联系电话：13911994845

联系邮箱：2021211013072@cuc.edu.cn

VR/AR-AdaptFace：面向虚拟现实与增强现实的自适应多模态面部替换模型

靳聪[1]，周满玲[1]，林美秀[1]，张佳一[2]，王晶[3*]，刘淼[3]

（1. 中国传媒大学信息与通信工程学院，北京 100024；2. 中国传媒大学广告学院，北京 100024；3.北京理工大学信息与电子学院，北京 100081）

摘要：随着 VR/AR 技术的迅猛进步，用户对于沉浸式体验的需求日益增长。同时，虚拟人脸技术亦趋成熟。基于此，本文探索将高度拟真的虚拟人脸融入 VR/AR，以增强用户体验的自然度与沉浸感。然而，在虚拟数字人领域，图像生成及换脸技术在 VR/AR 环境下仍遇诸多挑战，尤其是唇形合成模型在动态场景及多语言环境下的性能需进一步优化。为解决上述问题，本文提出 VR/AR-AdaptFace 模型，一个面向虚拟现实与增强现实的自适应多模态面部替换方案。该模型由两大模块构成：“文颜绘真”模块，采用先进的文本至图像转换技术和特定类别先验保存策略，优化虚拟人脸生成，并通过注意力机制大幅提升图像质量；“语唇映生”模块，依托强大的生成器、唇形同步判别器及视觉质量判别器，实现语音与唇形的精准同步，为 VR/AR 场景中的动态交互带来更加逼真的体验。

关键词：人脸合成；细节增强模型；动态视频唇形合成；虚拟现实；增强现实

中图分类号：TP181　　**文献标识码：A**

VR/AR-AdaptFace: Adaptive Multimodal Face Replacement Model for Virtual Reality and Augmented Reality

JIN Cong[1], ZHOU Manling[1], LIN Meixiu[1], ZHANG Jiayi[2], WANG Jing[3*], LIU Miao[3]

(1. School of Information and Communication Engineering, University of China, Beijing100024,China; 2.School of Advertising, Communication University, Beijing100024, China; 3. School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China)

Abstract: With the rapid advancement of VR and AR technologies, there is a growing demand for immersive experiences. At the same time, virtual face technology is also becoming mature. Based on this, this paper explores the integration of highly realistic virtual faces into VR/AR to enhance the naturalness and immersion of user experience. However, in the field of virtual digital human,

image generation and face-swapping techniques still encounter many challenges in VR/AR environments, especially the lip-synthesis model needs to be further optimised in dynamic scenes and multi-language environments. To solve the above problems, this paper proposes the VR/AR-AdaptFace model, an adaptive multimodal face replacement scheme for virtual reality and augmented reality. The model consists of two major modules: the "text-to-image" module, which uses advanced text-to-image conversion techniques and category-specific a priori retention strategies to optimise virtual face generation, and significantly improves the image quality through the attention mechanism; and the "speech-to-lip reflection" module, which The "speech-lip reflection" module, on the other hand, relies on a powerful generator, lip synchronisation discriminator and visual quality discriminator to achieve accurate synchronisation between speech and lip shape, bringing a more realistic experience for dynamic interaction in VR/AR scenes.
Keywords：Face synthesis; detail-enhanced modelling; Motion Video Lip Synthesis; virtual reality; augmented reality

1 引言

随着虚拟现实（virtual reality, VR）和增强现实（augmented reality, AR）技术的快速发展，用户对于高度逼真、沉浸式体验的需求日益增长。然而，在VR/AR应用中，真实人物的隐私保护和逼真度成为了挑战。与此同时，虚拟数字人技术的进步为该问题提供了新的解决方案。许多先进的模型已经能够生成栩栩如生的虚拟人脸，这引发了作者的思考：是否可以将这些技术应用于VR/AR场景，以替代真实人物？通过将虚拟人脸与唇形合成技术结合，有望在VR/AR环境中创造出既保护个人隐私又极具真实感的虚拟角色。这种方法将为VR/AR领域带来更多的创新空间和用户体验。然而，尽管深度学习在图像生成领域取得了显著进展，例如DualNet的快速学习和慢速学习系统在可持续学习中的应用，以及扩散模型在图像生成中的广泛使用，但在VR/AR应用中，仍面临一些挑战。目前，深度学习图像生成技术在人脸生成方面仍存在图像真实性、分辨率和样本多样性等问题。此外，现有的唇形合成模型在动态VR/AR场景合成方面表现不佳，且主要基于英文数据集训练，导致中文等其他语言的唇形合成效果并不理想。为了解决上述问题，本文提出了一个面向虚拟现实与增强现实的自适应多模态面部替换模型（VR/AR-AdaptFace）。该模型由“文颜绘真”和“语唇映生”两大核心模型组成，旨在提升VR/AR应用中虚拟角色的逼真度和动态交互效果。

“文颜绘真”模型包括两个主要部分：扩散模型生成人脸和数字化人脸细节增强。通过扩散模型，能够生成基础的人脸结构，同时结合数字化人脸细节增强技术，进一步提升人脸图像的真实感和细节表现。针对扩散模型生成人脸部分，本文分为两个模块，分别是文本生成图像的扩散模型[1]和特定类别的先验保存损失[2]，通过利用类生成样本和先验保存损失进行模型微调的方法解决了语言漂移问题[3][4]。针对数字化人物细节增强部分本文结合特征编码、通道注意力机制、空间注意力机制和特征重建等关键技术，以生成更加逼真和高分辨率的合成人脸图像。

“语唇映生”模型由唇形同步判别器、生成器和视觉质量判别器构成。唇形同步判别器类似SyncNet，

基金项目：国家自然科学基金(62207029，62271454)，北京市自然科学基金-小米联合基金(L223033)，中央高校基础科研经费(CUC230B018)
作者简介（*为通讯作者）：靳聪(1986-)，女，博士, 副教授，研究方向为音乐人工智能、智慧教育、人机混合表演等。E-mail：jincong0623@cucedu.cn；周满玲（2003-），女，本科生，主要从事图像处理、深度学习等研究。E-mail：zhoumanling@cuc.edu.cn；林美秀（2003-），女，本科生，主要从事计算机三维视觉研究。Email：lmeixiu8590@cuc.edu.cn；张佳一（2000-），女，硕士研究生，主要从事设计学-创意媒体设计等。E-mail：beverlyzhang@cuc.edu.cn；王晶*（1980-），女，博士，教授，主要从事语音和音频信号处理、多媒体通信、虚拟现实等研究。Email：wangjing@bit.edu.cn；刘淼（1998），男，博士研究生，主要从事音视频联合学习研究。Email：liumiao424@163.com

可判断面部帧与音频是否同步，通过二维卷积层编码面部与音频信息，使用最大边际损失训练以提升同步判别效果。生成器借鉴LipGAN，包含身份编码器、音频编码器和面部解码器，实现语音到唇形的转换。视觉质量判别器与生成器共同训练，优化生成视频的质量。在这个框架中，本文引入了检索与想象的概念，通过快速学习和慢速学习系统的相互作用能够在不断学习新任务时保留以往的经验。这种机制类似于人类大脑加深过去印象的方式，能够在有限的记忆资源下持续改善模型的性能。

VR/AR-AdaptFace模型的具体流程图如图1所示：

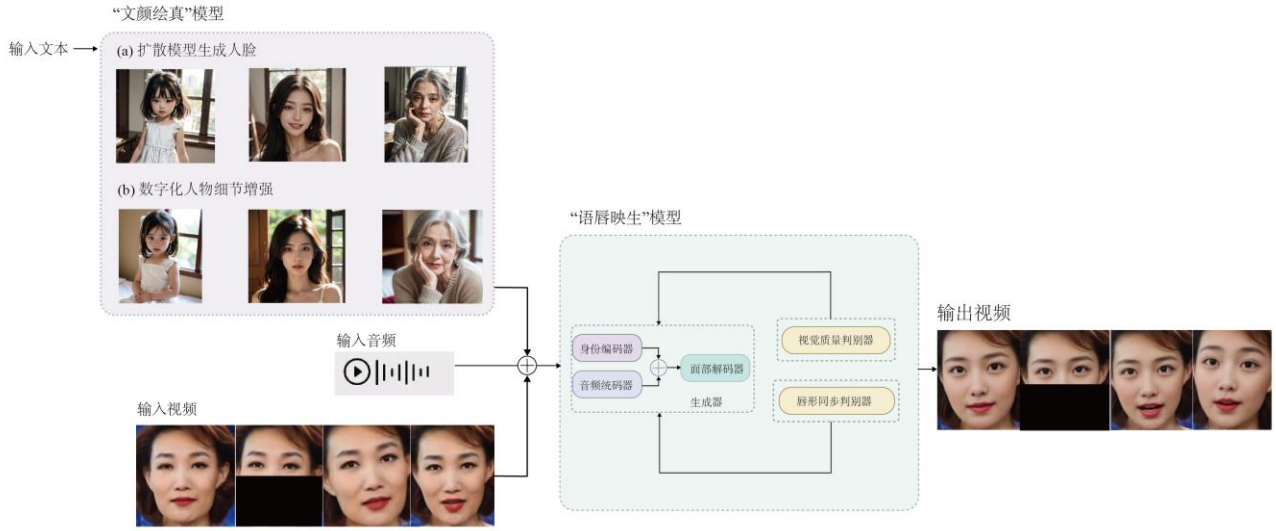

图 1 VR/AR-AdaptFace 模型流程图

2 相关工作

2.1 文生图模型

早期的文生图模型主要依赖于传统的生成模型和特征表示方法。其中，DDPM (denoising diffusion probalistic models)[5]作为扩散模型的先驱，通过前向加噪和反向去噪过程为文本到图像合成任务[6][7]提供了新的思路。然而，DDPM 的生成效率较低，限制了其在实际应用中的广泛使用。为了提高扩散模型的生成效率，研究者们提出了多种改进方法。DDIM(denoising diffusion implicit models)[8]通过更高效的采样策略避免了计算所有去噪步骤，显著提高了生成速度。此外，研究者们还探索了使用 Transformer[9]结构替代 CNN(convolutional neural network)[10]以提高模型的全局建模能力，进一步推动了文生图技术的发展。自编码器（AutoEncoder）[11]作为经典的生成模型，通过编码器和解码器学习数据的低维表示。为了引入离散性，研究者们提出了量化自编码器 VQ-VAE(variational quantized variational autoencoder)[12]，将连续的特征向量编码为离散的表示。VQ-VAE 在文生图任务中得到了广泛应用，为生成高质量的图像提供了可能。VQ-GAN[13]作为 VQ-VAE 的改进版本，在多个方面进行了优化。首先，它使用 Transformer 替代了 pixelCNN(pixel convolutional neural networks)，以捕捉图像中较远像素之间的依赖关系。其次，VQ-GAN 增加了一个 PatchGAN[14]作为判别器，并在训练过程中加入了判别损失，以提高生成图像的质量。最后，VQ-GAN 采用了感知损失替代了传统的 L2 损失，以更好地捕捉图像的高层语义信息。近年来，研究者们开始关注文生图模型的可控性和可解释性。VQGAN-CLIP(vector quantized generative adversarial networks - contrastive language-image pre-training)[15]模型通过结合 CLIP(contrastive language-image pre-training)[16]和 VQ-GAN，能够从复杂的文本提示[17]中生成高质量图像，无需额外训练。这种方法不仅提高了生成图像

的质量，还增强了模型的可控性和可解释性。LDM(Latent Diffusion Models)[18]通过引入隐空间机制和 Cross Attention 机制，使得模型能够在隐空间中捕捉更多的语义信息，实现对生成图像的精细控制。

2.2 唇形合成模型

针对唇形合成的模型诸如：基于 PaddlePaddle 深度学习框架的 PaddleGAN (Paddle Generative Adversarial Networks)广获应用，其在数字人脸、人体姿势与动作表现的高质量生成方面表现出色，同时亦在视频合成与唇形同步方面取得显著进展。进入 2023 年，硅谷的 Twinsync 项目引领了唇形合成领域的新潮流。该项目运用神经网络与先进渲染技术，旨在实现高度逼真的人物视频合成。Twinsync 提出的视频唇形同步算法，融合了神经辐射场（Neural Radiance Fields, NeRF）与网格变形（deform）技术，能从单一图像中精准预测演讲者的面部形态与纹理，进而将其应用于源视频，实现唇形同步。然而，这两项研究虽提升了视频的整体清晰度和匹配度，但在唇形合成准确度方面仍有待加强，存在改进空间。Prajwal 等人提出 Wav2Lip 模型，在结合唇形同步相关领域的研究基础上，一定程度上克服了以往的唇形生成模型的缺点，训练完毕后对任意说话者、语言和视频都可以进行合成，且合成视频的唇形与音频是非常同步的，进一步提升了唇形合成的精度。

本文研究工作主要围绕 LDM 与 Wav2Lip 模型展开，通过对其原理、构造等的深入了解，结合当前领域先进的研究成果，对 LDM 与 Wav2Lip 模型的性能提升及应用创新等问题展开探讨。

3 “文颜绘真”模型

“文颜绘真”模型由两大模块构成：一是基于扩散机制的面部合成模块，二是精细化的人脸数字细节强化模块。该模型首先利用先进的扩散模型技术，构建起人脸的基础形态框架，随后，融入数字化人脸细节增强机制，对生成的人脸进行深度优化，显著增强图像的逼真度与微观纹理表现力，从而实现人脸图像在视觉上的高度真实与自然。

3.1 扩散模型生成人脸

扩散模型作为一种概率生成模型，其核心思想在于通过对从高斯分布中抽取的变量进行渐进式的去噪处理，从而学习和模拟数据的内在分布规律。在本次研究中，本文特别关注那些已经预训练的文本到图像扩散模型 x_θ。该模型的工作原理如下：首先，引入一个初始噪声图 ε，该图基于标准正态分布 N(0, 1)生成；随后，利用文本编码器 Γ 和给定的文本提示 p [19]，构造出一个条件向量 $c=\Gamma(p)$。此条件向量与初始噪声图共同作为模型的输入，经过扩散模型的处理，最终生成目标图像 $x_{gen}=x_\theta(\varepsilon,c)$，其中 x 为真实图像，c 为条件向量。在这个过程中，条件向量扮演着关键的角色，它为模型提供了额外的信息，帮助生成符合预期的真实图像。这种基于文本的条件生成方法为图像生成任务提供了一种全新的范式，使得生成的图像更加丰富多样，并且更贴近真实世界的场景。

根据作者的实践经验，实现最大主题保真度的最佳途径在于对模型的所有层级[20]进行精细调整。这其中，以文本嵌入为基础的微调层尤为关键，然而，这也常常伴随着语言漂移的问题[21]。语言漂移是在语言模型中观察到的一种现象，表现为在大型文本语料库上预训练后，为特定任务进行微调的模型会逐渐丧失语言的语法和语义知识。此外，作者还面临着一个挑战，即输出多样性的潜在降低。文本到图像的扩散模型本质上具备丰富的输出多样性。然而，当针对少量图像进行微调时，本文期望模型能够生成具有新颖

视角、姿势和脸部形态的主题内容。但实践中存在降低输出姿态和主题视图多样性的风险。

为了应对上述挑战，在本文中提出了一种特定于类别的先验保持损失策略，旨在促进输出多样性和缓解语言漂移问题。该方法的核心思想是利用模型自身生成的样本进行监督，从而在少量样本微调时保留先验信息。这样做不仅有助于模型生成多样化的先验类别图像，还能保留关于先验类别的知识，进而与主题实例的知识相结合。

为了解决语言漂移和输出多样性潜在降低的问题，本文在具体实现上对具有随机初始噪声 $z_t \sim N(0,1)$ 和条件向量 c_{pr} 的冻结预训练扩散模型应用采样器生成数据 $x_{pr} = x(z_t, c_{pr})$。在训练流程启动之前，首先利用模型生成一系列聚焦于特定类别的图像样本。这些图像不仅展现了该类别对象的多样化形态，还涵盖了它们所处的丰富环境背景。随后，这些准备好的图像集被用作训练过程中的一种约束机制，旨在预防模型发生语言漂移现象。具体而言，即便在训练过程中模型参数持续更新，作者期望模型在接收到输入文本时，能够维持并增强生成该类别对象及其所处环境多样性的能力，而非局限于生成某种固定模式或特定实例的图像。这一设计确保了模型对输入文本的理解在训练期间保持一致性和稳定性，即模型没有发生语言漂移，仅当输入包含特定标识符的文本时，才会精确导向生成对应主题的图像。

凭借这样的设计，成功实现了从文本到图像的精确转换，同时确保了生成图像的高质量与逼真度。

3.2 数字化人物细节增强

针对当前的合成人脸图像通常存在分辨率较低、细节不足等问题。为了解决这些问题，本文提出了一种用于合成人脸细节增强的框架，旨在解决合成人脸图像中存在的低分辨率和缺乏细节等问题。该框架结合了特征编码、注意力机制和特征重建等关键技术，以生成更加逼真和高分辨率的合成人脸图像。首先，本文利用特征编码模块将输入图像转换为高维特征表示，捕捉到了图像的丰富信息。然后，基于窗口和基于通道的注意力机制有助于集中模型的注意力于特定区域和特定通道，提高了模型对人脸图像中重要特征的捕捉能力。接着，通过特征重建阶段，本文将重点放在重建高分辨率的图像特征上，实现了对细节信息的恢复和增强。最后，图像解码器将重建的高分辨率图像特征转换为最终的合成人脸图像，保持图像的真实性和逼真感。除了上述框架所涵盖的关键技术外，本文的方法还具有可扩展性和适应性。特征编码模块和图像解码器是基于深度学习模型构建的，因此可以轻松地对其进行修改和扩展，以适应不同类型的合成人脸任务，为直播电商领域的合成人脸应用提供了一种强大而灵活的解决方案。

“文颜绘真”模型的具体流程图如图 2 所示：

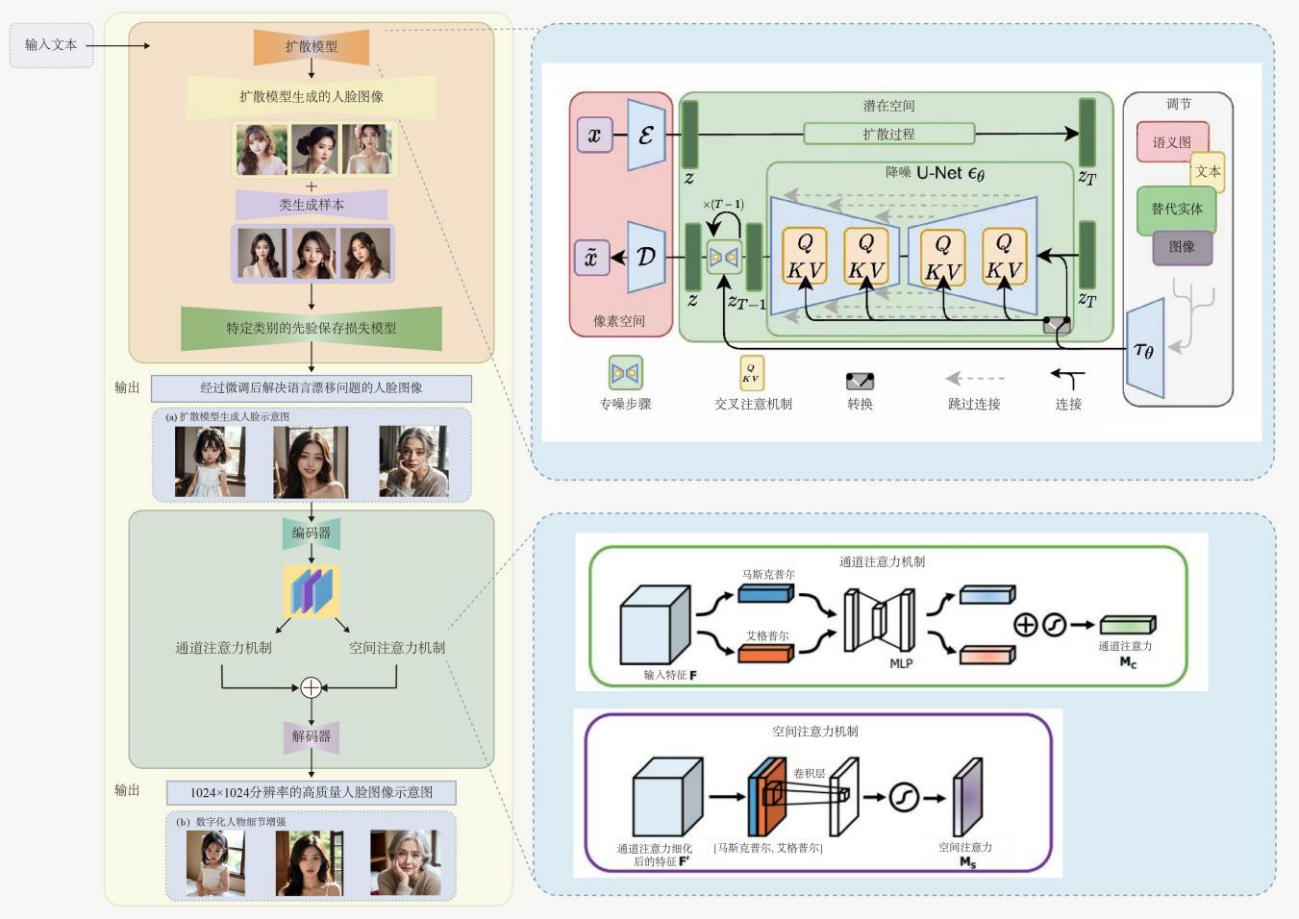

图 2 “文颜绘真”模型流程图

4 “语唇映生”模型

"语唇映生"模型的核心架构涵盖了三个预先训练的组件：唇形同步判别器、生成器以及视觉质量判别器，其详细的构造如图 3 所示。该模型中的唇形同步判别器采用了与 SyncNet 相似的结构设计，它接收连续的面部帧（特别关注下半脸）以及对应的音频片段作为输入，能够精确评估图像帧与音频之间的同步状态，并据此产生判别结果。其内部机制包含精心设计的面部编码器和音频编码器，二者均基于二维卷积层，通过最大化边际损失的训练策略，优化同步与不同步对之间的 L2 距离，从而确保了出色的同步判别效果。在生成器方面，该模型参考了 LipGAN 的设计理念，由身份编码器、音频编码器和面部解码器三个主要部分构成。其中，身份编码器利用残余卷积层结构，对随机选取的参考帧 R 进行有效编码，并与姿态先验 P（即目标面部下半部分的遮挡信息）进行融合。音频编码器则通过二维卷积层对输入的语音段 S 进行编码，随后与人脸特征表示进行集成。最后，面部解码器融合了卷积层与反卷积层，以两个编码器输出的特征图为输入，生成具有逼真唇形变化的视频图像。为了应对生成图像中可能出现的瑕疵或模糊问题，视觉质量判别器与生成器共同参与了训练过程，旨在提升生成视频的整体视觉质量。这一设计确保了即使在复杂多变的生成环境中，模型也能够产生清晰、准确的唇形同步视频。

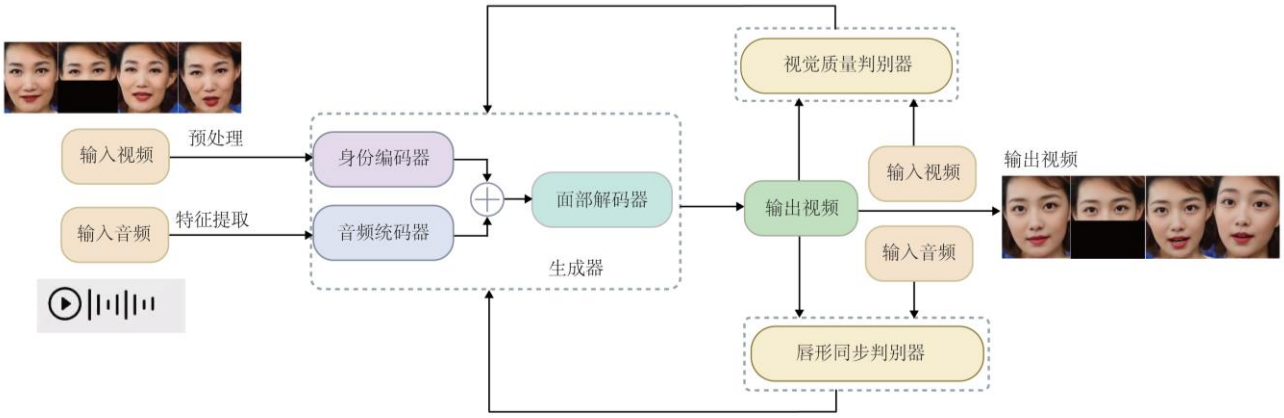

图 3 “语唇映生”模型流程

5 实验

5.1 数据集和实施细节

本文采用 FFHQ 数据集对“文颜绘真”进行评估，这是一个较新的数据集，由 70000 张高质量的人脸图像组成，这些图像经过对齐和裁剪处理，以 1024x1024 的分辨率呈现。为了准确衡量重建质量，需要确保所使用的测试图像在训练过程中未被使用过。因此，作者将 FFHQ 数据集划分为包含 50000 张图像的训练集和包含 20000 张图像的测试集。在 SGI-AE 模型逐步增长到新的分辨率水平时，为了有效促进模型的适应性和稳定性，作者在过渡阶段设计了训练方案。具体而言，这一阶段使用了 500k 的训练样本，这些样本并非简单地复制原始训练集中的 50,000 张图像，而是运用数据增强技术，生成了丰富多样的“新”图像。这些增强后的图像不仅保留了原始图像的核心特征，还引入了额外的变异性和复杂性，有助于模型学习实现更加全面和鲁棒的特征表示。此外，为了进一步增强训练过程的稳定性，作者还额外使用了 500k 的样本。这里的额外样本并非指单独准备的一套数据集，而是在整个训练过程中，数据被动态地分成多个批次进行处理。虽然每个批次可能只包含原始训练集或其增强版本中的一小部分图像，但整个训练过程会多次遍历这些图像，确保模型能够充分学习和吸收数据中的信息。通过这种方式，即使原始训练集规模有限，也能通过数据增强和分批处理的策略，实现高效且稳定的模型训练。一旦模型达到 1024x1024 的最大分辨率，继续利用 1M 张的图像进行训练。因此，整个训练过程中总共使用了 10M 张的图像样本。

5.2 “文颜绘真”模型的实验结果

首先，本文创新性地引入了一个特征编码模块，该模块能够高效地将输入的图像数据转换为高维度的特征向量，这一过程深刻挖掘并捕捉了图像中蕴含的丰富信息细节。随后，为了进一步优化模型对关键信息的识别能力，本文集成了两种先进的注意力机制：基于窗口的注意力机制与基于通道的注意力机制。这两种机制协同工作，引导模型更加聚焦于图像中的特定区域与特定通道，显著增强了模型对人脸图像核心特征的提取精度。紧接着，设计了一个精心构建的特征重建阶段，该阶段的核心目标聚焦于高精度地重建图像的高分辨率特征。通过这一环节，不仅实现了对图像细节信息的有效恢复，还进一步增强了这些细节的清晰度与表现力，为后续的图像处理奠定了坚实的基础。最终，利用一个高效的图像解码器作为桥梁，将经过精心重建的高分辨率图像特征无缝转换为最终的合成人脸图像。这一转换过程不仅确保了图像内容的完整性，还极大地保留了图像的真实质感与高度逼真的视觉效果，为用户提供了更为生动、自然的人脸图像呈现。下图中，样例一展示了利用 Diffusion 模型生成的人脸原图像经过数字化细节增强后的效果，而样例二和样例三则是基于 Diffusion 模型生成的人脸，在经过数字化细节增强处理后，再通过文本输入实现

场景与服饰的更换，从而呈现出不同的结果。

图 4 人脸细节增强扩散模型效果图

表 1 部分换脸模型的评估指标得分

模型名称	FID	PPL full
PGGAN	9.12	245.2
PIONEER	37.93	162.9
ALAE	20.45	43.2
“文颜绘真”	15.32	35.2

在本文的研究中，设计了一个基于类似生成对抗网络（GAN）的架构的换脸模型。表 1 中对比了目前几种主流的换脸模型，“文颜绘真”模型在 FID（Fréchet Inception Distance）和 PPL（Perceptual Path Length）指标下表现出色。具体来说，“文颜绘真”在 FID 值方面达到了 15.32，PPL 值方面达到了 35.2。这两个指标都是评估生成图像质量和多样性的重要标准。

首先，低 FID 值（15.32）显示出“文颜绘真”生成的人脸图像与真实图像之间的相似度较高，接近甚至优于部分主流模型，如 PGGAN (Progressive Growing of GANs)。其次，“文颜绘真”模型展现出了较低的

PPL 值（35.2），这表明在潜在空间中生成的图像变化更为平滑自然。相比之下，“文颜绘真”在质量和多样性方面都有明显的优势。在综合表现方面，“文颜绘真”采用了类似生成对抗网络（GAN）的架构，在生成图像的质量和多样性方面表现出色。它能够生成与真实图像相似度高、变化平滑自然的高质量人脸图像，这对于换脸技术来说至关重要。

本文的研究可能得益于采用了先进的生成网络结构和训练技术，能够更好地学习人脸特征的表示，并在生成过程中保持图像的连续性和自然性。这可能包括对抗训练中的改进、潜在空间的优化或者更有效的损失函数设计。

5.3 “语唇映生”模型的实验结果

Wav2Lip 模型，作为生成式对抗网络（GAN）架构的一项创新应用，虽然在视觉效果上显著展示了其优势，如维持视频人物面部五官的高清晰度，但在特定方面仍面临挑战。尽管该模型在唇形修改后确保了五官的清晰可辨，却在下巴区域偶现不明遮挡的问题，该问题轻微地影响了整体的视觉流畅度。更为显著的是，在处理中文语音输入时，Wav2Lip 模型生成的唇形动作变化频率异常偏高，这一异常现象与中文语境下自然唇形的变化规律大相径庭，如图 5 所示，这种差异不仅影响了口型同步的自然度，也削弱了模型在中文环境下的适用性和真实性。

在深入研究 Wav2Lip 模型的测试表现后，本文提出了一种名为 Sync_fromscratch+Wav2Lip_fromscratch（SSWS）的改良方法，旨在优化 Wav2Lip 模型在中文语境下的合成效果。

为了准确评估唇形同步的精准度，本文采用了两个关键指标：唇形同步误差距离（Lip Sync Error - Distance, LSE-D）和唇形同步误差置信度（Lip Sync Error - Confidence, LSE-C）。LSE-D 旨在量化生成的唇形序列与对应音频信号之间的时间偏差程度。其值越小，表明唇形与音频的同步性越佳，即唇形变化与语音内容更为吻合。而 LSE-C 则基于模型对唇形同步预测的自信程度来计算，较高的 LSE-C 值意味着音频与视频之间的关联性更为紧密。

本次实验所依赖的数据集来自浙江大学视觉智能和模式分析（VIPA）团队精心构建的 CMLR 数据集。该数据集专注于中文普通话唇读任务，收录了自 2009 年 6 月至 2018 年 6 月间的新闻联播视频，共计包含由多位主持人表述的 102,076 条句子。每条句子最多包含 29 个汉字，且不包含英文字母、阿拉伯数字和罕见标点符号。

图 5 Wav2Lip 原模型合成效果示意

针对处理中文语音输入时，Wav2Lip 模型合成的唇形动作变化频率异常频繁的问题，本文将对使用英文数据集训练的原始 Wav2Lip 模型与 SSWS 以及其他现有的唇形合成模型进行深入的比较分析。SSWS 模型特别采用 CMLR 数据集从头开始训练唇形同步判别器，随后基于这一训练完成的判别器，再次从零开始训练生成器及视觉质量判别器。

表 2 SSWS 各组数据对比

步数	绝对损失	同步损失	感知损失
9000	0.021	0.096	0.69
21000	0.016	0.074	0.71
33000	0.015	0.073	0.70

在训练唇形同步判别器时，本文基于 CMLR 数据集的 S2 部分进行。训练过程中，判别器的损失值呈现波动下降趋势，并在经过 51000 步训练后，验证集上的平均损失值降低至 0.278，这显著表明模型的性能得到了提升并逐渐趋于稳定。接下来，利用这一训练好的唇形同步判别器来训练生成器及视觉质量判别器。在训练过程中，生成器在验证集上的平均损失值如表 2 所示，其中感知损失保持稳定，这显示了模型在保持生成图像感知质量方面的一致性。同时，平均绝对损失从 0.021 降至 0.015，这反映了生成唇形图像准确性的显著提高。而同步损失在训练初期波动下降后，于 33,000 步后趋于平稳，这进一步验证了唇形同步判别器在训练过程中的有效性和稳定性。

图 6 SSWS 合成效果示意

图 6 展示了 SSWS 进行唇形合成的效果。从左至右，第一幅是输入的当前帧画面，第二幅为下一帧的姿势先验（即下半部分被遮盖的目标人脸），第三幅则是 SSWS 模型根据音频特征合成的预测画面，最后一幅是实际目标人脸（即真实的下一帧人脸画面）。通过对比可见，SSWS 模型合成的唇形与目标人脸图像的相似度极高，且过渡自然流畅。

5.3 VR/AR-AdaptFace 的效果

模型主要由三个核心部分构成：a.利用扩散模型生成人脸，该步骤确保了人脸的基本轮廓和结构的准确性；b.通过数字化人脸细节增强模型，进一步提升了生成人脸的清晰度和细节表现使得人脸图像更加逼真，所得到的数字人脸效果图如图 7 所示；c.采用了“语唇映生”唇形生成模型，将生成的数字人脸与视频中的原始人脸（如图 8 所示）进行替换，并同时实现唇形的精准合成后的示意图为图 9 所示。具体示例如下所示：

图 7“文颜绘真”模型生成的数字人脸

图 8 未使用数字人脸换脸技术的效果示意图

图 9 使用数字人脸换脸与合成唇形后的效果示意图

6 结论

本文提出了一个名为 VR/AR-AdaptFace 的面向虚拟现实（VR）与增强现实（AR）的自适应多模态面部替换模型，针对 VR/AR 领域中深度学习图像生成技术面临的挑战，特别是在虚拟角色的人脸生成和唇形同步方面存在的问题，进行了深入研究与改进。该模型由“文颜绘真”人脸合成模型和“语唇映生”唇形生成模型两大核心部分组成，显著提升了虚拟角色在 VR/AR 环境中的真实感和沉浸感。本模型经过微调后能够更精确地模拟人体器官，更准确地捕捉虚拟角色的面部特征，从而为用户提供更加自然和谐的 VR/AR 体验。然而，在数据预处理过程中，由于训练数据的过滤导致分布呈现出不均衡性，进而使得模型在应对训练集中较为稀少的特征时，其生成的虚拟角色在 VR/AR 环境中的表现稍显不自然。针对这一问题，作者计划通过扩大训练数据集的规模，以及优化数据分布来进一步提升模型在 VR/AR 应用中的性能。展望未来，作者将继续致力于克服这些局限性，为 VR/AR 领域带来更加逼真、生动的虚拟角色体验。

本文深度合成的内容符合社会主义核心价值观，合成的内容是真实的新闻播报内容而不是虚假信息；符合国家网信办出台的《生成式人工智能服务管理办法》，承诺仅用于科研，不提供其他服务。

参考文献（References）：

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.10684–10695.

[2] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 22500–22510.

[3] Mamoru Komachi, Taku Kudo, Masashi Shimbo, and Yuji Matsumoto. 2008. Graph-based analysis of semantic drift in Espresso-like bootstrapping algorithms. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. 1011–1020.

[4] Yuchen Lu, Soumye Singhal, Florian Strub, Aaron Courville, and Olivier Pietquin. 2020. Countering language drift with seeded iterated learning. In International Conference on Machine Learning. PMLR, 6437–6447.

[5] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. 2018. St-gan: Spatial transformer generative adversarial networks for image compositing. In Proceedings of the IEEE conference on computer vision and pattern recognition.

9455 – 9464.

[6] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. 2019. Gp-gan: To wards realistic high-resolution image blending. In Proceedings of the 27th ACM international conference on multimedia. 2487 – 2495.

[7] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020).

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems 30 (2017).

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 4700 – 4708.

[10] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. Neural computation 18, 7 (2006), 1527 – 1554.

[11] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. Advances in neural information processing systems 30 (2017).

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 12873 – 12883.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1125 – 1134.

[14] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Halla han, Louis Castricato, and Edward Raff. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. In European Conference on Computer Vision. Springer, 88 – 105.

[15] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems 35 (2022), 36479 – 36494.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning. PMLR, 8748 – 8763.

[17] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator ar chitecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 4401 – 4410.

[18] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In Proceedings of the 28th ACM international conference on multimedia. 484 – 492.

[19] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 1, 2 (2022), 3.

[20] Jason Lee, Kyunghyun Cho, and Douwe Kiela. 2019. Countering language drift via visual grounding. arXiv preprint arXiv:1909.04499 (2019).

[21] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sen gupta, and Anil A Bharath. 2018. Generative adversarial networks: An overview. IEEE signal processing magazine 35, 1 (2018), 53 – 65.