。

[13] Ali Varamesh 和 Tinne Tuytelaars，《用于目标检测和人体姿态估计的混合密集回归》，载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2020年，第 13086–13095 页。

[14]Tengda Han、Weidi Xie 和 Andrew Zisserman，《用于视频表征学习的记忆增强型密集预测编码》，载于 欧洲计算机视觉会议（*ECCV*）论文集，2020年，第 312–329 页。

[15] Joon Son Chung、Arsha Nagrani 和 Andrew Zisserman，《Voxceleb2：深度说话人识别》，载于国际语音通信协会（*ISCA*）年会，2018年，第 1086–1090 页。

[16] Arsha Nagrani、Joon Son Chung 和 Andrew Zisserman，《Voxceleb：大型说话人识别数据集》，*arXiv* 预印本：*1706.08612*，2017年。

[17] Arsha Nagrani、Joon Son Chung、Samuel Albanie 和 Andrew Zisserman，《利用跨模态自监督的解耦语音嵌入》，载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2020年，第6829–6833页。

[18] Lele Chen、Zhiheng Li、Ross K Maddox、Zhiyao Duan 和 Chenliang Xu，《一瞥唇部动作生成》，载于欧洲计算机视觉会议（*ECCV*）论文集，2018年，第520–535页。

[19] Zhou Wang、Alan C Bovik、Hamid R Sheikh 和 Eero P Simoncelli，《图像质量评估：从误差可见性到结构相似性》，*IEEE*图像处理汇刊，第13卷，第600–612页，2004年。

[20] NiranjanD Narvekarand LinaJKaram，《基于模糊检测累积概率的无参考感知图像锐度指标》，载于多媒体体验质量国际研讨会，2009年，第87–91页。