文章编号：2095-2163(2023)05-0151-05 中图分类号：TP391；TP183 文献标志码：A

基于 LSTM-CBAM 的音视频同步人脸视频生成

洪学敏，张海翔
（浙江理工大学 信息学院，杭州 310018）

摘 要：语音驱动的人脸视频生成是指通过视觉与听觉双模态的输入来生成唇音同步的高自然度人脸视频。人脸视频生成任务的主要挑战是如何在保证人脸面部真实性的同时，生成语音同步且连贯的人脸视频。传统方法仅将其考虑为多个单帧的视频生成，而不考虑视频帧间的时序关系，从而导致生成的视频存在不连贯性，容易出现像素抖动问题。本文提出了基于 LSTM-CBAM 的音视频同步生成模型来生成唇音同步的人脸视频，通过 LSTM 模块处理音频数据，可以对音频数据进行更好地特征编码，通过 CBAM 模块来推断网络中的注意力映射，可以实现对音频信息与人脸口型信息的特征细化，从而生成音频与人物口型同步的视频。实验结果表明，本文生成的人脸视频连续自然，指标较优。

关键词：视频生成；语音驱动；生成式对抗网络

LSTM-CBAM-based audio and video synchronization face video generation

HONG Xuemin, ZHANG Haixiang
(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

【Abstract】 Speech-driven face video generation is a high naturalness face video with labial synchronization through visual and auditory dual mode input. The main challenge of face video generation task is how to generate voice synchronous and coherent face video while ensuring face authenticity. Traditional methods only consider it as multiple single-frame video generation without considering the sequence relation between video frames, which leads to the inconsistency of the generated video and the problem of pixel jitter. We propose an audio and video synchronization generation model based on LSTM-CBAM to generate labial synchronization face video. LSTM module is used to process audio data, and better feature coding can be performed on audio data. CBAM module is used to infer attention mapping in the network. It can realize the feature refinement of audio information and face-mouth-shape information, so as to generate audio and mouth-shape synchronization video. Quantitative experiments on LRS2 data set show that the face video generated in this paper is natural and continuous, and the index is better.

【Key words】 Video generation; speech driven; generative adversarial network

0 引 言

在日常生活中，听觉和视觉是人类最主要的沟通方式，这两种信号之间有着密不可分的联系，两者之间可以互相提供丰富的特征信息。例如，在人与人之间交流时，面部表情、说话口型、头部和身体动作可以有效提高信息的可理解性。根据研究，人与人之间的交流有 3 种方式：文字信息、语音信息和动作信息。其中文字信息占 7%，语音信息占 39%，动作信息占 54%。与文字信息、普通语音信息相比，动作信息与语音信息的共同输入可以更好地提高人类互动交流的感受。因此，利用听觉与视觉双模态的数据输入进行跨模态学习来生成基于语音驱动的说话人脸视频成为目前的一大热门研究课题。

语音驱动的人脸视频生成具体实现过程就是输入一段人脸视频和一段音频，利用神经网络进行特征编码，使得神经网络不断学习音频特征和视频特征，从而生成新的与音频相匹配的说话人脸视频。其研究目的是为了挖掘音频特征与人脸之间的关联性，单张静态人脸图像之间有年龄、性别等多种属性关联[1]，连续多张动态人脸图像之间人脸嘴唇具有同步性，这意味着要求生成的说话人脸视频要自然真实，输入的语音要与生成的说话人脸视频口型一致。因此，语音驱动说话人脸视频生成方法需要综

作者简介：洪学敏(1996-)，女，硕士研究生，主要研究方向：计算机视觉；张海翔(1973-)，男，博士，副教授，主要研究方向：计算机视频图像处理、计算机视觉、深度几何学习方法。
通讯作者：张海翔 Email：zhhx@zstu.edu.cn

收稿日期：2020-05-29

合考虑上述两方面因素,才能更好地将其应用到实际生活中去。

本文工作的主要贡献概括为两个方面:一是给出了基于生成对抗式网络的人脸视频生成方法,可以有效地提高人脸视频生成质量。二是提出了基于 LSTM-CBAM 的音视频同步判别器,可以辅助生成语音与口型同步的人脸视频。本文方法比现有的其它方法性能更佳。

1 相关工作

1.1 生成式对抗网络

生成式对抗网络(Generative Adversarial Network, GAN)[2]是通过学习已有样本的分布,生成与已有样本相似的样本,该模型的训练使用对抗博弈的思想进行。对抗博弈思想是博弈双方通过互相约束与督促,使得博弈双方都在不断变强的过程。GAN 模型包含生成器和判别器,生成器与判别器通过对抗博弈使二者都更强大,最终使模型生成更真实的样本。其中,生成器可以训练学习到与真实样本相似的分布,从而得到虚假样本,而虚假样本可以欺骗鉴别器;判别器则可以区别数据分布是来自真实样本还是虚假样本。具体训练流程如图 1 所示。

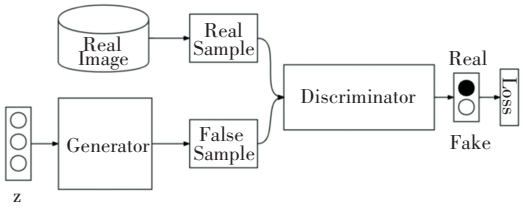

图 1 生成式对抗网络

Fig. 1 Generative adversarial network

原始 GAN 的优化目标函数如公式(1):

$$\min_G \max_D V(D,G) = E_{x\sim P_{data}(x)}[\log D(x)] + E_{Z\sim P_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

其中, E 是分布函数的期望值; x 是真实样本; $p_{data(x)}$ 是真实样本分布; z 是低维噪声; $p_z(z)$ 是低维噪声分布。

1.2 卷积块注意力模块

注意力机制[3]的作用,是告诉模型特征图中的哪些区域更应该被关注。注意力模块将某一位置的响应表达为所有位置对这一位置的特征加权,而权重和注意力向量的计算仅需要很小的计算成本。

卷积块的注意力模块(CBAM)[4],在前馈卷积神经网络中具有出色的效果。其在每个网络层都使用一个独立的深度学习框架来进行处理,并将注意力模块集成于整个模型之中。这种方法能够获得较好的结果,并且可以被广泛地应用于各种任务中。给定 CBAM 模块一个中间特征映射,沿通道和空间维度的注意力映射,可以实现特征细化。CBAM 是轻量级的模块,可以无缝集成到任何卷积网络架构中。

CBAM 模块由通道注意力模块(Channel Attention Module, CAM)和空间注意力模块(Spatial Attention Module, SAM)组成。通道注意力模块用于处理不同通道的特征图,并告知模型对这些特征图给予更多关注;空间注意力模块用于处理特征图上的特征区域,并通知模型应更多地注意这些特征区域。

2 音视频同步人脸视频生成方法

2.1 模型介绍

本文提出了一种基于 LSTM-CBAM 的音视频同步人脸视频生成方法,其中模型框架采用的是生成式对抗网络,由一个人脸视频生成器和一个音视频同步判别器组成,音视频同步人脸视频生成方法整体结构如图 2 所示。生成器与 LipGAN[5]方法的生成器类似,采用编码器-解码器结构,包含音频编码器、视频编码器、人脸解码器。SyncNet[6]是一种优秀的纠正人物口型与音频同步错误的方法,本文对该方法做了改进,并将其作为音视频同步判别器。

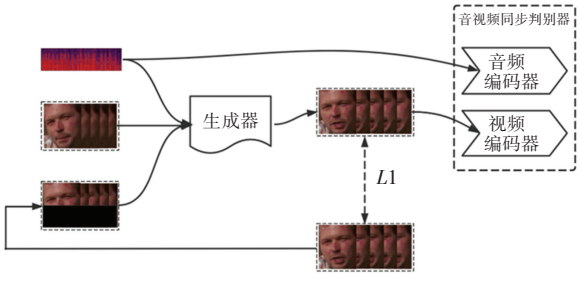

图 2 人脸视频生成方法整体结构

Fig. 2 Face video generation method overall structure diagram

音视频判别器模型由音频编码器和视频编码器组成,其结构如图 3 所示。音频编码器包含一个 LSTM[7]模块和多个卷积块,其中包含一些残差块,并且在每一卷积层后都添加了 CBAM 模块,这就意味着每一卷积层后得到的特征图多了通道注意力和空间注意力,可以更好地学习有意义的音频特征。视频编码器包含多个卷积块和残差块,与音频编码器一样,在每个卷积层后添加了 CBAM 模块,可以更好地学习人脸口型中有意义的视频特征。在音频编码器和视频编码器中,每个 CBAM 模块后都有一

个归一化层和 Relu 激活函数。

CBAM 模块由通道注意力模块和空间注意力模块共同组成，其完整的网络结构如图 4 所示。

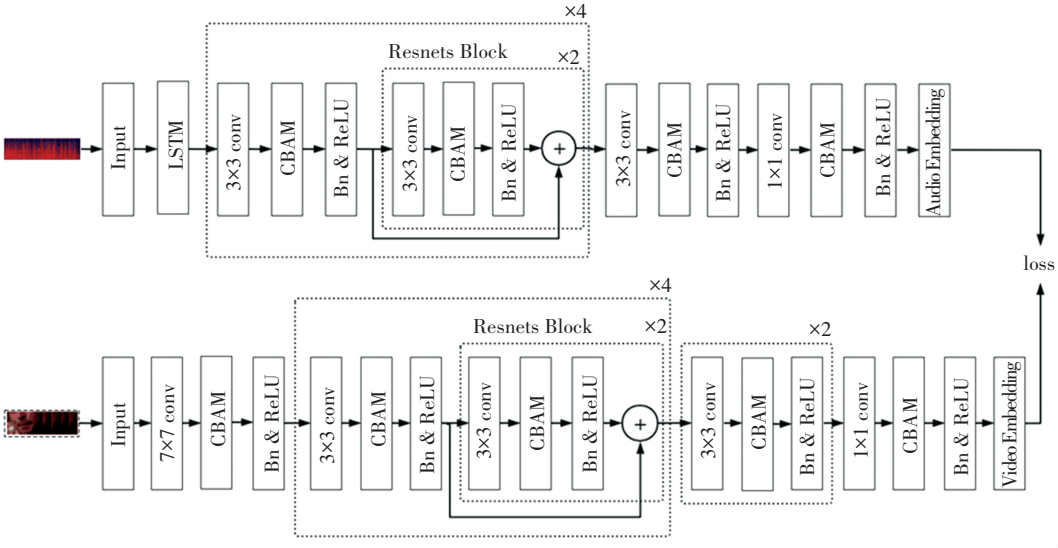

图 3　同步判别器模型结构图

Fig. 3　Structure diagram of synchronous discriminator model

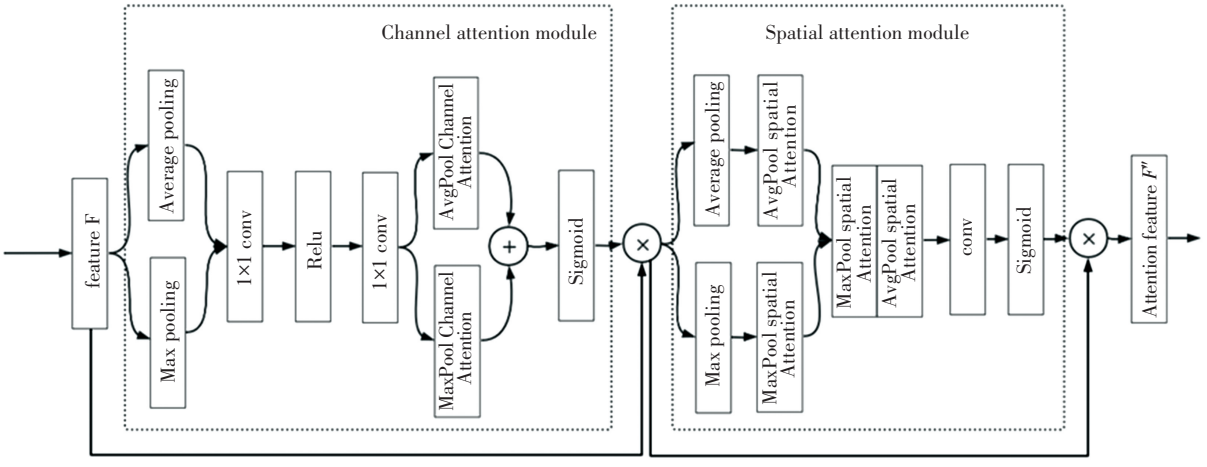

图 4　CBAM 模块模型结构图

Fig. 4　Structure diagram of CBAM module model

由于本文中在音视频同步判别器所有卷积层后都加入了 CBAM 模块，每一层的特征图不同，这里将特征图向量 $\boldsymbol{F}$ 表示为 $[C, H, W]$。将一个特征图输入到 CBAM 模块，依次计算通道注意力图 M_c 和空间注意力图 M_s，整体注意力过程如公式(2)：

$$F' = M_c(F) \otimes F$$
$$F'' = M_s(F') \otimes F' \tag{2}$$

其中，$\otimes$ 将特征图相乘，F'' 是带通道和空间注意力的新特征图。

2.2　训练过程

在训练过程中，判别器和生成器交替训练，通过训练让两个模型同时得到增强。两者都使用 Adam[8] 作为优化器，学习率为 1e-4。需要注意的是，Wav2Lip[9] 方法中采用了预训练同步判别器的方法，可以使同步判别器自身有更强大的判别力。因此，本文也采用预训练判别器的方式进行模型训练。

2.3　数据集

本文采用 LRS2[10] 数据集，该数据集是从 BBC 电视广播中收集的大规模视频数据组成。其中包括 100 万个单词的实例，由超过 3 700 个不同人物录制的短视频。数据集以说话人物分类，同一个人会有几个或几十个视频文件以及与视频对应的单词文

件,但是不包含音频数据。其中,训练、验证、测试集划分比例分别为 95%、2%、3%。

2.4 损失函数

本文实验总体采用了生成式对抗网络模型进行训练,采用 *L*1 损失与 *GAN* 损失结合来约束生成器训练生成音视频同步的人脸视频。

2.4.1 生成器的 *L*1 重构损失函数

生成器与 LipGAN 等模型类似,相当于一个自编码器,对生成视频帧与真实样本帧中的每一帧计算 *L*1 损失,使得生成的帧与真实视频帧之间的 *L*1 重构损失最小化,如公式(3):

$$L1 = \frac{1}{N}\sum_{i=1}^{N} \| L_g - L_t \|_1 \tag{3}$$

该公式表明, *L*1 重构损失越小,生成的视频帧与真实样本帧越相似,生成的视频越真实。

2.4.2 同步判别器损失函数

本文使用 wav2Lip 中提出的 P_{sync} 损失函数,该函数用二元交叉熵损失的余弦相似度,为每个样本生成一个[0,1]之间的值。在音视频判别器中,视频编码器对视频序列提取特征,语音编码器对语音序列提取特征,并通过二元交叉熵损失的余弦相似度来计算音频特征与视频特征之间的损失。输入音频与视频对同步的概率如公式(4),音视频同步判别器损失如公式(5)。

$$P_{sync} = \frac{v \cdot s}{\max(\| v \|_2 \cdot \| s \|_2, \varepsilon)} \tag{4}$$

$$E_{sync} = \frac{1}{N}\sum_{i=1}^{N} -\log(P_{sync}) \tag{5}$$

2.4.3 总体损失函数

模型的总体损失函数通过结合 *L*1 重构损失(式(3))、Esync 同步损失(式(5))的加权和来得到,如公式(6):

$$L = \lambda_1 \cdot L_1 + \lambda_2 \cdot E_{sync} \tag{6}$$

其中, λ_1 是 *L*1 重构损失惩罚权重, λ_2 为同步损失惩罚权重。

2.5 评价指标

2.5.1 图像感知相似度评价指标

在生成任务时,使用生成对抗网络通常会引入随机噪声,以增加生成样本的多样性,虽然生成样本与真实样本有所不同,但分布是相同的。因此,视频质量评价指标(PSNR、SSIM)则不适合对抗网络生成样本的评价指标。FID[11] (Frechet Inception Distance)是一种图像感知相似度评价指标,其是计算真实图像和生成图像特征向量之间距离的一种度量,常常用来评估生成式对抗网络生成的图像的真实性。因此,本文将采用 FID 作为评判图像真实性的指标。FID 越低,两组图像就越相似,代表得到的视频帧更具有真实性。真实图像分布与生成器生成分布之间的差异,即 FID 分数如公式(7)所示:

$$FID(g,r) = \| \mu_g - \mu_r \| + T_r(\sum\nolimits_x + \sum\nolimits_g - 2(\sum\nolimits_g \sum\nolimits_r)^{\frac{1}{2}}) \tag{7}$$

其中, *g* 代表生成图像; *r* 代表真实图像; μ_r、μ_g 分别表示真实图像与生成图像特征向量的均值; $\sum_g$、$\sum_r$ 分别表示生成图像与真实图像特征向量的协方差矩阵; *Tr* 表示矩阵的迹,矩阵开根如果为复数,则只取实部。

2.5.2 口型-语音同步评价指标

语音驱动人脸视频生成的重要目标是音频与视频中人物口型保持同步。本文使用 SyncNet 方法中评价口型-语音同步的方法作为评价指标,该方法通过训练视频片段的语音特征和视频特征,计算其欧式距离,然后再由视频片段组成的原视频中找到最小欧式距离,这个最小欧式距离将作为人脸口型与语音的偏差指标(LSE-D)。当 LSE-D 越低,表示人脸口型时序上越连贯。LRS2 数据集中的方法是使用欧式距离的最小值和中位数之差作为人脸口型与语音的置信度分数(LSE-C)。当 LSE-C 越高,表示人脸口型与语音相关程度越高。

3 实验结果分析

为了验证所提出的 LSTM 模块与 CBAM 模块在音视频同步判别器在模型中的效果与性能,并使用 FID、LSE-D、LSE-C 来衡量生成质量,并将在 LRS2 数据集上进行消融实验,实验结果见表 1。表 1 中,"ours"表示本文提出的基于 LSTM-CBAM 的音视频同步判别器的生成方法;"w/o LSTM&CBAM"表示缺少 LSTM 模块和 CBAM 模块的同步判别器;"w/o LSTM"表示缺少 LSTM 模块的同步判别器;"w/o CBAM"表示缺少 CBAM 模块的同步判别器。

表 1 消融实验结果

Tab. 1 Ablation experiment results

算法	LSE-D ↓	LSE-C ↑	FID ↓
w/o LSTM&CBAM	6.723	6.802	4.387
w/o LSTM	6.537	7.081	4.456
w/o CBAM	6.552	7.111	4.449
Ours	6.207	7.455	4.806

由表 1 可见，“w/o LSTM”和“w/o CBAM”都比“w/o LSTM &CBAM”取得了更低的 LSE-D 值、更高的 LSE-C 值，而“ours”比“w/o LSTM”和“w/o CBAM”取得了更低的 LSE-D 值、更高的 LSE-C 值；“ours”相比“w/o LSTM &w/o CBAM”，LSE-D 值下降了 7.7%，LSE-C 值提升了 9.6%，证实了 LSTM 模块与 CBAM 模块可以有效提高音视频同步判别器性能。

Speech2Vid 模型使用了传统的编码器-解码器结构，用音频与人脸图像的联合嵌入，分别用音频编码器和身份编码器进行特征提取，将人物特征和音频特征输入到人脸图像生成解码器，用来生成说话人脸的视频帧，不足之处在于该方法只对每帧视频帧计算 $L1$ 损失。LipGAN 模型使用生成对抗方法，并使用同步判别器，但该方法的同步判别器每次仅处理一帧视频，虽然有效的保证了单帧视频音视频同步，但视频缺乏连贯性，容易出现视频抖动问题。Wav2Lip 模型也使用生成对抗方法，模型有较好的音视频同步能力，生成的视频视觉质量也较好，但整体性能相比本文提出的模型有所欠缺。见表 2，本文方法有较低的 FID 值，这意味着本文方法生成的人脸视频和真实视频在特征层面的距离最接近，即有更高的质量。本文方法有较低的 LSE-D 值，以及相对较高的 LSE-C 值。虽然 LSE-D 值相较于 Wav2Lip 模型略差一点，但整体结果相对较好，这意味着本文模型能有较好的口型-音频同步能力，具有更好的性能。

表 2　对比实验结果

Tab. 2　Comparative experiment results

数据集	算法	LSE-D ↓	LSE-C ↑	FID ↓
LRS2	Speech2vid	14.23	1.587	12.32
	LipGAN	10.33	3.199	4.861
	Wav2Lip	6.386	7.789	4.887
	Ours	6.207	7.455	4.806

4　结束语

本文提出了基于生成对抗网络的音视频同步人脸视频生成方法，并提出了基于 LSTM-CBAM 的音视频同步判别器，在 LRS2 数据集上进行对比实验和消融实验，结果表明本文方法生成的人脸视频生成效果在定量评估上结果较好，证明了本文方法的有效性。

参考文献

[1] 赵璐璐. 基于静态属性和动态关联的语音驱动人脸生成方法研究[D]. 合肥：合肥工业大学，2021.

[2] GOODFELLOW I，POUGET-ABADIE J，MIRZA M，et al. Generative adversarial nets[C]//Advances in neural information processing systems. 2014：2672-2680.

[3] BAHDANAU D，CHO K，BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv：1409.0473，2014.

[4] WOO S，PARK J，LEE J Y，et al.Cbam：Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018：3-19.

[5] KR P，MUKHOPADHYAY R，PHILIP J，et al. Towards automatic face-to-face translation[C]//Proceedings of the 27th ACM International Conference on Multimedia. 2019：1428-1436.

[6] CHUNG J S，ZISSERMAN A. Lip reading in the wild[C]//Asian conference on computer vision. Springer，Cham，2016：87-103.

[7] HOCHREITER S，SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation，1997，9(8)：1735-1780.

[8] KINGMA D P，BA J. Adam：A method for stochastic optimization[J]. arXiv preprint arXiv：1412.6980，2014.

[9] 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.

[10] AFOURAS T，CHUNG J S，SENIOR A，et al. Deep audio-visual speech recognition[J]. IEEE transactions on pattern analysis and machine intelligence，2018，44(12)：8717-8727.

[11] HEUSEL M，RAMSAUER H，UNTERTHINER T，et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in neural information processing systems，2017，2(3)：30.