基于模态仿射融合的语音控制说话人脸视频对抗生成

陈诗航，孙玉宝

（南京信息工程大学计算机学院，江苏 南京 210044）

摘　要：语音生成说话人脸视频是当前一个研究热点，涉及音频和视觉两个模态的处理，需要着重解决说话时唇部运动和输入音频对齐的问题。针对该问题提出一种端到端的语音控制说话人脸视频生成对抗模型，主要包括模态仿射融合的生成器、视觉质量判别器和唇形同步判别器，基于仿射融合的生成器通过模态仿射融合模块（MAFBlock），在人脸特征解码过程中添加音频信息，有效地融合音频信息和人脸信息，使得音频能够更好地控制说话人脸视频生成。引入空间注意力和通道注意力机制，增强模型对于局部区域的关注。基于双判别器提高模型生成质量和唇形同步率，唇形同步判别器用于约束唇部运动，对音频和唇形进行相似性判断，在不改变整体轮廓和脸部细节的前提下更精细地控制唇部动作生成，视觉质量判别器判断生成图片的真实性，提高生成图片质量。在两个视听数据集上与多个现有的代表性模型进行对比实验，结果表明：该模型在 LRS2 验证集上具有 8.128 的 LSE-C 分数和 6.112 的 LSE-D 分数，相比于 Baseline 分别提升了 4.3%和 4.4%；在 LRS3 验证集上具有 7.963 的 LSE-C 分数和 6.259 的 LSE-D 分数，相比于 Baseline 分别提升了 6.2%和 6.9%。

关键词：说话人脸生成；视频生成；唇形同步；音频驱动生成；空间注意力；通道注意力

中图分类号：TP391　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0069992

Adversarial Generation of Voice-Controlled Speaking Face Videos Based on Modal Affine Fusion

CHEN Shihang, SUN Yubao

(School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, Jiangsu, China)

【Abstract】Generating speaking face videos from speech, involving the processing both audio and visual modalities, is a current research hotspot. A key challenge is achieving precise alignment between lip movements in the video and the input audio. To address this problem, this study proposes an end-to-end, speech-controlled speaking face video generation adversarial model, which mainly consists of a modal affine fusion-based generator, a visual quality discriminator, and a lip synchronization discriminator. The affine fusion-based generator adds audio information during face feature decoding through the Modal Affine Fusion Block (MAFBlock), effectively fuses audio information with face information and enables the audio to be better controlled for speaking face video generation. Spatial and channel attention mechanisms are incorporated to enhance the model's focus on local facial regions. The model employs a dual-discriminator strategy to enhance both visual quality and lip synchronization accuracy. The lip synchronization discriminator constrains lip movements by evaluating the similarity between the audio and the generated lip shapes without changing the overall contour and face details, thereby providing finer control over lip movement generation. The visual quality discriminator assesses the realism of the generated image frames to improve image quality. A comparative experimental analysis is conducted with several existing representative models on two audiovisual datasets. On the LRS2 validation set, the proposed model achieves an LSE-C score of 8.128 and an LSE-D score of 6.112, which are 4.3% and 4.4% higher than those of the baseline, respectively. On the LRS3 validation set, it achieves LSE-C and LSE-D scores of 7.963 and 6.259, representing improvements of 6.2% and 6.9% over the baseline scores, respectively.

【Key words】speaking face generation; video generation; lip synchronization; audio-driven generation; spatial attention; channel attention

0　引言

视频是最常见的娱乐方式之一，人类对于视频的感知通常涉及两种模态，分别是听觉和视觉。通常一个视频往往伴随着各种听觉信息，例如人说话会伴随嘴部和面部运动。对于一个视频来说，如果

基金项目：国家自然科学基金（U2001211，62276139）。

作者简介：陈诗航（CCF 学生会员），男，硕士研究生，主研方向为图像生成；孙玉宝（通信作者），教授、博士。

收稿日期：2024-06-12　　**修回日期：**2024-09-22　**E-mail：**sunyb@nuist.edu.cn

想扩展其在全球的受众群体,将视频翻译通常是最简单的方式之一,但直接的翻译无法提供自然的视听体验,翻译后的语音和嘴唇缺乏同步。因此,生成与音频同步的唇部运动对于增强观看体验至关重要。

基于音频生成说话人脸是一项很有前景的技术,它能根据目标身份合成出唇音同步的逼真视频。将该技术应用于视频配音,可以显著提升唇形的一致性,从而改善整体视听体验。区别于传统的视频生成任务,说话人脸生成不需要从零生成一段视频,它只需要改变已经存在视频的唇部区域,而且直接从原视频中修改也保证了帧之间的连贯性。

现有的编码器-解码器结构[1-2]为生成逼真的说话人脸视频做出了关键贡献,极大地推动了该领域的发展。Wav2Lip[1]把音频编码器和人脸编码器的特征进行拼接并将其作为联合特征,通过多层上采样生成说话人脸视频,图像的生成质量较高,但唇形同步仅靠损失函数约束,使得部分唇形动作错误。MITTAL 等[3]将音频表示分离为语音内容、情感音调和其他因素。通过特征解耦去除在唇形同步过程中无关的因素。然而,解耦后的音频特征缺乏明确包含唇部运动的视觉信息,这些缺乏的信息可以帮助解码器从音频映射到图像中。在文生图任务中,跨模态注意力机制虽然能够有效融合文本和图像信息,但其计算复杂度较高。考虑到本文的任务专注于生成唇形区域而非整幅图像,直接采用跨模态注意力机制可能会导致大量计算资源的浪费,因此需要更加高效的方式对视觉信息和音频信息进行深度融合。

为了解决现有方法存在的问题,基于已有的数据集[4-5],本文提出了一种说话人脸合成框架,首先单独训练一个可以辨别音频和视频中唇部运动是否相符的判别器,利用该判别器为后续唇部运动进行约束;然后进行人脸生成,提出了一种仿射变换的融合方式,同时为了提高模态间的交互,添加了通道注意力机制,将多尺度人脸和音频特征进行融合;最后使用 styleGAN[6]跳跃连接的方式来提高图片生成质量。本文的贡献主要有以下 3 个方面:

1)提出一种高效的基于仿射变换的视觉听觉融合模块,该模块在深层次对视觉和音频进行融合,充分利用两种模态的信息来促成说话人脸视频的生成。

2)为了进一步提高模型唇部生成的准确性,对模态融合的特征添加唇形判别器来提高模型对唇部特征的关注。

3)为了提高模型训练的稳定性,对视觉质量判别器和唇形同步判别器添加了谱归一化。

1 相关工作

1.1 生成对抗网络(GAN)

自 GOODFELLOW 等[7]提出 GAN 以来,GAN 便成为图像生成领域的主要模型之一,它通过解决生成器和判别器之间的最小-最大优化问题来生成图片。但是,GAN 也存在训练困难、生成样本缺乏多样性等问题,许多研究尝试通过改善损失函数、网络结构设计和训练策略等方面去缓解这些问题。针对 GAN 的训练困难这一问题,WGAN[8]引入了 Wasserstein 距离减少生成器在训练过程中遇到的梯度消失的问题,缓解了训练不稳定的问题。但是,WGAN 中的权重裁剪却带来了一些问题,例如可能会导致梯度爆炸或消失。后来 WGAN-GP[9]做出了改进,添加了梯度惩罚,通过在目标函数中增加一个额外的惩罚项来满足利普希茨连续。MIYATO 等[10]提出谱归一化的使用,主要是使用判别器权重,使判别器满足利普希茨连续性,函数变化更加平稳,模型更加稳定。在另外一些工作中[11]也通过实验验证了谱归一化能够有效约束判别器的性能,间接提高模型的生成能力。

作为图像生成领域的代表性 GAN,StyleGAN 系列模型[12]在图像生成任务上展现了出色的能力,可以生成保真度非常高的人脸,具有逼真的脸部细节。StyleGAN 本身很难实现可控的图像合成效果,在其他作品中通过限制 StyleGAN 来执行有条件的图像生成,例如 StyleCLIP[13]将 StyleGAN 的图片生成能力和 CLIP[14]的多模态能力相结合,通过文本驱动图像生成。本文以 GAN 网络作为基础架构,将音频特征通过仿射变换与视频特征相结合来实现有条件图片生成,同时为了保证训练的稳定性,添加谱归一化和梯度惩罚。

1.2 语音驱动的说话人脸生成

从模型的泛化性角度,说话人脸视频生成分为两种:一种是受约束的说话人脸视频生成,一些方法展示了特定说话人的高质量视频的生成[15-18],SUWAJANAKORN 等[15]通过奥巴马每周演讲的几个演讲片段合成了高质量视频,后续方法[19-20]在缩短数据所需的时间的同时保证了高质量视频生成;另一种是根据任意人物生成语音驱动的视频,该任务[2,21]给定一个随机面部以及一个随机音频片段,生成一个唇形和音频同步的视频,这类任务的应用更加广泛。

生成说话人视频的典型方法是采用编码器-解码器结构以实现端到端的面部视频合成。LipGAN 输入了下半部分被遮掩的目标面部，将其作为先验姿势，这对于人脸生成至关重要，可以直接将裁剪的人脸生成后覆盖原视频，不需要经过额外的处理。Wav2Lip 是 LipGAN 的进阶版本，额外采用以唇语同步为条件的预训练鉴别器，通过鉴别器来约束唇形同步，通过对抗学习来提高图片生成质量。随着三维(3D)面部重建技术的发展，许多工作探索从视频中提取 3D 可变形模型(3DMM)来表示面部运动，通过使用 3DMM 进行建模，模型能够控制头部姿势和面部运动。ZHOU 等[22]提出了一种新颖的联想和对抗训练过程，用于音频和视频在潜在空间中隐式建模。之后，ZHOU 等[23]又提出了另一种用于说话面部生成的隐式建模方法，使用姿势参考视频来控制头部姿势。

在当前先进的方法中，特征融合仅通过模态间简单拼接的方式实现，这种简单的拼接操作虽然计算效率较高，但无法充分捕捉模态间的复杂关联。此外，基于交叉注意力机制的模态融合和 3D 面部重建的方法与本文的任务契合度较低。前者尽管能够捕获模态间的复杂交互，但由于其计算复杂度高，在基础模型上增加数倍的参数，浪费了宝贵的计算资源；后者则需要在每次生成新的人脸视频时重新训练模型，导致效率低下并增加了实现难度。本文以编码器-解码器结构作为说话人面部生成的基本结构，利用模态仿射变换模块融合音频和视频，以一种高效的方法深度融合视频和音频信息。

2 本文模型

本文提出的模型结构基于编码器-解码器架构，如图 1 所示(彩色效果见《计算机工程》官网 HTML 版，下同)。人脸编码器和音频编码器由多层卷积层构成，能够有效捕捉视觉和音频输入的复杂特征。在解码器中，模型引入了模态仿射融合模块(MAFBlock)，深度融合图像和音频特征，确保唇部动作与语音内容的精确同步。为了增强生成图像的真实性和质量，模型结合了唇形同步损失，提升了唇形生成的准确性，并通过视觉质量判别器和 L1 重建损失生成真实、高质量的面部图像。

2.1 唇形区域视频掩码提取

在视频处理部分，将视频逐帧提取，利用人脸检测模型裁剪出人脸部分尺寸为 96×96 像素的区域。参考 LipGAN 中人脸拼接方式，选取连续 T 帧作为目标人脸，对目标人脸的下半部分进行掩码操作，避免向模型泄露唇部信息。模型的目标是将遮掩人脸生成为目标人脸，但是遮掩人脸仅包含部分特征，模型需要额外的面部细节为模型生成提供特征，所以添加其他帧作为参考帧。

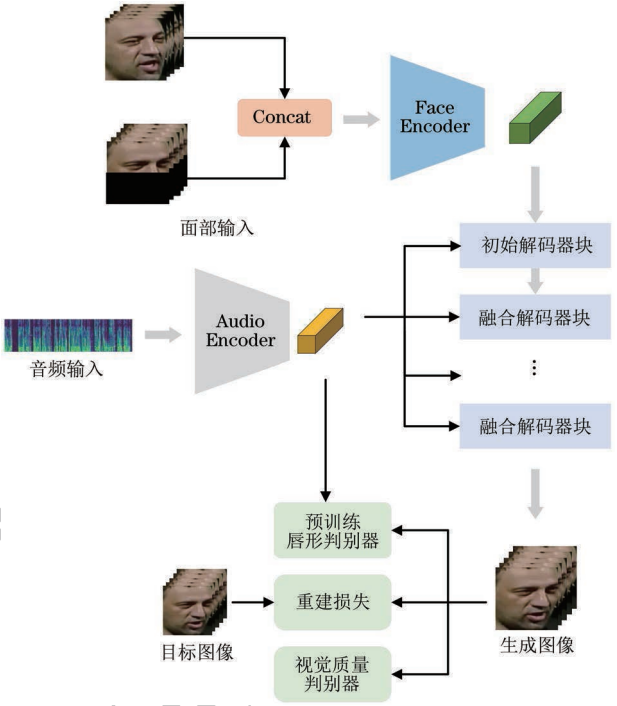

图 1 基于模态仿射融合的说话人脸视频生成模型

Fig. 1 Speaking face video generation model based on modal affine fusion

本文选取和目标人脸不同的连续 T 帧人脸作为参考人脸，为模型生成目标人脸提供整体特征。如图 2 所示，模型根据两者拼接的输入生成目标人脸，获得视频输入窗口 W_V。为了使用唇形同步判别器约束唇形生成，视频窗口长度设置为与唇形判别器中相同的窗口长度 T。

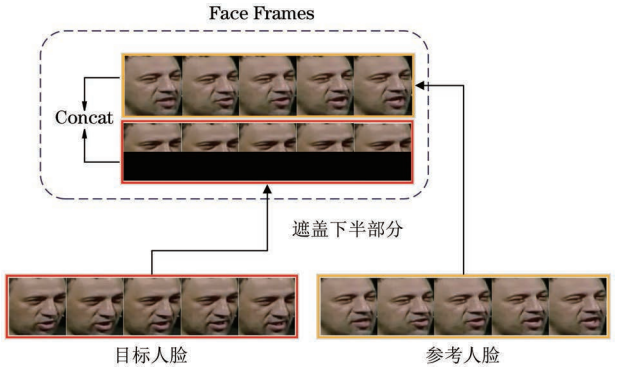

图 2 人脸预处理

Fig. 2 Face pre-processing

在音频处理中，从视频源中提取音频部分，并使用梅尔频谱图表示，如图 3 所示，根据选取的目标人脸裁剪出对应时间步长的音频。视频对应帧数和视频的帧率有关，而音频的时间步长取决于音频的采

样率，训练数据中的音频采样率都是 16 kHz 的，在本文设置中，视频窗口 W_V 的时间步长设置为 5，视频的帧率设置为 25 帧/s，音频帧移设置为 200，窗口大小设置为 800，音频窗口 W_A 的长度为 16。

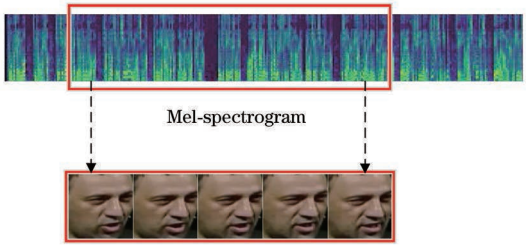

图 3 音频预处理

Fig. 3 Audio pre-processing

2.2 MAFBlock

为了高效融合音频和图像特征，提出了 MAFBlock。相比于基于 GAN 的说话人脸视频生成方法，MAFBlock 深化了音频和图像的融合过程，实现了更加细粒化的音频图像融合。MAFBlock 由多个仿射融合和激活函数组成，每个仿射融合由两个结构相同的 MAP 层构成，仿射融合如图 4 所示，分别通过 MAP 层从音频特征向量 $\boldsymbol{F}_A$ 中提取出模态通道缩放参数 λ 和模态通道偏移量 ε，如式(1)和式(2)所示。对于视觉特征图的缩放操作如式(3)所示。

$$\mathrm{MAP}(\boldsymbol{F}_A)=\mathrm{MLP}(\mathrm{ReLU}(\mathrm{MLP}(\boldsymbol{F}_A))) \tag{1}$$

$$\lambda=\mathrm{MAP}_1(\boldsymbol{F}_A)$$

$$\varepsilon=\mathrm{MAP}_2(\boldsymbol{F}_A) \tag{2}$$

$$\mathrm{AFF}(\boldsymbol{F}_A|F_i)=\lambda_i\cdot F_i+\varepsilon_i \tag{3}$$

式中：AFF 表示仿射融合；ReLU 表示 ReLU 激活函数的改进版本 LeakyReLU，正数部分保持不变，负数部分乘上一个较小的值，这样可以缓解死亡 ReLU 问题，后面所有未特殊说明的均为该改进版本的 ReLU 函数；模态通道缩放参数 $\lambda\in\mathbb{R}^{B\times1\times1\times W}$，$\lambda_i$ 表示视觉特征图第 i 个通道的模态通道缩放参数，与视觉特征图第 i 个通道数相乘来进行缩放操作；模态通道偏移量 $\varepsilon\in\mathbb{R}^{B\times1\times1\times W}$，$\varepsilon_i$ 表示第 i 个通道的模态通道偏移量，用来对通道进行细粒度调整，将对应特征图的对应通道中每个数和该数进行相加；F_i 是视觉特征图的第 i 个通道；$\boldsymbol{F}_A$ 表示音频特征向量。

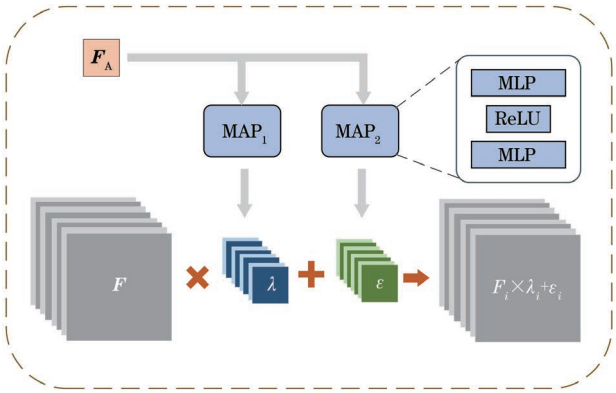

图 4 仿射融合

Fig. 4 Affine fusion

仿射融合对音频和图像进行深度融合，扩展了条件表示空间，但是由于仿射融合本身是线性运算，因此在每个 MAFBlock 后引入 ReLU 激活函数，为网络注入非线性，从而增强其表征能力，使其能够捕捉音频与视觉特征间更复杂的关联。

MAFBlock 结构如图 5 所示，在其中包括两个仿射融合和 ReLU 激活函数，通过 MAFBlocak 能够让生成器在生成过程中更加充分地融合音频特征。

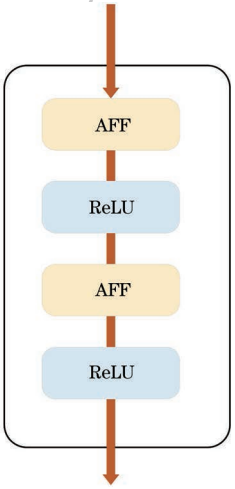

图 5 MAFBlock 结构

Fig. 5 MAFBlock structure

2.3 基于模态仿射融合的生成器

如图 6(a)所示，生成器有 3 个模块：人脸编码器，音频编码器和人脸解码器。音频编码器的输入就是在预处理阶段获得的音频窗口，根据一系列卷积网络进行下采样操作得到音频特征向量 $\boldsymbol{F}_A$。人脸编码器的输入是在预处理阶段得到的人脸拼接视频窗口，编码器是一系列具有残差的卷积网络，视频窗口经过编码器得到视频特征。音频特征和视频特征经过解码器块生成连续人脸帧。重建损失 L_{recon} 通过计算生成人脸帧与真实人脸窗口间的平均绝对误差(MAE)得到。生成器的重建损失如式(4)所示：

$$L_{\mathrm{recon}}=\frac{1}{N}\sum_{i=1}^{N}|L_g-L_G| \tag{4}$$

式中：L_g 是来自生成器生成的图片；L_G 是真实的图片。

解码器部分由 6 个解码器块构成，每个解码器

块中包含一个 MAFBlock。解码器块分为初始解码器块和融合解码器块。两者整体结构相似，唯一的不同点在于初始解码器中需要根据人脸编码器生成最初的 RGB 图。

在初始解码器块中，如图 6(b)所示，将 $\boldsymbol{F}_{\mathrm{V}}$ 上采样并经过 ToRGB 模块得到初始的 RGB 图，ToRGB 模块将特征图转换成指定分辨率的 RGB 图，上采样部分由一个双线性上采样和两层卷积构成。$\boldsymbol{F}_{\mathrm{V}}$ 和 $\boldsymbol{F}_{\mathrm{A}}$ 通过 MAFBlock 融合，经过一层卷积块注意力模块(CBAM)[24]和 ToRGB 模块生成融合 RGB 图，两个 RGB 图逐元素求和得到中间 RGB 图，中间 RGB 图给融合解码器块使用，如图 6(c)所示。每层融合解码器块融合的 RGB 图和中间 RGB 图逐元素求和得到新的中间 RGB 图，经过 5 层融合解码器块得到最终分辨率的 RGB 图，如式(6)所示：

$$G_{\mathrm{ToRGB}}=\mathrm{Conv2d}(I_{\mathrm{channel}},3) \tag{5}$$

$$G_{\mathrm{RGB},N}=\mathrm{UpSample}(G_{\mathrm{RGB},N-1})+\mathrm{ToRGB}(\boldsymbol{F}'_{\mathrm{V}}) \tag{6}$$

式中：UpSample 表示上采样操作，这里使用双线性插值作为上采样方式；Conv2d 表示二维卷积操作；I_{channel} 表示输入通道；$\boldsymbol{F}'_{\mathrm{V}}$是经过 CBAM 模块生成的特征图，该特征图既用于新的 RGB 图生成，又作为下一层解码块的输入；N 表示当前解码器块的层数。

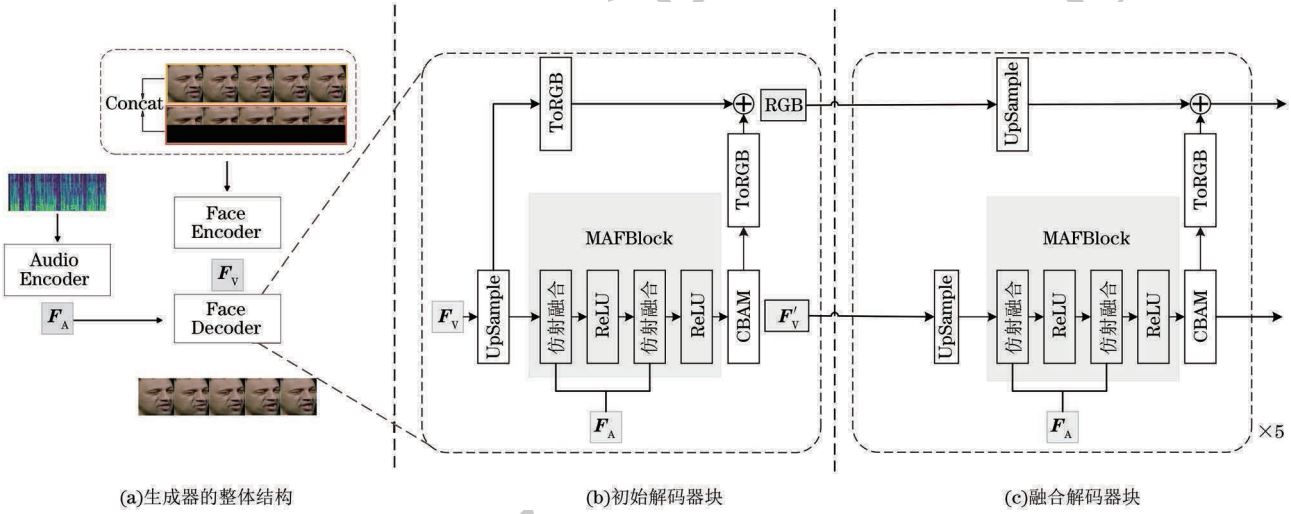

图 6 基于仿射融合的生成器结构

Fig.6 Generator structure based on affine fusion

对于给定的特征图 $\boldsymbol{X}\in\mathbb{R}^{B\times C\times H\times W}$，$\boldsymbol{X}$ 表示的是经过 MAFBlock 后的特征图，通过 CBAM 模块中的通道注意力模块对融合特征的通道进行自适应调整，使用空间注意力模块提高模型对于唇部区域的关注。

CBAM 将通道注意力和空间注意力进行结合，特征图先经过通道注意力得到新的特征图后再经过空间注意力得到最终的特征图。

经过通道注意力的特征图如式(7)表示：

$$\boldsymbol{X}_{\mathrm{C}}=\boldsymbol{X}\times M_{\mathrm{C}}(\boldsymbol{X}) \tag{7}$$

式中：$M_{\mathrm{C}}(\boldsymbol{X})$表示的是通道注意力的计算。

为了保证注意力权重在 0 和 1 之间，使用 Sigmoid 激活函数，如式(10)所示：

$$\boldsymbol{F}_{\mathrm{Avg}}=W_1(W_0(\mathrm{Avg}(\boldsymbol{X}))) \tag{8}$$

$$\boldsymbol{F}_{\mathrm{Max}}=W_1(W_0(\mathrm{Max}(\boldsymbol{X}))) \tag{9}$$

$$M_{\mathrm{C}}(\boldsymbol{X})=\sigma(\boldsymbol{F}_{\mathrm{Avg}}+\boldsymbol{F}_{\mathrm{Max}}) \tag{10}$$

式中：Avg 表示平均池化层，Max 表示最大池化层，通过基于宽度和高度的全局平均池化和全局最大池化得到两个 $1\times1\times C$ 的向量，最大池化层和平均池化层后的特征共享两个全连接层；$W_0\in\mathbb{R}^{C/r\times C}$ 和 $W_1\in\mathbb{R}^{C\times C/r}$ 表示全连接层；r 表示缩放率，设置为 8；C 表示通道数；σ 表示 Sigmoid 激活函数。

经过 $M_{\mathrm{C}}(\boldsymbol{X})$得到注意力的权重，将得到的注意力权重与原始特征图的每个通道相乘，得到通道特征图，输入空间注意力层，将特征图基于通道维度分别经过最大池化和全局池化，得到两个 $H\times W\times1$ 的特征图，将两个特征图在通道维上拼接得到 $H\times W\times2$ 的特征图，经过 7×7 的卷积降维成 $H\times W\times1$ 的特征图，最后使用 Sigmoid 激活函数控制注意力权重范围，如式(11)所示：

$$M_{\mathrm{S}}(\boldsymbol{X})=\sigma(f^{7\times7}(\mathrm{Concat}[\mathrm{Avg}(\boldsymbol{X}),\mathrm{Max}(\boldsymbol{X})])) \tag{11}$$

式中：$f^{7\times7}$ 表示 7×7 的卷积；Concat 表示对特征进行拼接操作；Avg 和 Max 分别表示平均池化和最大池化。

$M_{\mathrm{S}}(\boldsymbol{X})$计算出空间注意力权重，将其与输入特

征图相乘获得最终输出，如式(12)所示：

$$\boldsymbol{F}'_{\mathrm{V}}=\boldsymbol{X}_{\mathrm{C}}\times M_{\mathrm{S}}(\boldsymbol{X}) \tag{12}$$

2.4 视觉质量判别器和唇形同步判别器

视觉质量判别器如图 7 所示，判别器接收真实图片和虚假图片，它的任务是将两者进行区分，用来辅助生成模型生成更加逼真的图片。本文采用的判别器由多个卷积块组成，每个卷积块由卷积层和 LeakyReLU 激活函数组成，通过下采样处理图片，最终获得置信度分数。在训练初期，往往判别器过于强大会导致训练出现模式坍塌的问题，所以在训练中做出了一些改进[10]，针对判别器的每一层，将原有的批归一化去除，改为权重谱归一化，更改过后相较于原来训练更稳定，没有出现模式坍塌的问题。判别器的损失为生成样本被判别为假的概率与真实样本被判别为真的概率之和，如式(13)所示：

$$L_{\mathrm{disc}}=E_{x\sim L_{\mathrm{G}}}[\log_a(\mathrm{D}(x))]+E_{x\sim L_{\mathrm{g}}}[\log_a(1-\mathrm{D}(x))] \tag{13}$$

式中：D 表示判别器。

除了视觉质量判别器以外，为了适应本文任务，还需要额外的约束条件，利用 SyncNet[25] 中提到的唇形-音频同步模型结构作为唇形判别器的基本结构。如图 7 所示，唇形判别器由两个 CNN 编码器组成：一个是人脸编码器；另一个是音频编码器。

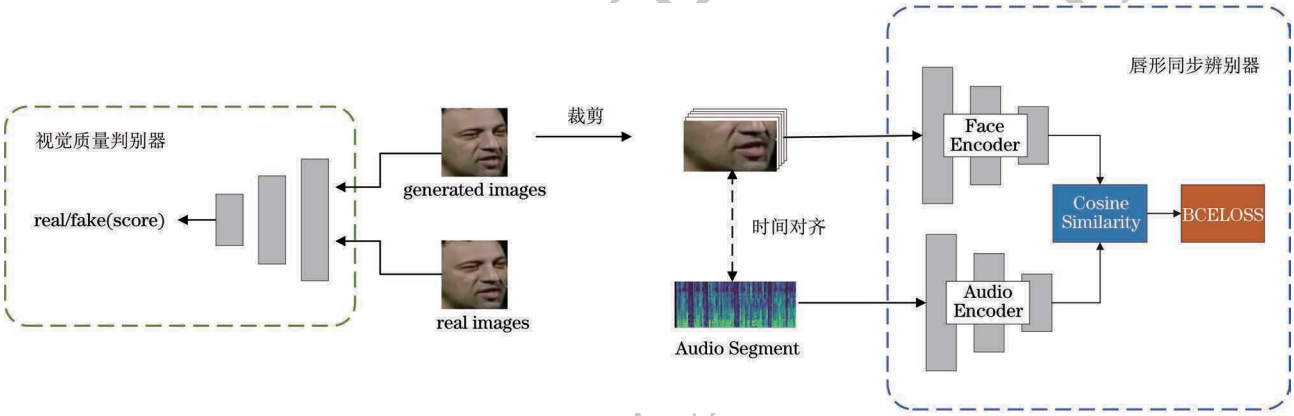

图 7 视觉质量判别器和唇形同步判别器

Fig. 7 Visual quality discriminator and lip synchronization discriminator

人脸编码器的输入是连续 T 帧的人脸，音频输入是与之对应时间的音频梅尔频谱图，将人脸的下半部分截取出来作为人脸编码器的输入，以减少无效信息对同步判别的干扰，最终的输入为(B，$H/2$，W，$3\cdot T$)。音频编码器的输入为与时间对齐的音频梅尔频谱图特征。本文还使用具有余弦相似度 D_{sync} 的二元交叉熵损失 L_{sync} 作为唇形同步的损失函数代替原本的对比损失，如式(14)、式(15)所示：

$$D_{\mathrm{sync}}=\frac{\boldsymbol{F}_{\mathrm{V}}\cdot\boldsymbol{F}_{\mathrm{A}}}{\max(\|\boldsymbol{F}_{\mathrm{V}}\|_2\cdot\|\boldsymbol{F}_{\mathrm{A}}\|_2,\varepsilon)} \tag{14}$$

$$L_{\mathrm{sync}}=-\frac{1}{N}\sum_{i=1}^{N}\log_a(D^i_{\mathrm{sync}}) \tag{15}$$

唇形辨别器的损失和重建损失进行加权求和构成了生成器的损失，权重 α 的设置会在实验部分进行说明，生成器损失如式(16)所示：

$$L_{\mathrm{Gen}}=(1-\alpha)L_{\mathrm{recon}}+\alpha L_{\mathrm{sync}} \tag{16}$$

唇形同步判别器在使用时分为两步：第一步是在整个数据集上单独训练，将损失函数训练到一定程度的时候停止训练；第二步是将训练好的唇形同步判别器用在生成器中，并且在此时唇形同步判别器会冻结所有权重。这样预训练好的唇形同步判别器会使生成器生成逼真的唇形同步人脸。

3 实验

3.1 数据集

本文在多个大型唇读数据集上验证了模型的有效性，包括 LRS3、LRS2、LRW 和 HDTF 数据集，LRS3 数据集包含 400 多小时来自 5 594 场 TED 和 TEDx 英语演讲的视频，所有视频均从 YouTube 下载，且只包含说话人脸部轨迹，视频格式为 224×224 像素、25 帧/s 的 MP4，音轨为单通道 16 位、16 kHz 格式，由于其中有些视频过于模糊，人脸检测模型检测时存在大量人脸丢失情况，为了提高训练稳定性，针对人脸检测模型在人脸丢失较多的视频中进行了筛选，视频由连续的图片构成，图片按连续顺序命名以此来和音频对齐，对于模糊的视频，根据人脸图片数量与视频帧数的比较，剔除了丢失图片数量超过 10%的视频。LRS2 数据集中的视频片段主要源自英国广播公司(BBC)的电视节目，涵盖了各种场景与说话者。视频帧率为 25 帧/s，音频采样率为 16 kHz，相比于 LRS3，数据总时长更短。LRW 是分辨率更低的数据集，除分辨率只有 112×

112 像素以外，其余规格和 LRS2 类似。

LRS3 包含来自各个国家的演讲者，内容涵盖更广泛的话题，并且各个演讲者有不同的口音和表达方式，而 LRS2 和 LRW 数据集包含 BBC 的节目，通常是在正式场合的表达，整体来说缺乏变化，更加容易训练，根据实验也能看出在 LRS3 数据集上训练的模型具有更好的泛化性，因此最终本文所有实验只在 LRS3 进行训练。

除了这 3 个数据集以外，本文还在 HDTF 数据集上进行评估，HDTF 是一个更高质量的唇读数据集，其中包括了 720P 甚至更高分辨率的数据集，每个视频中的说话人面部清晰可见，且背景相对干净。

3.2 实验设置

实验所使用的操作系统为 Ubuntu 20.04，GPU 配置为 NVIDIA RTX 4090(24 GB)，使用 PyTorch 深度学习框架构建网络模型并训练。第一阶段唇形同步判别器训练时设置 Batch 大小为 64，使用 Adam 优化器训练模型，Adam 中 β_1 设置为 0.5，β_2 设置为 0.99，学习率设置为 5×10^{-5}，在训练周期内通过余弦退火的方法逐步衰减至 1×10^{-6}，共训练了 40 个 Epoch。第二阶段生成模型训练 Batch 大小设置为 32，用 Adam 优化器训练模型，Adam 中 β_1 设置为 0.5，β_2 设置为 0.999，初始学习率为 1×10^{-4}，通过余弦退火的方法逐步衰减至 5×10^{-5}，训练过程中将第一阶段中预训练完成的唇形判别器损失的初始权重 α 设置为 0，在后续训练时如果唇形判别器损失降低至 0.75 就将权重 α 上升到 0.03。使用 Adam 优化器[26]来优化模型，其中，β_1 设置为 0.5，β_2 设置为 0.99。在模型的输入处理中，选择了连续的 5 帧图像作为目标图像。为了避免模型在生成过程中对固定的参考图像产生依赖，引入了一种随机化策略，即在每次输入时随机选取 5 帧图像作为模型的参考图像。通过这种方法，模型能够在不同的上下文中学习，具有更好的泛化性能，从而提升生成效果的多样性与鲁棒性。

为了确保唇形辨别器在生成任务中能够准确发挥作用，首先对其进行了充分的预训练，使其能够有效判别唇形和音频是否同步。在随后的面部生成过程中，将唇形辨别器的权重进行冻结，避免其参与进一步训练。这一策略有助于保持唇形辨别器的稳定性，使其在生成过程中持续提供高质量的唇形同步性评估，进而提高生成图像唇形同步的准确性。

为了验证视觉质量判别器是否会影响唇形同步判别器在唇部拟合方面的表现，对包含视觉质量判别器和不包含视觉质量判别器的模型进行了独立训练。对比两者性能，评估视觉质量判别器是否在提升图像质量的同时抑制唇形同步的准确性。

对于包含视觉质量判别器的模型，由于其需要在生成过程中对图像的视觉质量进行额外的判别与优化，拟合难度增加，因此在训练过程中设置更多的迭代次数。包含视觉质量判别器的模型共训练了 200 个 Epoch，以确保其能够充分学习并提高图像的视觉质量。对于不包含视觉质量判别器的模型，其训练过程相对简化，设置为 100 个 Epoch。

3.3 评估指标

实验评估方向是唇形同步和视觉质量。在唇形同步方面，使用 LSE-D 和 LSE-C 两个指标。LSE-D：音频和视频之间的距离得分，表示唇形和音频特征之间的距离的平均误差，该指标越低表示视频和音频越匹配。LSE-C：置信度得分，该指标越高表示唇形与音频同步率越高。通过加载预训练的 SyncNet 模型来计算两种指标。在视觉质量方面，使用结构相似性(SSIM)作为评估指标，SSIM 可以衡量图片的失真程度，主要考量图片的亮度、对比度和结构，相对于均方误差(MSE)更符合人眼的直观感受。峰值信噪比(PSNR)通过计算两张图像之间的信噪比来衡量图像质量，反映的是像素级的误差。SSIM 和 PSNR 越高表明生成图像的整体效果越好。

3.4 实验效果对比

在 LRS3 数据集上针对两种不同的训练方式进行了比较，训练设置和上文所述保持一致，具体结果如表 1 所示，其中 w/o visual disc 表示去除了视觉质量判别器。

表 1 在 LRS3 数据集上不同训练方式对比

Table 1 Comparison of different training methods on the LRS3 dataset

Model	LSE-C	LSE-D	SSIM	PSNR/dB
Ours	7.963	6.259	0.916	32.772
w/o visual disc	8.012	6.153	0.871	29.824
Wav2Lip[1]	7.497	6.693	0.921	32.221

在表 1 中可以看出，缺乏视觉质量判别器的情况下，模型的唇形同步指标有略微提升，但视觉指标出现显著下降，当模型只包含唇形同步判别器时，模型的训练目标更加集中，主要关注音频与唇形的同步性。这使得模型的学习重点完全放在如何更好地匹配音频和唇形动作上，而不必分散资源去考虑图像的整体视觉质量，然而唇形同步率的提升并未在生成图像中带来显著的视觉质量改善。如图 8 所示，尽管唇形同步的客观指标有所

下降，但从主观感受上看，整体唇形动作的趋势变化不大。然而，生成图像的下半部分质量却显著不同，尤其在细节处理上。如图 9 所示，缺乏视觉质量判别器的模型生成的图像存在明显的缺陷，唇部区域显得过于平滑，缺少细节处理。此外，生成的嘴唇中出现了不自然的牙齿，显得突兀并缺乏真实感。这些问题表明，视觉质量判别器对于图片生成质量有显著的影响。

图 8 两种训练方式生成图像对比

Fig. 8 Comparison of images generated by two training methods

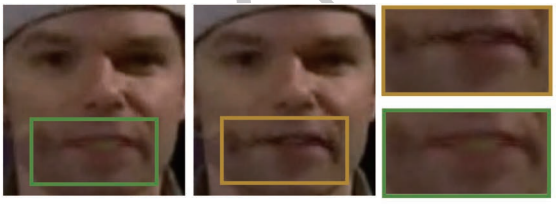

图 9 唇部细节对比

Fig. 9 Comparison of lip details

视觉质量判别器能够在仅对唇形同步率产生微小影响的情况下显著提升图像的整体质量。因此，在后续实验中，决定保留包含视觉质量判别器的模型，并将其作为最终模型进行测试与评估。

使用最终模型在不同情形下生成说话人脸视频并展示结果，图 10 展示了模型在不同测试集上的生成效果。在每个示例中从左至右依次为模型输入中的遮挡人脸、模型的输出以及真实图片。图像包括了在侧脸和嘴部有复杂细节等不同场景下的生成效果。可以看出，模型在各类场景中均具有良好的视觉效果，并准确呈现对应的唇形。

此外，为了验证模型的泛化能力，选取来自其他来源的视频，生成对应的说话人脸视频，并展示这些视频中唇形对应的单词，分别包含中文和英文，其中单词中具体发音部分使用下划线表示，如图 11 所示，其中音频采用其他视频中提取的音频。结果表明，唇形运动基本符合音频内容，生成视频分辨率与

图 10 说话人脸生成图片

Fig. 10 Speaking face generation images

模型训练数据集分辨率一致。如何生成高质量同时保持高唇形同步率的视频也是后续工作将继续改进和提升的部分。

图 11 单词发音对应的人脸生成图像

Fig. 11 Face images generated from word pronunciations

将最终模型与多个相关模型在多个数据集上进行比较评估。为了测量模型的泛化能力，在 LRS3 数据集上进行训练，并在其他数据集上与相关模型在 LRW 和 LRS2 数据集上进行比较，如表 2 所示，其中，最优指标值用加粗字体标示，“—”表示原文献中没有该指标值，下同。可以看出，本文模型均在唇形同步指标 LSE-C 和 LSE-D 上取得了最好的结果，同时在 SSIM[27] 和 PSNR 指标上也均取得了较好的结果，说明 MAFBlock 在保持视觉质量的前提下有效提升了唇形同步率，与真实视频的指标非常接近，表明模型生成的效果与真实视频之间的差异较小。

表 2 不同模型在 LRW、LRS2 数据集上的性能对比

Table 2 Performance comparison of different models on LRW and LRS2 datasets

Model	LRW				LRS2			
	LES-C↑	LSE-D↓	SSIM↑	PSNR/dB↑	LES-C↑	LSE-D↓	SSIM↑	PSNR/dB↑
Real Videos	6.876	6.968	**1.000**	—	8.247	6.259	**1.000**	—
Wav2Lip[1]	7.237	6.617	0.875	32.147	7.789	6.386	0.925	32.195
LipGAN[2]	3.350	10.050	—	—	3.199	10.33	—	—
PC-AVS[23]	6.420	7.344	0.778	30.440	6.928	7.101	0.784	30.683
SyncTalkFace[28]	6.619	7.013	0.893	**33.126**	7.925	6.352	0.876	32.529
ATVG[29]	5.735	7.664	0.781	31.409	5.584	8.223	0.735	30.427
Ours	**7.784**	**6.432**	0.901	33.012	**8.128**	**6.112**	0.916	**32.585**

此外，由于 HDTF 数据集本身具有较高的视频清晰度，因此在 HDTF 数据集上单独评估了模型的性能，如表 3 所示。

表 3 不同模型在 HDTF 数据集上性能对比

Table 3 Performance comparison of different models on the HDTF dataset

Model	LES-C	LSE-D	SSIM	PSNR/dB
Real Videos	8.993	6.498	1.000	—
Wav2Lip[1]	8.290	6.871	0.907	29.287
PC-AVS[23]	8.135	6.613	0.638	23.630
ATVG[29]	7.061	8.422	0.907	25.543
DINet[30]	6.841	8.377	**0.942**	**30.008**
Ours	**8.314**	**6.512**	0.901	29.012

从表 3 中可以看出，本文模型设计并非专门针对高清数据集，但在高清数据集上仍保持了良好的唇形同步率。然而，SSIM 和 PSNR 指标主要评估两幅图像的差异性，并未反映图像的分辨率。由于模型在 LRS3 数据集上训练，而 HDTF 数据集的人脸细节更为清晰，因此模型生成的人脸在整体视频中显得较为模糊，拼接到原图中可能显得突兀。相比之下，DINet 和 ATVG 等模型是专门为高清训练集设计的，因此生成的图像清晰度较高。然而，从表 3 中也可以看出，尽管 DINet 生成的图像清晰度更高，且更接近真实图像，但其唇形同步指标较低，未能兼顾唇形同步和视频质量。因此，如何平衡这两者将是后续研究的重点。

3.5 消融实验

为进一步验证本文提出模块的有效性，通过消融实验评估了模型的每个组成部分。在 LRS2 和 LRS3 数据集上进行了 3 种消融实验。首先，去掉模型的 CBAM 模块，具体而言，在本文模型结构中去掉每个 MAFBlock 后的通道注意力模块和空间注意力模块，如表 4 所示模型的整体性能有略微下降，这主要是 CBAM 模块能够捕捉特征图的关键部分，但整体属于轻量级，添加 CBAM 模块后视频生成的速率平均每 10 s 的差距在 0.5 s 之内，能够以较低的代价提升模型性能。

表 4 各模块的消融实验结果

Table 4 Results of ablation experiments for each module

Model	LRS3				LRS2			
	LES-C	LSE-D	SSIM	PSNR/dB	LES-C	LSE-D	SSIM	PSNR/dB
w/o CBAM	7.868	6.368	0.914	31.857	7.961	6.223	0.915	31.818
w/o MAFBlock	7.621	6.731	0.915	31.752	7.783	6.521	0.913	31.782
w/o lip disc	5.931	8.172	**0.941**	**33.712**	6.012	7.931	**0.932**	**33.421**
Ours	**7.963**	**6.259**	0.916	32.772	**8.128**	**6.112**	0.916	32.585

然后，移除了 MAFBlock 模块，将两个编码器生成的特征直接拼接后输入解码器中以生成图像，这种方式类似于 Wav2Lip 中的方法。结果显示，如表 4 所示，模型的唇形同步性能显著降低。这表明 MAFBlock 模块在深度融合音频和视频特征方面起到了至关重要的作用。具体来说，MAFBlock 中的仿射融合机制通过音频生成的缩放值和偏移量能够动态调整视频特征图中每个通道的激活值。这种动态调整相当于为视频特征图引入了与音频信息相关的注意力机制，使得视频特征中的关键信息能够更好地与音频信号进行对齐。因此，MAFBlock 不仅提升了模型对多模态信息的理解能力，而且还有效增强了唇形同步的准确性。这一实验结果进一步验证了 MAFBlock 模块在保持唇形同步和提升生成

视频的准确性方面的有效性，使得模型能够在多样化的场景中生成更自然、更协调的图像。

最后，移除了唇形同步判别器。如表 4 所示，虽然模型图片的生成质量出现了一定的提高，但模型的唇形同步指标出现了显著下降。这一结果表明唇形同步判别器在约束唇部运动方面起到了关键作用。具体来说，当移除唇形同步判别器后，模型的唇形同步损失在训练过程中会下降到约 0.65 时出现较大波动，难以进一步优化。当唇形同步判别器存在时，尽管其权重仅设置为 0.03，却能够有效地引导模型训练，使得唇形同步损失逐渐降低，并最终稳定于约 0.1 的较低水平。唇形同步判别器虽然占比小，却对模型在唇形同步上的表现具有显著影响。不仅为模型提供了有效的监督信号，还确保了生成视频中的唇部运动与音频信息高度一致。

4 结束语

当前的说话人脸视频生成研究在深度融合音频与视频信息方面存在一定的局限性。为解决这一问题，本文提出一种基于仿射变换、通道注意力和空间注意力的 GAN，将其用于生成更加同步的说话人脸视频。通过仿射变换融合模块，音频特征直接作用于视频特征的通道维度，模型在解码器的每个层级中利用仿射变换来捕捉音频与视频之间的关联，克服了现有模型在音频信息利用方面的不足，同时通道注意力和空间注意力机制进一步引导其关注人脸的关键区域，特别是唇部，从而实现更精确的唇形同步。

与其他相关模型相比，本文模型在唇形同步率方面表现出显著提升，然而这种提升也伴随着生成速度的下降，而视频的视觉质量未显著提高。因此，未来的工作将着重设计和优化模型，使其能够在更高分辨率的视频上进行训练，以满足实际应用中对视频分辨率的要求，同时确保唇形同步性能不受影响。

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. New York, USA: ACM Press, 2020: 484-492.

[2] PRAJWAL K R, MUKHOPADHYAY R, PHILIP J, et al. Towards automatic face-to-face translation[C]//Proceedings of the 27th ACM International Conference on Multimedia. New York, USA: ACM Press, 2019: 1428-1436.

[3] MITTAL G, WANG B Y. Animating face using disentangled audio representations[C]//Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). Washington D.C., USA: IEEE Press, 2020: 3290-3298.

[4] AFOURAS T, CHUNG J S, ZISSERMAN A. LRS3-TED: a large-scale dataset for visual speech recognition[EB/OL]. [2024-05-11]. https://arxiv.org/abs/1809.00496.

[5] AFOURAS T, CHUNG J S, SENIOR A, et al. Deep audio-visual speech recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 8717-8727.

[6] KARRAS T, LAINE S, AILA T M. A style-based generator architecture for generative adversarial networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2019: 4401-4410.

[7] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[8] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]//Proceedings of the International Conference on Machine Learning. [S. l.]: PMLR, 2017: 214-223.

[9] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of Wasserstein GANs[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2017: 5769-5779.

[10] MIYATO T, KATAOKA T, KOYAMA M, et al. Spectral normalization for generative adversarial networks[EB/OL]. [2024-05-11]. https://arxiv.org/abs/1802.05957.

[11] 张慧妍, 梁勇, 兰景宏, 等. 基于记忆模块与过滤式生成对抗网络的入侵检测方法[J]. 计算机工程, 2024, 50(6): 197-207.
ZHANG H Y, LIANG Y, LAN J H, et al. Intrusion detection method based on memory module and filtered generative adversarial network[J]. Computer Engineering, 2024, 50(6): 197-207. (in Chinese)

[12] KARRAS T, LAINE S, AITTALA M, et al. Analyzing and improving the image quality of StyleGAN[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2020: 8110-8119.

[13] PATASHNIK O, WU Z Z, SHECHTMAN E, et al. StyleCLIP: text-driven manipulation of StyleGAN imagery[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Washington D.C., USA: IEEE Press, 2021: 2065-2074.

[14] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the International Conference on Machine Learning. [S. l.]: PMLR, 2021: 8748-8763.

[15] SUWAJANAKORN S, SEITZ S M, KEMELMACHER-SHLIZERMAN I. Synthesizing Obama[J]. ACM Transactions on Graphics, 2017, 36(4): 1-13.

[16] GUO Y D, CHEN K Y, LIANG S, et al. AD-NeRF: audio driven neural radiance fields for talking head synthesis[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Washington D.C., USA: IEEE Press, 2021: 5764-5774.

[17] YE Z, JIANG Z, REN Y, et al. GeneFace: generalized and high-fidelity audio-driven 3D talking face synthesis[EB/OL]. [2024-05-11]. https://arxiv.org/abs/2301.13430.

[18] LAHIRI A, KWATRA V, FRUEH C, et al. LipSync3D: data-efficient learning of personalized 3D talking faces from video using pose and lighting normalization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2021: 2755-2764.

[19] FRIED O, TEWARI A, ZOLLHÖFER M, et al. Text-

based editing of talking-head video[J]. ACM Transactions on Graphics, 2019, 38(4): 1-14.
[20] THIES J, ELGHARIB M, TEWARI A, et al. Neural voice puppetry: audio-driven facial reenactment[C]//Proceedings of the 16th European Conference on Computer Vision. Berlin, Germany: Springer International Publishing, 2020: 716-731.
[21] CHUNG J S, JAMALUDIN A, ZISSERMAN A. You said that?[EB/OL]. [2024-05-11]. https://ludwig.guru/s/you+said+that.
[22] ZHOU H, LIU Y, LIU Z W, et al. Talking face generation by adversarially disentangled audio-visual representation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2019: 9299-9306.
[23] ZHOU H, SUN Y S, 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 (CVPR). Washington D.C., USA: IEEE Press, 2021: 4176-4186.
[24] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV). Berlin, Germany: Springer International Publishing, 2018: 3-19.
[25] CHUNG J S, ZISSERMAN A. Out of time: automated lip sync in the wild[C]//Proceedings of ACCV'16. Berlin, Germany: Springer International Publishing, 2016: 251-263.
[26] KINGMA D P, BA J. Adam: a method for stochastic optimization[EB/OL]. [2024-05-11]. https://arxiv.org/abs/1412.6980.
[27] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.
[28] PARK S J, KIM M, HONG J, et al. SyncTalkFace: talking face generation with precise lip-syncing via audio-lip memory[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2022: 2062-2070.
[29] CHEN L L, MADDOX R K, DUAN Z Y, et al. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D.C., USA: IEEE Press, 2019: 7824-7833.
[30] ZHANG Z M, HU Z P, DENG W J, et al. DINet: deformation inpainting network for realistic face visually dubbing on high resolution video[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2023: 3543-3551.

文字编辑 陆燕菲
栏目编辑 宋 圆