arXiv:2305.05445v1 [cs.CV] 9 May 2023

# StyleSync：基于 Style 的生成器实现高保真通用与个性化唇形同步

Jiazhi Guan[1], [2*] Zhanwang Zhang[1*] HangZhou[1†] Tianshu Hu[1†] KaisiyuanWang[3]
DongliangHe[1] HaochengFeng[1] Jingtuo Liu[1] Errui Ding[1] Ziwei Liu[4] Jingdong Wang[1]
[1]百度公司计算机视觉技术部（VIS）[2]清华大学[3]悉尼大学[4]S-Lab，南洋理工大学
guanjz20@mails.tsinghua.edu.cn {zhangzhanwang,zhouhang09,hutianshu01}@baidu.com




图 1. **由我们的 StyleSync框架生成的个性化唇形同步结果**。我们的方法不仅支持根据条件音频对任意目标模板视频进行高保真修改，还能通过个性化优化进一步适配特定风格。在此图中，我们的唇形同步结果应与条件音频对应的唇形同步视频具有相同的嘴型。

## 摘要

尽管近期在将唇部运动与任意音频波形同步方面取得了进展，现有方法仍难以兼顾生成质量与模型的泛化能力。以往研究要么需要长期数据进行训练，要么会为所有人物生成相似且质量较低的运动模式。本文提出 *StyleSync*，这是一种能够实现高保真唇同步的有效框架。我们发现，基于风格的生成器能够充分实现这一出色特性，适用于单样本以及少样本场景。具体而言，我们设计了一种掩码引导的空间信息编码模块，用于保留给定人脸的细节。通过调制卷积，音频能够准确地修改嘴型。此外，我们的设计还通过引入风格空间，并仅在有限帧上进行生成器细化，实现个性化唇形同步。因此，能够准确保留目标人物的身份和说话风格。大量实验表明，我们的方法能够在各种场景中生成高保真结果。资源可在 *https://hangz-nju-cuhk.github.io/projects/StyleSync* 获取。

*贡献相同。
†通讯作者。

## 1. 引言

根据条件音频生成唇形同步视频，对数字人创作、音频配音、电影制作和娱乐等领域至关重要。近年来，该领域发展迅速，但大多数方法 [6,8,9,16,20,25,28,29,40,41,49,54,59,62,64–66] 都专注于生成完整的动态说话人头部。在这种设定下生成的结果很难融入现有场景。在音频配音等现实场景中，一个关键需求是在不改变场景其他部分的前提下，无缝修改嘴部或面部区域，因此这些方法并不适用。

以往方法通过两种不同途径实现无缝的嘴部修改。一些研究 [40,46]致力于在针对特定人物的场景中取得逼真效果，这需要使用长时段视频片段对目标进行建模。此外，这些方法依赖先验的 3D 面部结构信息。3D 拟合过程中累积的不确定性和误差会极大影响其性能。另一方面，人们希望构建能够突破数据限制、适用于更广泛场景的模型。因此，少数方法 [32, 33, 43] 设计了不依赖 3D 或 2D 结构先验的人物无关模型。然而，这种设定极具挑战性。

为了在任意长度的视频中生成高保真唇形同步结果，需要应对两个关键挑战。**1)** 如何高效设计强大的生成式骨干网络，同时支持准确表达音频信息和无缝的局部替换。直观而言，唇形同步质量与保留原始目标帧信息 [33, 65]天然存在矛盾。**2)**如何有效利用尽可能多的已提供信息，并将个性化特征融入通用模型。尽管少样本元学习已被证明有助于生成会说话的人脸 [5, 7, 61, 66],，但如何将这种能力融入唇形同步流程尚未得到探索。

在本文中，我们提出了一种高度简洁且全面的框架，称为**StyleSync**，能够在通用和个性化场景下生成高保真的唇形同步结果。其关键在于我们对基于风格的生成器进行简单但面向唇形同步的修改。尽管基于风格的生成器 [23, 24] 已被应用于各种说话头部生成方法 [5, 60, 65], , 但这些方法取得的成果仅具有部分借鉴意义。它们旨在生成完整的头部，这会导致背景不稳定和失真，而这些问题在我们的场景中是不可接受的。

通过重新审视基于风格的生成器的细节，我们发现了一些简单但至关重要的修改，使我们的框架适用于唇形同步。与上述方法不同，我们采用掩码嘴部建模方案 [33, 43] ，并精心设计了一种基于掩码的 *Spa-*信息编码策略，其中目标帧和参考帧的信息均根据不同的掩码方案编码到生成器的噪声空间 [56, 57] 中。而音频动态和高层参考帧的信息则以与 [26, 65] 类似的方式注入风格调制卷积。这样，我们的方法既能受益于基于风格的生成器强大的生成能力，也保留了易于实现、训练速度快的优势。

此外，我们的网络修改能够保留个性化信息（例如说话风格以及嘴部和下颌的细节）。我们受到近期反演 StyleGAN 先验研究的启发 [1, 2, 36, 47]，并提出了一种个性化优化方案。由于音频配音通常是在说话视频上进行的，我们的模型只需利用几秒钟的个人信息，就可以优化额外的特定人物参数，包括 $W^+$ 和生成器。大量实验表明，在单样本设置下，我们的框架以较大优势明显优于以往最先进的方法，而针对目标的优化进一步提升了结果的保真度。

我们的贡献可概括如下：**1)**我们提出了 **StyleSync** 框架，通过对基于风格的生成器进行简单而有效的改进，包括基于掩码的空间信息编码。**2)** 我们提出了个性化优化流程，将少样本人物专属优化融入我们的框架。**3)** 大量实验表明，我们的框架能够直接生成准确且高保真的单样本唇形同步结果。此外，我们提出的个性化优化进一步提升了生成质量。我们的方法明显优于以往方法。

## 2. 相关工作

### 2.1. 音频驱动的面部动画

音频驱动的面部动画长期以来一直是计算机视觉和图形学领域的研究课题。相关研究针对数字 3D 人脸 [15, 21, 34, 67] 以及逼真人头 [14, 18, 44]展开。我们关注现实世界中的人头。

**说话头部生成**。大多数关于唇形同步的研究，旨在生成说话者的完整头部 [6, 8,9、 16–19、 26、27、 37、 42、44、 48、 50、 58、 59、 63–66]。具体而言，许多研究利用二维 [9]关键点、三维关键点 [66] 和三维网格 [6] 等结构信息。这类表示的不确定性和不准确性会导致说话头部生成过程中出现误差累积。大多数个体专属建模方法 [20,28,44] 依赖这些信息，因此生成结果的

泛化能力或唇形同步质量较差。最近，NeRF 也被用于个体专属建模 [16, 27, 37, 45], , 但在数据有限的情况下，其表现同样不佳。

**面部唇形同步**。 另一类研究聚焦于对视频中的嘴部进行唇形同步，同时保持其他信息不变 [32, 33, 41, 46]。我们的工作属于同一研究范畴。尽管 Thies 等人 [46]和 Song 等人 [40]能够生成逼真的结果，但他们依赖于基于特定人物的训练，且需要超过 2 分钟的数据。

具体而言，Wav2Lip [33]能够生成与人物无关且高度准确的唇形同步结果。然而，他们基于低分辨率图像构建生成模型，导致结果模糊。我们发现，如果需要恢复的图像质量信息较少，模型将更容易学习嘴部运动。此外，他们的方法无法捕捉给定目标模板视频中的个性化模式。

### 2.2. 基于风格的面部生成器

StyleGAN 模型 [22–24] 在图像生成任务中取得了巨大成功，尤其是在面部图像生成和编辑方面 [1, 2, 35, 38, 39, 47]。

**StyleGAN反演**。 StyleGAN反演是指利用预训练的StyleGAN恢复给定图像潜在空间的过程。Abdal 等人 [1, 2] 提出，不仅使用风格空间 $W$对图像进行反演，还将其扩展至 $\mathcal{W}^+$，从而更好地保留细节。近期研究提出学习编码器 [3, 35, 47] ，以实现更好的可编辑性和更快的优化效果。此外，近期研究还尝试通过关键点微调 [36]来优化生成器的参数。在我们的工作中，我们借鉴了以往StyleGAN反演研究的思路，并提出了一种个性化优化过程。我们按照 [47] 从参考帧中编码出一个 $\mathcal{W}^+$ 空间，从而实现个性化唇部运动的学习。

**应用**。 StyleGAN 架构也被用于人脸修复 [52,57] 和人脸交换 [55,56]。这些任务需要保留原始的面部情绪和表情。另一方面，基于风格的生成器也被用于创建面部动画 [5,26,65],，这与我们的任务密切相关。然而，它们依赖于 $W$ 或 $W^+$ 空间中的风格向量来控制外观和运动动态。这种设置的一个主要缺点是， $W^+$ 空间无法轻松确保背景的空间一致性，从而导致不真实的结果或明显的伪影。

## 3. 方法

在本节中，我们将介绍 StyleSync 框架的具体细节，如图 2 所示。该框架通过借鉴以往基于风格的生成器所取得的若干成功成果，以简洁的方式进行设计。我们将在第 3.1 节中介绍对基于风格的生成器所做的改进，这些改进使其成为一个成功的骨干网络。

以及我们在第 3.2 节中泛化模型的训练目标。具体而言，我们将在第 个性化优化 节介绍该流程，该流程通过针对个人进行优化进一步提升了我们的效果。3.3。

**任务表述**。我们的目标是使目标人物的唇部运动与任意提供的音频同步，并将其无缝融合到原始目标视频中。我们采用典型的自重建方式进行训练。训练设置与 Wav2Lip [33]类似。给定目标视频 $\mathbf{V} = \{I_1, \ldots, I_T\}$，包含 $T$帧，我们用掩码遮住包括下颌和脸颊在内的面部下半部分 $M$。目标是根据对应的音频 $\mathbf{a} = \{a_1, \ldots, a_T\}$ (**a** 会被处理为频谱图 )，恢复完整面部。由于掩码目标帧 $I_t^m = (\mathbf{1} - M) * I_t$未提供任何有关嘴部和下颌形状的信息，因此我们在训练期间利用随机参考帧 $I_{ref} \in \mathbf{V}$ 提供所需的上下文。

### 3.1. 修改基于风格的生成器以实现唇同步

**基于风格的生成器概述**。 我们对其进行面向唇同步的修改，以构建一个成功的唇同步骨干网络，并采用 StyleGAN2 [24] 架构。原始 StyleGAN2 接收一个风格向量 $w$ 作为输入，并在训练期间用它调制总计 $L$ 个生成层中的卷积操作。在推理期间，可以在不同层使用不同的 $w$，从而构成 $W^+$ 空间$\mathbf{W} = \{w_1, \ldots, w_L\}$。原始 StyleGAN2 的基本组成部分，包括 $W^+$ 空间以及 StyleGAN 生成器 G 的各层，在图 **2 中以蓝色**标示。为简化起见，省略了某些细节结构。

**基于掩码的空间信息编码**。 我们的目标是在参考图像的辅助下，将同步的嘴部无缝融合到目标帧中。正如第2.2 节所述，以往利用基于风格的生成器的方法并不能直接应用于我们的场景，因此我们寻求一种不同的方式来编码人脸的空间信息。

近期关于人脸修复 [57] 和换脸 [56] 的研究已证明，可以将内容引导的特征图以类似于噪声 的方式附加到 StyleGAN 层，而不会影响 **生成模型的表达能力。这些包含丰富上下文信息的**噪声 $\mathbf{N} = \{N_1, \ldots, N_L\}$ $= \{N_1, \ldots, N_L\}$ 能够很好地保留编码视觉信息的原始空间结构和属性。这一特性为利用参考图像中包含的属性信息提供了一种途径。因此，遵循 Wav2Lip [33], , 我们将经过掩码处理的输入与参考输入拼接起来，作为框架的视觉输入，并对相应的特征图进行编码 $\mathbf{F}$ $_f = \{F_1, \ldots, F_L\}$。最简单的设置是将$\mathbf{N}$ **= 视为**F。

然而，生成唇部运动的目标是创造出不同于原始面部结构的动态效果，同时保留面部身份；这与以保持结构不变为目标的以往方法有着显著区别。

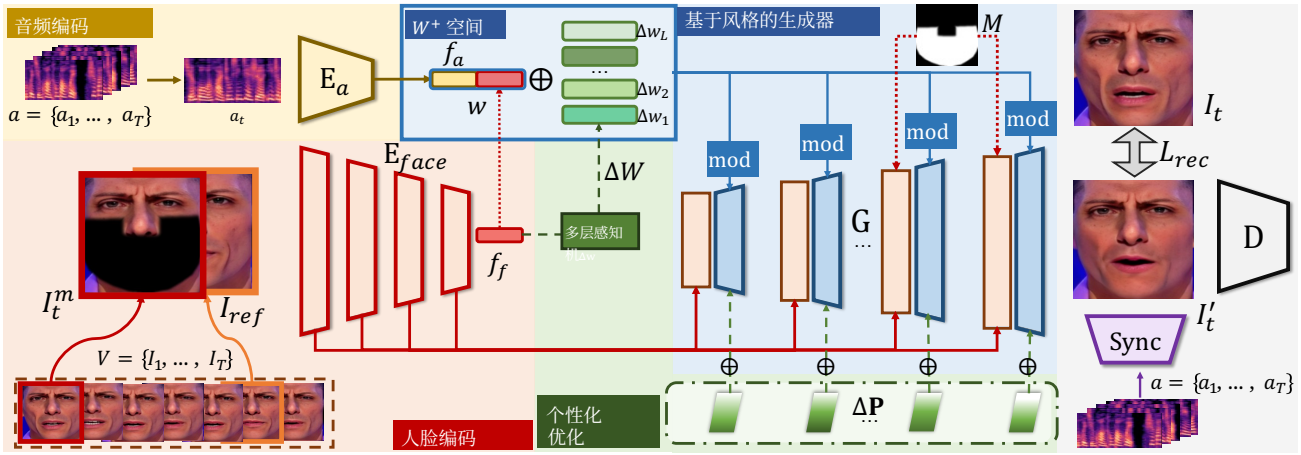


图2. **我们的 StyleSync框架。 蓝色** 的模块表示基于风格的生成器。掩码目标帧 $I_t^m$与参考帧 $I_{ref}$拼接后，由 $E_{face}$ （**红色**）编码为 $f_f$。音频信息被编码为音频特征 $f_a$ （**黄色**）。这些特征拼接在一起，形成风格（$W$）空间。具体而言，我们设计了个性化优化过程（**绿色**），包括学习 $\Delta\mathbf{W}$ 和 $\Delta\mathbf{P}$。这部分在初始骨干网络训练期间不会进行训练。

采用相同方案时，嘴型会受到参考图像 $I_{ref}$ 的显著影响。我们发现，**F** （高分辨率层）的低层信息中融入了超出预期的面部结构信息。因此，我们直观地屏蔽面部特征中的低层信息部分。这里我们简单地定义

$$F'_l = (\mathbf{1} - M) * F_l, \text{for } l > \lfloor \frac{L}{2} \rfloor. \tag{1}$$

面部信息在图**2** 中以红色标示。 **1** 是一个与 $M$大小相同的全一矩阵。

**风格信息编码。** 我们沿用以往研究 [5, 42, 65] 将音频动态和面部信息编码到基于风格的生成器的风格空间（$W$ 空间）中。音频特征 $f_a$ 由编码器 E 编码得到$a$ ，面部特征 $f_f$ 则是面部编码器 $E_{face}$的瓶颈层 $w =$ ，拼接($f_a$， $f_f$)。由于参考帧中的信息已通过空间噪声编码融合到生成器中，因此融合面部特征的必要性较低。实验表明，无论是否使用 $f_f$，差异都很细微，但我们仍遵循常见做法保留这一设计。

**个性化要素。** 我们上述的设计不仅简单，还具备个性化唇形同步建模的外部潜力。详情见第 3.3 节，我们将在此简要讨论使模型更加完善的其他要素。

具体而言，我们的扩展受到 StyleGAN反演领域进展的启发 [1, 2, 36]。相关研究指出，将 $w$ 扩展到 $W^+$ 包含 $\mathbf{W} = \{w + \Delta w_1, \dots, w + \Delta w_L\}$ 的空间对于恢复特定图像至关重要。此外，近期研究 [4, 36] 还探索了微调生成器的参数，以提升网络对特定目标的拟合能力。因此，我们的生成器有望学习有限的参数偏移 $\Delta\mathbf{P}$ ，从而拟合特定人物。

这些个性化要素如图**绿色**所示（图2 ），且在泛化训练过程中不会进行优化。为了使两种设置下的流程保持一致，我们在泛 $\Delta\mathbf{W} = \mathbf{0}$ 和 $\Delta\mathbf{P} = \mathbf{0}$ 化模型中设置。

### 3.2. 骨干网络训练目标

在训练过程中，整个骨干网络以同一时间步的 6 通道拼接结果 $[I_t^m, I_{ref}]$和音频片段 $a_t$作为输入，并预测 $I'_t = \mathrm{G}(\mathbf{N}, \mathbf{W})$，旨在恢复未被掩码的目标图像 $I_t$。为了使设计保持简洁，训练目标大多与 StyleGAN 2 [24] 和 Wav2Lip [33]保持一致。

**重建损失。**像素级重建损失$\mathcal{L}_{rec}$ 是训练我们任务的基础。这里我们采用常用的 $L_1$ 损失和感知损失 [51]。

$$\mathcal{L}_{rec} = \|I'_t - I_t\|_1 + \sum_{m=1}^{N_{vgg}} \|\mathrm{VGG}_m(I'_t) - \mathrm{VGG}_m(I_t)\|_1, \tag{2}$$

其中，$\mathrm{VGG}_m$ 是预训练 VGG19 网络第 $m$ 层的输出。

**对抗损失。**我们直接采用 StyleGAN2 [24] 中相同的判别器 D 进行对抗训练:

$$\mathcal{L}_{adv} = \min_{\mathrm{G}} \max_{\mathrm{D}} (\mathbb{E}_{I_t}[\log \mathrm{D}(I_t)] + \mathbb{E}_{I'_t}[\log(1 - \mathrm{D}(I'_t))]). \tag{3}$$

具体而言，我们使用 StyleGAN2 [24] 中的预训练版本初始化判别器的权重。这种做法能使结果明显更加清晰。

**唇形同步损失**。 此外，我们还训练了 SyncNet [12]，它由视觉编码器 $S_v$ 和音频编码器 $S_a$组成。它通过对比损失判断视觉与音频片段在时间上是否对齐。训练生成器时，我们在每个推理步骤中预测一个批次内连续的 5 帧 $I'_{t:t+4}$，并借助 SyncNet 对训练进行监督。目标函数为：

$$\mathcal{L}_{sync} = -\frac{S_v(I'_{t:t+4})^{\mathbf{T}} \cdot S_a(a_{t:t+4})}{\|S_v(I'_{t:t+4})\|_2 \|S_a(a_{t:t+4})\|_2} \tag{4}$$

广义训练过程中的整体损失函数可写为：

$$\mathcal{L}_g = \mathcal{L}_{adv} + \lambda_r \mathcal{L}_{rec} + \lambda_s \mathcal{L}_{sync}. \tag{5}$$

### 3.3. 个性化优化

骨干网络训练完成后，我们的方法便可为任意主体生成高保真唇同步结果。然而，由于模型采用广义方式训练，不同人物生成的嘴部运动模式基本相同。已有研究证实，不同身份具有不同的说话风格 [54, 59]。在此，我们利用原始模板视频捕捉这种个性化特征。值得注意的是，我们关注少样本设置，即仅提供不到一分钟的原始视频，而先前的特定人物模型 [16, 20, 40, 46]无法处理这种情况。下面，我们将介绍如何基于以上讨论成功设计个性化优化模块。

**基本学习设置**。 个性化优化的基本学习过程与骨干网络训练类似。对于带有音频 $\mathbf{a} = \{\bar{a}_1 \ldots, \bar{a}_T\}$ 的模板视频 $\mathbf{V} = \{I_1, \ldots, I_T\}$，我们随机选取两帧，分别作为目标 $I_t$ 和参考帧 $I_{ref}$，作为一个样本进行训练。在训练过程中，通用损失 $\mathcal{L}_g$ 持续监督整个过程。

由于编码器 $E_a$ 编码的是与人物无关的语音内容信息，因此在个性化优化步骤中将其固定。虽然人脸编码器 $E_{face}$ 编码的是目标人物的属性信息，但为了避免过拟合，我们不会针对单个目标人物优化其参数。

**风格空间优化**。 根据我们的公式，$W^+$ 空间包含唇部运动和高层面部风格信息。因此，对其进行优化理想情况下会产生个性化的唇部运动模式。E4E [47]证实，学习原始 $W$ 空间附近的小幅位移，既能提升图像的反演质量，也能实现强大的编辑能力。因此，我们提出以类似的方式，根据编码后的面部特征 $f_f$，通过几层多层感知机 $MLP_{\Delta w}$学习一组 $\Delta\mathbf{W} = MLP_{\Delta w}(f_f)$。

**生成器调优**。 生成器对合成和融合质量的影响最大。我们允许生成器 G 的参数**P** 在使用个性化数据时稍微偏移至 $\Delta\mathbf{P}$。

**学习目标**。 个性化优化过程的最终学习目标可总结如下：

$$\mathcal{L}_p = \mathcal{L}_g + \lambda_p\left(\sum_{i,j} |\Delta\mathbf{W}_i^j|_2^2 + \sum_{m,n} |\Delta\mathbf{P}_n^m|_2^2\right), \tag{6}$$

其中，$\mathbf{W}_i^j$ 表示$\mathbf{W}$中第 $i$个元素的第 $j$个参数。$\mathbf{P}_n^m$同理。我们将位移限制在有限步长内。

## 4. 实验

**数据集**。 我们在两个常用的音视频数据集 LRW [11] 和 VoxCeleb2 [10] 上进行实验。我们遵循数据集原有的训练集/测试集划分，并使用这两个数据集的混合数据训练模型。

- **LRW** [11] 是从英国广播公司新闻中收集的唇读音视频数据集。它是最早的大规模高质量音视频数据集之一。该数据集包含 500 个单词的 1,000 段一秒钟长的语音片段。

- **VoxCeleb2** [10]VoxCeleb [31]和 VoxCeleb2 是为说话人验证任务收集的大规模音视频数据集。我们重新下载了约五分之一的 VoxCeleb2 用于训练，这些数据质量较高。官方测试集经过 GPEN [57]处理后用于评估。

**实现细节**。 我们以 25 帧/秒处理所有视频，并根据预先检测到的眼部关键点对齐所有人脸。所有人脸均裁剪为 $256 \times 256$ 的尺寸。如图 2 所示，我们采用相同的 U形掩码，擦除目标帧中的嘴部、脸颊和下颌。所有音频均按照 Wav2Lip 的相同方式处理 [33]。为了简化设计并避免额外调整超参数，我们采用了先前研究中的大部分设置 [24, 33, 57]。具体而言，生成器总共包含 $L = 14$ 个风格卷积层。 $\lambda_r$ 的经验设定值为 10，所有其他 $\lambda$ 均设为 1。对于个性化优化，我们使用该人物的数据训练 5 个训练轮次。延长训练时间并不会带来更好的结果。特别是，我们发现，将个性化模型与通用训练数据一同训练会带来更稳定的结果；在我们的实验中，个人数据与通用数据的比例设为 1:1。

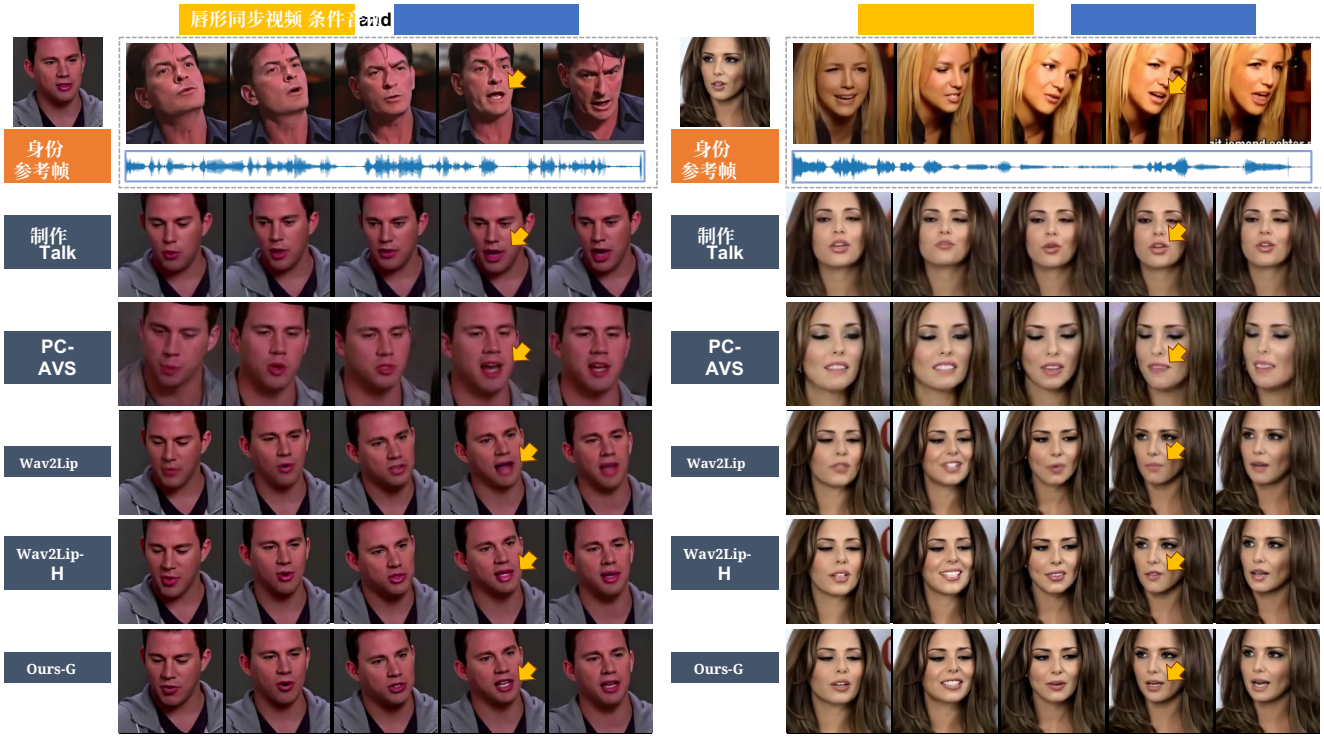


图 3. **定性结果**。第一行展示了条件驱动音频对应的唇形同步视频。MakeitTalk [66] 无法生成准确的嘴部形状，且缺乏头部动态。PC-AVS [65] 和 Wav2Lip [33] 生成的唇形同步结果与音频配合良好，但画面明显模糊。特别是，我们复现的 Wav2Lip-H 能生成高质量结果，且唇形同步效果良好。不过，它们仍会产生一些伪影。相比之下，我们的泛化模型可直接生成高保真结果，其唇部运动与原始同步视频一致。

**比较方法**。由于我们的模型基于与人物无关的设置构建，因此我们将 StyleSync框架与三种最先进的方法进行比较，包括 MakeitTalk [66]，PC-AVS [65] 和 **Wav2Lip** [33]。**MakeitTalk** [66] 合成具有自然头部姿态的说话头部视频。**PC-AVS** [65] 使用姿态源视频来实现姿态可控的说话人脸生成。MakeitTalk 和 PC-AVS 会生成完整的说话头部，而专注于嘴部的 Wav2Lip 是我们主要的竞争方法。由于 Wav2Lip 最初是在低分辨率下使用小型网络训练的，我们在与我们的方法相同的数据和损失函数上对其进行了精心复现。我们将该模型记为 **Wav2Lip-H** （高质量）。为进行公平比较，所有基础比较均在泛化设置下进行 **Ours-G**。

我们还在测试集上进行个性化优化，其中两个数据集中的每帧时长均少于 10 秒。结果记为 **我们-P**。我们同样对上述方法进行生成器微调，但它们的表现均不如原始模型，因此下文不再列出相关结果。此外，我们将我们的个性化设置与 ADNeRF [16], a 一种近期基于神经辐射场的说话头部技术 [30]进行比较。

### 4.1. 定量评估

由于定量评估只能在自建设置下进行，我们避免直接利用帧内的全部信息进行重建。对于 PC-AVS、Wav2Lip、Wav2Lip-H 和我们的方法，我们统一随机选取同一帧作为参考帧。

**评估指标**。我们遵循以往研究 [8, 9, 64,65]，采用常用的结构相似性指数（SSIM） [53], 和峰值信噪比（PSNR）指标评估生成质量，并采用嘴部周围的标志点距离（唇部标志点距离，LMD）以及 SyncNet 的 [12] 置信度分数评估唇形同步质量。请注意，SyncNet 指标是在官方发布的版本上计算的，这与我们在损失函数中的实现不同。

特别是，由于无法完全恢复目标人物的细节，我们利用 ArcFace [13] 网络，逐帧计算特征余弦距离 $\mathcal{D}_{ID}$，以评估生成结果是否很好地保留了身份。

**评估结果**。定量实验在 LRW [11]和 VoxCeleb2 [10] 数据集的测试集上进行。结果请参见表 1 。可以看出，在 LRW 数据集上，通用版本

| 方法 | LRW | | | | | VoxCeleb2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 结构相似性指数 ↑ | 峰值信噪比 ↑ | 唇部标志点距离 ↓ | 同步$_{conf}$ ↑ | $\mathcal{D}_{ID}$ ↑ | 结构相似性指数 ↑ | 峰值信噪比 ↑ | 唇部标志点距离 ↓ | 同步$_{conf}$ ↑ | $\mathcal{D}_{ID}$ ↑ |
| MakeItTalk | 0.69 | 29.83 | 2.75 | 3.88 | 0.79 | 0.63 | 28.38 | 6.94 | 2.15 | 0.71 |
| PC-AVS | 0.79 | 30.26 | 1.84 | 7.19 | 0.82 | 0.71 | 29.53 | 2.75 | 8.16 | 0.74 |
| Wav2Lip | 0.79 | 30.54 | 1.28 | 7.39 | **0.90** | 0.80 | 30.53 | 1.92 | 8.90 | **0.90** |
| Wav2Lip-H | 0.80 | 31.38 | 1.20 | 7.19 | 0.88 | **0.81** | 30.53 | 1.87 | 8.35 | **0.90** |
| GT | 1.00 | 100.00 | 0.00 | 7.65 | 1.00 | 1.00 | 100.00 | 0.00 | 7.71 | 1.00 |
| Ours-G | **0.85** | **31.78** | **1.18** | 7.25 | 0.89 | 0.79 | **31.00** | **1.47** | 8.25 | **0.90** |
| 我们-P | **0.88** | **32.66** | **0.86** | 6.35 | **0.93** | **0.82** | **31.54** | **1.15** | 7.26 | **0.93** |

表 1. **LRW 和 VoxCeleb 上的定量结果**。对于唇部标志点距离，数值越低越好；对于其他指标，数值越高越好。

的 StyleSync 在生成质量方面已大幅优于以往方法。而在 VoxCeleb2 数据集上，差距则不那么明显。这可能是因为我们的方法所用的大多数训练数据基本上都是从 LRW 中选取的正面人脸，而 VoxCeleb2 包含更复杂的场景。另一方面，经过个性化优化后，我们模型的性能再次提升。这表明此阶段生成的结果与目标明显更加相似，这一点也可以通过查看身份距离得到验证。

与此同时，我们的方法在两个数据集上的唇同步指标方面也取得了具有竞争力的性能。我们的唇部标志点距离得分略优于其他方法。至于 SyncNet 得分，我们在 LRW 上取得了相当的结果，并且更接近真实值的 SyncNet 得分。我们认为，一旦该指标超过真实值，继续参考 SyncNet 得分来验证唇形同步质量就没有意义了。$Sync_{conf}$ 该指标仅反映视听对与所学习的 SyncNet 模型的匹配程度，而非真实的感知质量。因此，尽管生成结果在该指标上可能优于真实值，但这并不意味着同步质量更好。

个性化处理后，唇形同步得分有所下降。我们推测，说话风格的独特性会降低张口程度。图5也显示了这一点。该女性不太会张大嘴巴，因此嘴部运动幅度较小

## 4.2. 定性评估

主观评估对于判断生成模型的能力至关重要，尤其是在视频方面。我们强烈建议读者观看我们的**补充视频**，网址：https://hangz-nju-cuhk.github.io/projects/StyleSync。

图中展示了两个交叉驱动示例及其与最先进方法的比较。

3。我们使用从测试集中任意选取的一位人物的条件音频来驱动模板。可以看出，MakeitTalk [66]无法生成准确的唇部运动。此外，PC-AVS 和 Wav2Lip 都会产生明显的伪影或模糊




图 4. 自驱动结果与个性化方法的比较。

结果。我们复现的 Wav2Lip-H 采用与我们相同的训练设置，因此在大多数情况下都能生成合理的结果。尽管如此，我们的泛化模型仍能生成最准确、保真度最高的嘴型。

此外，我们在图中展示了一个针对 50 秒视频片段进行个性化优化的案例。1可以看出，我们的方法能够在实现精准唇同步的同时，保留目标人物 [16]（图4）。与 AD-NeRF 相比，我们的个性化结果（我们-P）在保留更多人物特有细节方面具有更高保真度，尽管我们使用的个人数据显著更少（微调使用 10 秒，而 [16]中的训练使用 5 分钟）。

**用户研究**。我们还邀请了 15 名参与者开展用户研究，以进行进一步的主观评估，结果见表2。具体而言，我们从 LRW 和 VoxCeleb2 数据集中随机选取 63 个视频作为测试集，并使用我们的泛化版 StyleSync 和相应的对比方法生成视频。我们采用常用的平均意见分（MOS）评分方案，请所有参与者从三个方面对每个生成视频进行评分（1 至 5 分，分数越高越好）：（1）唇形同步质量；（2）生成质量；（3）视频真实感。

如表2所示，MakeitTalk [66]由于缺乏合理的唇部运动和头部姿态，在所有方面的得分都最低。PC-AVS [65], Wav2Lip [33]和 Wav2Lip-H 的唇部同步得分相对较高，但只有 Wav2Lip-H 能够合成嘴部周围纹理较不模糊的视频。总体而言，我们的 StyleSync

| \ 方法的平均意见分 | | | | | |
|---|---|---|---|---|---|
| 唇形同步质量 | 2.06 | 3.00 | 3.49 | 3.67 | **4.24** |
| 生成质量 | 2.63 | 2.16 | 1.87 | 3.42 | **4.52** |
| 视频真实感 | 1.89 | 2.16 | 2.17 | 2.98 | **4.06** |

表2. 用户研究采用平均意见分进行测量。分数范围为1（最差）至5（最好）。

| 方法 | 结构相似性指数 ↑ | 峰值信噪比 ↑ | 唇部标志点距离 ↓ | 同步$_{conf}$ ↑ | $\mathcal{L}_{ID}$ ↑ |
|---|---|---|---|---|---|
| 无掩码 | 0.81 | 31.20 | 1.80 | 7.89 | 0.90 |
| 无同步 | 0.80 | 31.34 | 1.55 | 7.41 | 0.90 |
| Ours-G | 0.79 | 31.00 | **1.47** | **8.25** | 0.90 |
| 我们-P 无**W** | 0.82 | 31.51 | 1.32 | 7.31 | 0.92 |
| 我们-P | 0.82 | **31.54** | **1.15** | 7.26 | **0.93** |

表 3. VoxCeleb 上的定量消融实验结果　　2.

在这三个方面，我们的方法都以较大优势胜过其他方法，表明了我们方法的有效性。

### 4.3. 消融研究

为进一步展示我们新颖设计的贡献，我们在 VoxCeleb2 数据集上进行了消融研究，涵盖广义设置和个性化优化设置（分别记为“Ours-G”和“Ours-P”）。具体而言，我们构建了“Ours-G”的两个变体（记为“无掩码”和“无同步”），分别移除了公式 1 中的掩码操作和训练过程中的唇形同步损失。对于“Ours-P”，我们构建了一个变体（记为“Ours-P 无 $\Delta\mathbf{W}$”），方法是将 $\Delta\mathbf{W} = \mathbf{0}$ 固定。我们也尝试了使用固定生成结果进行个性化优化，但这会导致结果模糊，因此这里省略了该消融实验。

定量和定性结果如表3 和图5所示。如表3所示，我们观察到，“无掩码”和“无同步”在图像质量指标上的表现相当，但唇同步指标有所下降。与“Ours-G”相比，“无掩码”在唇部标志点距离和同步方面的得分较低，这表明所提出的掩码策略能够显著减轻参考帧带来的影响。与此同时，“无同步”的唇形同步效果有所下降，图5 中的结果也显示，与唇形同步视频相比，唇部运动不一致。这些结果证明了 SyncNet 提供的额外监督是有效的。

从图 5 可以看出，经过个性化优化后，生成帧中的身份和嘴部张开的模式都变得更加接近原始模板视频。显然，经过个性化训练后，嘴部张开的幅度不如泛化模型明显。我们认为这是

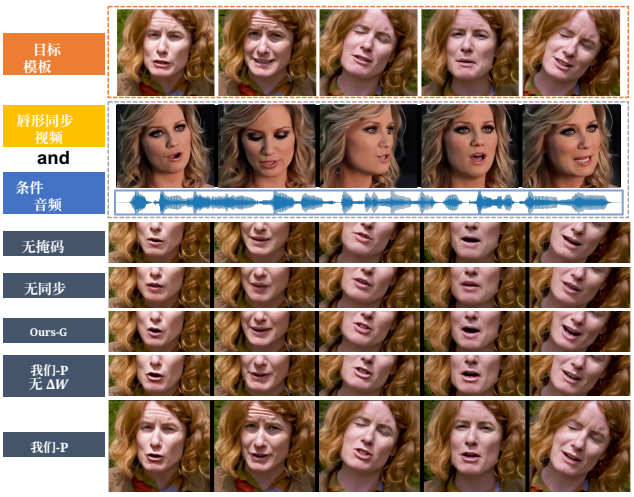


图 5. 消融研究的可视化结果。放大查看细节。

该目标人物说话风格的一部分。这也导致 SyncNet 上的指标值较差。

将我们的个性化模型与“Ours-P 无 $\Delta\mathbf{W}$”进行比较，在图像质量和唇形同步质量方面，两者得分相近。不过，不使用 $\Delta\mathbf{W}$ 时，身份得分略有下降，这表明在训练过程中加入 $\Delta\mathbf{W}$ 有助于更好地保留身份。

## 5. 结论与讨论

**结论。** 本文提出了**StyleSync**，能够在单样本和少样本设置下生成高保真唇同步结果。我们强调本方法的独特之处：**1)** 通用模型生成的视频结果明显优于以往最先进方法。**2)** 我们的模型以近期基于风格的生成器的成功为基础，只需进行简单修改即可构建。该模型易于实现，也便于训练。**3)** 通过引入受 StyleGAN 启发的个性化优化过程，只需提供少量视频片段，我们的模型就能针对特定人物进一步优化。

**局限性。** 由于我们的方法使用固定掩码，将唇同步后的面部融合到现有视频中，因此无法改变目标人物的头部姿态和表情。此外，在某些极端情况下，如果目标人物的下颌特别大，可能会超出我们的遮罩区域。

**伦理考量。** 我们的方法可能被用于制作并不存在的谈话和演讲内容，并可能遭到恶意使用。我们将严格限制仅向研究机构提供核心模型。

**致谢。** 本研究由新加坡教育部通过其教育部学术研究基金第二层级（MOE-T2EP20221-0012）、南洋理工大学南洋助学金，以及RIE2020产业协同基金–产业合作项目（IAF-ICP）资助计划提供支持，同时还获得了产业合作伙伴的现金及实物捐助。

## 参考文献


[1]Rameen Abdal、Yipeng Qin 和 Peter Wonka。《Image2stylegan：如何将图像嵌入 StyleGAN 潜在空间？》载于 IEEE/CVF国际计算机视觉会议论文集，第4432–4441页，2019年。2、3、4[2]Rameen Abdal、Yipeng Qin 和 Peter Wonka。《Image2stylegan++：如何编辑嵌入的图像？》载于 IEEE/CVF计算机视觉与模式识别会议论文集，第8296–8305页，2020年。2、3、4[3] Yuval Alaluf、Or Patashnik 和 Daniel Cohen-Or。《Restyle：一种通过迭代细化实现的基于残差的 StyleGAN 编码器》。载于 IEEE/CVF国际计算机视觉会议论文集，第6711–6720页，2021年。3[4] Yuval Alaluf、Omer Tov、Ron Mokady、Rinon Gal 和 Amit Bermano。《Hyperstyle：使用超网络进行 StyleGAN 反演以实现真实图像编辑》。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第18511–18521页，2022年。4[5] Egor Burkov、Igor Pasechnik、Artur Grigorev 和 Victor Lempitsky。《使用潜在姿态描述符进行神经头部重演》。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第13786–13795页，2020年。2、3、4[6] Lele Chen、Guofeng Cui、Ziyi Kou、Haitian Zheng 和 Chenliang Xu。《什么构成了优质的说话人头部视频生成？：综述与基准》。arXiv预印本 arXiv:2005.03201，2020年。2[7] Lele Chen、Guofeng Cui、Celong Liu、Zhong Li、Ziyi Kou、Yi Xu 和 Chenliang Xu。《具有节奏性头部运动的说话人头部生成》。载于欧洲计算机视觉会议论文集，第35–51页。施普林格，2020年。2[8] Lele Chen、Zhiheng Li、Ross K Maddox、Zhiyao Duan 和 Chenliang Xu。《唇部运动生成一览》。载于欧洲计算机视觉会议（ECCV）论文集，第520–535页，2018。2、6[9] Lele Chen、Ross K Maddox、Zhiyao Duan 和 Chenliang Xu。《具有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论文集，第7832–7841页，2019年。2、6[10] J. S. Chung、A. Nagrani 和 A. Zisserman。《VoxCeleb2：深度说话人识别》。载于 INTERSPEECH，2018。5、6[11] Joon Son Chung 和 Andrew Zisserman。《野外唇读》。载于亚洲计算机视觉会议，第87–103页。施普林格，2016年。5、6[12] Joon Son Chung 和 Andrew Zisserman。《超越时间：野外自动唇同步》。载于亚洲计算机视觉会议，2016年。5、6[13] Yu Deng、Jiaolong Yang、Sicheng Xu、Dong Chen、Yunde Jia 和 Xin Tong。《基于弱监督学习的精准三维人脸重建：从单幅图像到图像集》。载于 IEEE/CVF计算机视觉与模式识别会议研讨会论文集，第0–0页，2019年。6[14] Bo Fan、Lijuan Wang、Frank K Soong 和 Lei Xie。《基于深度双向长短期记忆网络的逼真说话人头部》。载于 2015 IEEE 国际声学、语音与信号处理会议（ICASSP），第4884–4888页。IEEE，2015年。2[15] Yingruo Fan、Zhaojiang Lin、Jun Saito、Wenping Wang和Taku Komura。Faceformer：使用Transformer的语音驱动三维面部动画。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第18770–18780页，2022年。2[16] Yudong Guo、Keyu Chen、Sen Liang、Yongjin Liu、Hujun Bao和Juyong Zhang。Ad-NeRF：用于说话头部合成的音频驱动神经辐射场。载于IEEE/CVF国际计算机视觉会议（ICCV），2021年。2、3、5、6、7[17] Siddharth Gururani、Arun Mallya、Ting-Chun Wang、Rafael Valle和Ming-Yu Liu。SpaceX：具有可控表情的语音驱动肖像动画。arXiv预印本 arXiv:2211.09809，2022年。2[18] Amir Jamaludin、Joon Son Chung和Andrew Zisserman。你说什么？：从音频合成说话人脸。《国际计算机视觉杂志》，127(11)：1767–1779，2019年。2[19] Xinya Ji、Hang Zhou、Kaisiyuan Wang、Qianyi Wu、Wayne Wu、Feng Xu和Xun Cao。EAMM：通过基于音频的情感感知运动模型实现单样本情感说话人脸。SIGGRAPH，2022年。2[20] 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《IEEE/CVF计算机视觉与模式识别会议论文集》，第14080–14089页，2021年。2、5[21] Tero Karras、Timo Aila、Samuli Laine、Antti Herva和Jaakko Lehtinen。通过姿态与情感的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）》，36(4)：1–12，2017年。2[22] Tero Karras、Miika Aittala、Samuli Laine、Erik H¨ark¨onen、Janne Hellsten、Jaakko Lehtinen和Timo Aila。无混叠生成对抗网络。载于NeurIPS论文集，2021年。3[23] Tero Karras、Samuli Laine和Timo Aila。用于生成对抗网络的基于风格的生成器架构。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第4401–4410页，2019年。2、3[24] Tero Karras、Samuli Laine、Miika Aittala、Janne Hellsten、Jaakko Lehtinen和Timo Aila。分析并改进StyleGAN的图像质量。载于CVPR论文集，2020年。2、3、4、5[25] Avisek Lahiri、Vivek Kwatra、Christian Frueh、John Lewis和Chris Bregler。LipSync3D：利用姿态与光照归一化从视频中高效学习个性化三维说话人脸。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第2755–2764页，2021年。2[26] Borong Liang、Yan Pan、Zhizhi Guo、Hang Zhou、Zhibin Hong、Xiaoguang Han、Junyu Han、Jingtuo Liu、Errui Ding和Jingdong Wang。具有细粒度音视频控制的表现力说话头部生成。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第3387–3396页，2022年。2、3

[27] Xian Liu、Yinghao Xu、Qianyi Wu、Hang Zhou、Wayne Wu和Bolei Zhou。语义感知的隐式神经音频驱动视频肖像生成。ECCV，2022年。2、3[28] Yuanxun Lu、Jinxiang Chai和Xun Cao。实时语音肖像：实时照片级说话头部动画。《ACM图形学汇刊（TOG）》，40(6)：1–17，2021年。2[29] Yifeng Ma、Suzhen Wang、Zhipeng Hu、Changjie Fan、Tangjie Lv、Yu Ding、Zhidong Deng和Xin Yu。Styletalk：具有可控说话风格的单样本说话头部生成。AAAI，2023年。2[30]Ben Mildenhall、Pratul P Srinivasan、Matthew Tancik、Jonathan T Barron、Ravi Ramamoorthi和Ren Ng。NeRF：将场景表示为用于视图合成的神经辐射场。载于《欧洲计算机视觉会议》，第405–421页。施普林格，2020年。6[31] A. Nagrani、J. S. Chung和A. Zisserman。VoxCeleb：大规模说话人识别数据集。载于INTERSPEECH，2017年。5[32] Se Jin Park、Minsu Kim、Joanna Hong、Jeongsoo Choi和Yong Man Ro。SyncTalkFace：通过音频-唇形记忆实现精准唇形同步的说话人脸生成。载于AAAI人工智能会议。人工智能促进协会，2022年。2、3[33] KR Prajwal、Rudrabha Mukhopadhyay、Vinay P Namboodiri和CV Jawahar。野外语音到唇形生成只需一个唇形同步专家。载于第28届ACM多媒体国际会议论文集，第484–492页，2020年。2、3、4、5、6、7[34] Alexander Richard、Michael Zollh¨ofer、Yandong Wen、Fernando de la Torre和Yaser Sheikh。MeshTalk：通过跨模态解耦实现基于语音的三维人脸动画。载于《IEEE/CVF国际计算机视觉会议（ICCV）论文集》，2021年。2[35] Elad Richardson、Yuval Alaluf、Or Patashnik、Yotam Nitzan、Yaniv Azar、Stav Shapiro和Daniel Cohen-Or。以风格编码：用于图像到图像转换的StyleGAN编码器。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第2287–2296页，2021年。3[36] Daniel Roich、Ron Mokady、Amit H Bermano和Daniel Cohen-Or。真实图像基于潜变量的编辑的枢轴调优。《ACM图形学汇刊》，2021年。2、3、4[37] Shuai Shen、Wanhua Li、Zheng Zhu、Yueqi Duan、Jie Zhou和Jiwen Lu。面向小样本说话头部合成的动态人脸辐射场学习。载于ECCV，2022年。2、3[38] Yujun Shen、Ceyuan Yang、Xiaoou Tang和Bolei Zhou。InterfaceGAN：解读生成对抗网络学习到的解耦人脸表示。《TPAMI》，2020年。3[39] Yujun Shen和Bolei Zhou。生成对抗网络中潜在语义的闭式分解。载于CVPR，2021年。3[40] Linsen Song、Wayne Wu、Chen Qian、Ran He和Chen Change Loy。大家都在说话：让我按你的意愿说话。arXiv预印本arXiv:2001.05201，2020年。2、3、5[41] Yang Song、Jingwen Zhu、Dawei Li、Xiaolong Wang和Hairong Qi。通过条件循环生成对抗网络实现说话人脸生成。arXiv预印本arXiv:1804 04786，2018年。2、3。

[42] Yasheng Sun、Hang Zhou、Ziwei Liu 和 Hideki Koike。Speech2talking-face：从同步的音视表征中推断并驱动人脸。在 IJCAI，2021 年。2，4[43] Yasheng Sun、Hang Zhou、Kaisiyuan Wang、Qianyi Wu、Zhibin Hong、Jingtuo Liu、Errui Ding、Jingdong Wang、Ziwei Liu 和 Hideki Koike。通过在 Transformer 中利用音视上下文进行掩码唇同步预测。在 SIGGRAPH Asia 2022 Conference Papers，第 1–9 页，2022 年。2[44] Supasorn Suwajanakorn、Steven M Seitz 和 Ira Kemelmacher-Shlizerman。合成奥巴马：从音频中学习唇同步。ACM图形学汇刊（TOG），36(4)：1–13，2017 年。2[45] Jiaxiang Tang、Kaisiyuan Wang、Hang Zhou、Xiaokang Chen、Dongliang He、Tianshu Hu、Jingtuo Liu、Gang Zeng 和 Jingdong Wang。通过音频-空间分解实现实时神经辐射说话肖像合成。arXiv 预印本 arXiv:2211.12368，2022 年。3[46] 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。在欧洲计算机视觉会议，第 716–731 页。施普林格，2020 年。2，3，5[47] Omer Tov、Yuval Alaluf、Yotam Nitzan、Or Patashnik 和 Daniel Cohen-Or。为 StyleGAN 图像操控设计编码器。ACM图形学汇刊（TOG），2021 年。2，3，5[48] Kaisiyuan Wang、Qianyi Wu、Linsen Song、Zhuoqian Yang、Wayne Wu、Chen Qian、Ran He、Yu Qiao 和 Chen Change Loy。MEAD：用于情感说话人脸生成的大规模音视数据集。在 ECCV，2020 年。2[49] Suzhen Wang、Lincheng Li、Yu Ding、Changjie Fan 和 Xin Yu。Audio2Head：具有自然头部运动的音频驱动单样本说话头生成。arXiv 预印本 arXiv:2107.09293，2021 年。2[50] Suzhen Wang、Lincheng Li、Yu Ding 和 Xin Yu。通过单说话人音视相关性学习实现从单说话人音频生成单样本说话人脸。在 AAAI人工智能会议论文集，2022 年。2[51] Ting-Chun Wang、Ming-Yu Liu、Jun-Yan Zhu、Andrew Tao、Jan Kautz 和 Bryan Catanzaro。使用条件生成对抗网络进行高分辨率图像合成与语义操控。在 IEEE 计算机视觉与模式识别会议论文集，第 8798–8807 页，2018 年。4[52] Xintao Wang、Yu Li、Honglun Zhang 和 Ying Shan。面向真实世界的基于生成式人脸先验的盲人脸修复。在 IEEE 计算机视觉与模式识别会议（CVPR），2021 年。3[53] 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，13(4)：600–612，2004 年。6[54] Haozhe Wu、Jia Jia、Haoyu Wang、Yishun Dou、Chao Duan 和 Qingshan Deng。模仿任意说话风格，实现逼真的音频驱动说话人脸合成。在第 29 届 ACM 多媒体国际会议论文集，第 1478–1486 页，2021 年。2，5[55] Chao Xu、Jiangning Zhang、Miao Hua、Qian He、Zili Yi 和 Yong Liu。区域感知人脸交换。在

IEEE/CVF计算机视觉与模式识别会议论文集，第 7632–7641 页，2022 年。3[56] Zhiliang Xu、Hang Zhou、Zhibin Hong、Ziwei Liu、Jiaming Liu、Zhizhi Guo、Junyu Han、Jingtuo Liu、Errui Ding 和 Jingdong Wang。StyleSwap：基于风格的生成器赋能稳健的人脸交换。在欧洲计算机视觉会议，第 661 677 2022 2, 3 – 页。施普林格，。[57] Tao Yang、Peiran Ren、Xuansong Xie 和 Lei Zhang。用于野外盲人脸修复的嵌入式 GAN 先验网络。在 IEEE/CVF计算机视觉与模式识别会议论文集，第 672–681 页，2021 年。2，3，5[58] Zhenhui Ye、Ziyue Jiang、Yi Ren、Jinglin Liu、Jinzheng He 和 Zhou Zhao。GeneFace：通用且高保真的音频驱动三维说话人脸合成。ICLR，2023年。2[59] Ran Yi、Zipeng Ye、Juyong Zhang、Hujun Bao 和 Yong-Jin Liu。通过基于学习的个性化头部姿态实现音频驱动的说话人脸视频生成。arXiv 电子预印本，第 arXiv–2002 页，2020 年。2，5[60] Fei Yin、Yong Zhang、Xiaodong Cun、Mingdeng Cao、Yanbo Fan、Xuan Wang、Qingyan Bai、Baoyuan Wu、Jue Wang 和 Yujiu Yang。StyleHEAT：通过预训练 StyleGAN 实现单样本高分辨率可编辑说话人脸生成。ECCV，2022 年。2[61] Egor Zakharov、Aliaksandra Shysheya、Egor Burkov 和 Victor Lempitsky。逼真神经说话头模型的少样本对抗学习。在 IEEE/CVF国际计算机视觉会议论文集，第 9459–9468 页，2019年。2[62] Chenxu Zhang、Yifan Zhao、Yifei Huang、Ming Zeng、Saifeng Ni、Madhukar Budagavi 和 Xiaohu Guo。FACIAL：通过隐式属性学习合成动态说话人脸。在 IEEE/CVF国际计算机视觉会议（ICCV）论文集，第 3867–3876 页，2021 年。2[63] Zhimeng Zhang、Zhipeng Hu、Wenjin Deng、Changjie Fan、Tangjie Lv 和 Yu Ding。DINet：用于高分辨率视频逼真人脸视觉配音的形变修复网络。AAAI，2023年。2[64] Hang Zhou、Yu Liu、Ziwei Liu、Ping Luo 和 Xiaogang Wang。通过对抗式解耦音视表征实现说话人脸生成。在 AAAI人工智能会议论文集，第 33 卷，第 9299–9306 页，2019年。2，6[65] Hang Zhou、Yasheng Sun、Wayne Wu、Chen Change Loy、Xiaogang Wang 和 Ziwei Liu。通过隐式模块化音视表征实现姿态可控的说话人脸生成。在 IEEE/CVF计算机视觉与模式识别会议论文集，第 4176–4186 页，2021 年。2，3，4，6，7[66] Yang Zhou、Xintong Han、Eli Shechtman、Jose Echevarria、Evangelos Kalogerakis 和 Dingzeyu Li。MakeItTalk：说话人感知的说话头动画。ACM 图形学汇刊（TOG），39(6)：1–15，2020 年。2，6，7[67] Yang Zhou、Zhan Xu、Chris Landreth、Evangelos Kalogerakis、Subhransu Maji 和 Karan Singh。VisemeNet：音频驱动的动画师中心语音动画。ACM 图形学汇刊（TOG），37(4)：1–10，2018 年。2