

基于多级残差映射器的文本驱动人脸图像生成和编辑^{*}

李宗霖¹, 张盛平¹, 刘杨¹, 张兆心¹, 张维刚¹, 黄庆明²



¹(哈尔滨工业大学 计算机科学与技术学院, 山东 威海 264209)

²(中国科学院大学 计算机科学与技术学院, 北京 100049)

通信作者: 张盛平, E-mail: s.zhang@hit.edu.cn

摘要: 尽管生成对抗网络在人脸图像生成和编辑领域取得了巨大的成功, 但在其潜在编码空间中寻找可以操作人脸语义属性的方向仍然是计算机视觉的一大挑战, 这一挑战的实现需要大量标记数据不断进行网络调优, 而搜集、标注类似数据存在诸多难点, 比如较高的技术门槛以及大量的人工成本。最近的一些工作都在试图借助预训练模型来克服标记数据短缺的问题。虽然这种做法已经被验证能够完成上述任务, 但在操作的准确性和结果的真实性上都无法满足真实人脸编辑场景的需求。借助对比语言-图像预训练模型(CLIP)的图像文本联合表示能力将图像和文本内容编码在一个共享的潜在编码空间中, 借助于精心设计的网络结构和损失函数, 所提框架可以精准识别相关面部属性并学习一个多级残差映射网络, 所提网络可根据图像和文本内容编码预测潜在编码残差, 再借助图像生成预训练模型 StyleGAN2 完成高质量的人脸图像生成和编辑任务。大量实验也证明了所提方法在操作准确性、视觉真实性和无关属性保留方面的优异表现。

关键词: 多模态学习; 预训练模型; 人脸图像生成; 人脸图像编辑; 对抗生成网络

中图法分类号: TP391

中文引用格式: 李宗霖, 张盛平, 刘杨, 张兆心, 张维刚, 黄庆明. 基于多级残差映射器的文本驱动人脸图像生成和编辑. 软件学报, 2023, 34(5): 2101–2115. <http://www.jos.org.cn/1000-9825/6767.htm>

英文引用格式: Li ZL, Zhang SP, Liu Y, Zhang ZX, Zhang WG, Huang QM. Text-driven Face Image Generation and Manipulation via Multi-level Residual Mapper. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2101–2115 (in Chinese). <http://www.jos.org.cn/1000-9825/6767.htm>

Text-driven Face Image Generation and Manipulation via Multi-level Residual Mapper

LI Zong-Lin¹, ZHANG Sheng-Ping¹, LIU Yang¹, ZHANG Zhao-Xin¹, ZHANG Wei-Gang¹, HUANG Qing-Ming²

¹(School of Computer Science and Technology, Harbin Institute of Technology, Weihai 264209, China)

²(School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Although generative adversarial networks (GANs) have achieved great success in face image generation and manipulation, discovering meaningful directions in the latent encoding space of GANs to manipulate semantic attributes of faces is a great challenge in computer vision. The solution to this challenge requires a large amount of labeled data and several hours of network fine-tuning. However, many difficulties are confronted in the collection and annotation of similar data, such as great technical barriers and high labor costs. Recent studies have been attempting to overcome the problem of lacking labeled data by pre-trained models. Such efforts are proved capable of accomplishing the above task, but the accuracy of the manipulation and the authenticity of the results cannot meet the needs of real face editing scenarios. To address these problems, this study encodes the image and text descriptions into a shared latent encoding space by leveraging the joint representation capability of contrastive language-image pre-training (CLIP). With carefully designed network structures and loss functions, the proposed framework can accurately recognize relevant face attributes and learn a residual mapping network. The network can predict the latent code residuals according to image and text description codes and perform high-quality image

* 基金项目: 国家自然科学基金 (61872112, 61976069)

本文由“融合预训练技术的多模态学习研究”专题特约编辑宋雪萌副教授、聂礼强教授、申恒涛教授、田奇教授、黄华教授推荐。

收稿时间: 2022-04-14; 修改时间: 2022-05-29, 2022-08-03; 采用时间: 2022-08-24; jos 在线出版时间: 2022-09-20

CNKI 网络首发时间: 2023-03-17

generation and manipulation by the pre-trained model StyleGAN2. Extensive experiments demonstrate the superiority of the proposed approach in terms of manipulation accuracy, visual realism, and irrelevant attribute preservation.

Key words: multimodal learning; pre-trained model; face image generation; face image manipulation; generative adversarial network (GAN)

作为计算机视觉领域中艰难但有意义的任务,人脸图像的生成和编辑引起了研究者的广泛兴趣。随着生成对抗网络(generative adversarial network, GAN)^[1]尤其是StyleGAN^[2,3]的发展,一个全新的图像生成范式实现了高质量的图像生成。为了使这个过程更易于交互,大量研究创新性地提出了基于一系列条件输入快速实现图像生成和编辑的设想,例如人脸草图^[4,5]、语义标签^[6-8]或文本描述^[9-11]等。其中基于文本驱动的人脸图像生成和编辑技术是在联合自然语言处理和生成对抗网络的基础上发展起来的,利用自然语言处理从输入文本中提取语义表示向量,并将其作为生成对抗网络模型的监督信号生成图像,这不同于传统的图像生成技术,生成对抗网络在训练过程中没有复杂的变分下界也不需要使用马尔可夫链方法^[12]以及各种近似推导,大大改善了生成式模型的训练难度和训练效率^[13]。由于自然语言和图像分属两个不同的模态,一段文本描述可以对应图像中的许多位置,一张图像也可以对应许多种文本描述,因此生成高质量且符合文本语义的图像具有非常高的研究和应用价值^[14]。

现阶段基于文本驱动的人脸图像生成和编辑方法大都是采用堆叠多个生成对抗网络的方法,通过融合每个阶段生成的不同分辨率的图像,模型最终输出基于文本描述生成的图像^[15]。然而这种架构带来的缺点是训练过程非常不稳定,生成的图像更像是各种文字属性的堆叠,缺乏真实性,特别是当需要基于文本内容生成复杂图像时,判别器无法提供充分的监督信息。具体来说,此类方法通常有3个阶段,每个阶段包含1个生成器和1个判别器。3个阶段同时进行训练,依次生成3个不同尺度的图像,比如 64×64 、 128×128 和 256×256 。在这个过程中带有粗略形状和颜色的初始图像将被细化为更加细致的基于文本描述的图像。然而,此类方法中常用的图像-文本匹配模型无法利用跨模态的属性信息,导致从文本生成图像时属性匹配不成功,或在处理图像时对不相关的属性产生不期望的操作。本文采用的是将预训练StyleGAN2模型^[3]和对比语言-图像预训练模型CLIP^[16]结合作为基础框架的方法。其中StyleGAN2相比于上一代模型,它着重于修复特征伪影,并进一步提高了生成图像的质量,是目前最先进的高分辨率图像合成方法。CLIP是一个由图像编码器和文本编码器构成的预训练模型,通过对4亿个图像-文本对的联合训练,他们可以精确测量输入图像和文本描述之间的语义相似性。

对于文本驱动人脸图像生成和编辑技术,该问题的主要难点有:(1)自然语言处理。虽然近年来自然语言处理技术在很多应用中取得了重大突破,但由于自然语言文本中广泛存在歧义性和多义性,实现自然语言理解仍存在诸多难点。自然语言的形式(字符串)与其意义之间存在多对多的关系,其中由单词可组成词组,由词组可组成句子,一个英文文本或一个词组可能有多个含义。反过来,一个相同或相近的意义同样可以用多个英文文本或多个词组来表示。为了消解歧义性和多义性,自然语言处理模型需要极其大量的知识进行推理。如何将这些知识较完整地收集和整理出来;又如何找到合适的形式,将它们存入模型中;以及如何有效地利用它们来消除歧义,都是工作量极大且十分困难的工作。因此,将人类语言文字编码成高质量的语义向量实现不同程度的语义理解仍是一个技术挑战。(2)图像合成。图像合成是计算机视觉、计算机图形学等领域的重要研究方向,目前已经广泛应用于由文字生成图像、不同模态间的图像转化、图像修复、编辑、去模糊、超分辨率等任务。当前图像合成的主要难点在于如何保证图像的真实性、多样性和输入条件一致性。近年来生成对抗网络的出现虽提升了合成图像的真实性,但其本身存在的训练不稳定、生成样本质量差、评价体系不够健全、可解释性差等问题是目前生成对抗网络研究的重点和难点。考虑到StyleGAN可以通过无监督式地对图像的高层语义属性做一定解耦分离,例如人脸图像的姿势和身份、所生成图像的随机变化如雀斑和头发等,也可以做到一定程度上的控制图像合成。因此,如何基于StyleGAN语义丰富、解耦性能强的潜在空间实现人脸图像编辑仍是一个技术挑战。(3)文本-图像一致性。基于文本生成或编辑图像最棘手的问题是文本和图像两种不同模态的信息如何实现精准匹配。文本和图像通常存在一对多的关系,也即一段文本描述可以对应多个不同的图像位置,一张图像也可以对应多个文本描述。现在的做法通常是将文本表示向量嵌入到条件生成对抗网络中,作为图像生成过程的监督信号,比如将文本表示通过词嵌入和句子嵌入的方式实现对于生成过程的语义监督,通过比对嵌入向量的差异对语义进行区分,但是在生成对抗网络中

缺乏文本嵌入向量和视觉特征之间的有效匹配,使得通过文本作为条件生成的图像难以保证视觉语义的连贯性,特别是当使用多个生成对抗网络作为模型架构的时候,其生成的图像的质量往往取决于生成对抗网络的训练稳定性而文本嵌入方式对于图像生成效果的提升不明显。因此,如何实现图像-文本语义一致性和如何确保网络训练收敛从而生成高质量的图像是当前文本驱动图像生成和编辑任务面临的严峻挑战。

针对上述问题,我们借助预训练 StyleGAN2 模型的图像生成能力和 CLIP 模型的跨模态语言-图像表达能力,为不同语义级别的特征学习了一个单独的残差映射器将输入条件映射为相应的潜在编码变化,最终实现文本驱动人脸图像生成和编辑功能。更具体地说,基于给定要编辑的真实图像和文本描述,我们首先使用 StyleGAN2 模型反演方法获得其潜在编码,再将图像和文本内容输入 CLIP 模型获得人脸属性编码和属性状态编码,此后借鉴 StyleCLIP 的网络设计,使用多级残差映射器网络将不同级别的潜在编码迁移至共享潜在空间,通过比对差异预测潜在编码残差,将修改后的潜在编码将被反馈到大规模人脸数据集上预先训练过的 StyleGAN2 模型中,以获得目标编辑结果。其中最为关键的部分是学习一个残差映射器网络,将输入条件映射到相应的潜在编码变化中。与 StyleCLIP 不同的是,我们探索了 CLIP 超越测量图像文本相似性的潜力,以及一些新的设计: 1) 共享条件编码。为了将文本和图像条件统一到同一领域,我们利用 CLIP 的文本编码器和图像编码器分别提取它们的潜在编码并转换到同一个 StyleGAN 潜在空间下,作为残差映射器网络的输入; 2) 分离信息注入。我们明确地将人脸属性和属性状态信息分开,并将它们输入与它们的语义级别对应的不同的子残差映射器。这有助于我们的方法实现解耦编辑; 3) 分级调制模块。设计了一个条件调制模块通过预测潜在编码残差来实现对潜在编码的直接控制,提高方法的操作能力。

本文第 1 节介绍文本驱动人脸图像生成和编辑的相关方法和研究现状。第 2 节介绍本文所需的基础知识,包括语言-图像联合表达和预训练 StyleGAN2 模型潜在编码操作。第 3 节介绍本文构建的基于预训练图像生成模型 StyleGAN2 和对比语言-图像模型 CLIP 的文本驱动人脸图像生成和编辑方法。第 4 节通过定性、定量的对比试验和关键部件的消融实验来证明提出方法的有效性。第 5 节总结全文。

1 相关工作及研究现状

近年来,由于强大的特征学习和特征表达能力,生成对抗网络被广泛用于各种视觉任务,包括图像生成^[10,17,18]、图像编辑^[9,19]等。为了实现对结果的控制,一些方法创新性地引入了各种用户定义的引导条件,例如手绘草图^[4,5]、语义标签^[6,7]或文本描述^[9,10,15,20,21]等。其中基于文本的图像生成方法大致分为两类。第 1 类由 1 个生成器和 1 个判别器直接从文本生成图像。例如, Reed 等人^[22]使用条件生成对抗网络基于给定的文本描述生成图像。Tao 等人^[23]提出了一种简化的主干结构,可直接通过 Wasserstein 距离将文本信息融合到视觉特征图中,提高图像质量和文本图像的一致性。尽管该方法直接有效,但在某些情况下,单阶段模型在照片真实性和文本相关性方面都无法满足真实场景的需要。因此,另一类研究方向采用多阶段的方式进行基于文本的图像生成。Zhang 等人^[24]通过不断细化草图并进行叠加,从文本描述中生成满足文本描述的图像。Zhang 等人^[25]进一步提出了一种三阶段架构,该架构堆叠了多个生成器和判别器,以多尺度的方式逐步生成满足文本描述的图像。Xu 等人^[10]从两个方面改进 Zhang 等人^[25]的工作。首先,他们引入注意机制来探索细粒度文本和图像表示。其次,他们提出了一个深层注意力多模态相似模型(DAMSM)来计算生成的图像和句子之间的相似性。后续的研究基本上遵循了 Xu 等人^[10]提出的框架,通过引入不同的机制,如自注意机制^[26]或动态记忆模型^[27],提出了几种变体。然而,多阶段框架产生的结果看起来像是来自不同图像尺度的可视化属性的简单组合,缺乏真实性,难以满足真实场景的需求。

与文本到图像生成类似,使用文本编辑图像的目标是生成文本描述包含视觉属性的结果。不同的是,编辑后的结果只应更改文本中提及需要调整的部分,保留原始图像中与文本描述无关的内容。例如, Dong 等人^[28]提出了一种编码器-解码器结构,根据给定文本修改图像。Nam 等人^[9]通过引入文本自适应判别器,将不同的视觉效果分离开,该判别器可以为生成器提供更好的训练反馈。Li 等人^[29]介绍了一个多阶段网络,该网络带有一个新颖的文本图像组合模块,可以实现基于文本描述产生图像。与文本图像生成类似,性能最好的基于文本的图像编辑方法大多数也是使用多阶段框架进行集联。与现有的所有方法不同,我们提出了一种新的框架,该框架将文本引导的图像生

成和操作方法统一起来,过程中无需多阶段处理以及复杂的机制,直接生成满足文本描述的高分辨率图像。

尽管上述方法均一定程度地实现了文本驱动人脸图像生成和编辑的任务,但如何控制生成对抗网络实现更加准确地操作仍然是一个活跃的研究问题。之前关于控制生成过程的研究表明,通过训练条件生成对抗网络,可以生成属于特定类别或具有特定属性的图像^[30]。然而,条件生成模型需要为每个目标属性提供大量标记数据。Chen 等人^[19]提出的 InfoGAN 旨在生成一个分离的潜在空间,其中每个潜在维度控制一个特定属性。然而,这些方法仅提供取决于可用监督信息的粒度的有限控制。

针对这些问题并以可控生成为目标的最新研究包括简单的方法,如修改图像的潜在编码,以及更复杂的方法,如在预先训练的对抗生成网络模型中搜索方向和插值潜在向量。利用潜在空间实现图像编辑的方法可以分为两大类:有监督的方法和无监督的方法。监督方法通常受益于预先训练的属性分类器,这些分类器通过在潜在空间中引导优化过程发现更有意义的方向,或者使用标记数据来训练新的分类器,这些分类器直接用于学习期望的操作。无监督方法是在不依赖成对图像-文本数据集的前提下在潜在空间中找到有意义的方向。Härkönen 等人^[31]提出将主成分分析(PCA)应用于 BigGAN^[32]和 StyleGAN2 模型中间层的随机抽样潜在向量。Shen 等人^[33]以闭环的形式优化了对抗生成网络的中间权重矩阵。Yüksel 等人^[34]提出了一种对比学习方法,以发现可转移到不同类别的指令。

无监督方法的目标是在潜在空间中找到能够实现属性解耦的操作方向。这个过程中发现了大量针对域独立且可解释的方向,如放大、旋转和平移等,通过使用找到的方向一定程度地修改潜在编码,增强或消除生成图像中的目标属性同时,难免会对其他属性也造成一定的影响,无法实现面部属性的完全解耦。一些最近的工作利用 StyleGAN2 的风格空间来解耦属性以处理粗糙(如性别、身份)和精细(如发型、眼睛)的视觉特征。这些图像处理方法的前提是能够找到一个潜在编码准确地重构输入图像,以便直接对真实图像执行编辑操作。为了执行文本驱动的图像编辑,TediGAN^[35]通过训练编码器将图像和文本映射到 StyleGAN2 的潜在空间,根据文本内容执行样式混合以生成相应的图像。考虑到 CLIP 是一个由图像编码器和文本编码器构成的预训练模型,通过对 4 亿个图像文本对的联合训练,他们可以测量输入图像和文本描述之间的语义相似性。StyleCLIP^[36]将 CLIP 强大的图像文本表示功能作为损失监控,使得生成结果与文本内容尽可能地匹配。

2 基础知识

本文所提方法主要基于图像-文本联合模态、StyleGAN2 模型反演和 StyleGAN2 潜在编码操作,下面就相关概念和基本知识予以介绍。

2.1 图像-文本联合模态

实现文本驱动图像生成和编辑的一个关键前提是将视觉属性与相应的文本内容进行匹配。为了实现视觉与语义之间的对齐,目前的方法是利用判别器获取明确的单词级训练反馈以及一系列关于图像-文本相似度计算,其目的在于利用匹配关系在文本和图像之间建立紧密的对应关系,从而为生成器提供监督信号。根据表示粒度,大多数方法可分为两个分支深层架构,即全局或局部表示。第 1 类是利用深度神经网络来提取两种模式的全局特征,并计算它们的相似性。第 2 类是进行实例级别的图像文本匹配,学习单词和图像区域之间的对应关系。

第 1 类中,许多工作如基于语言的图像检索^[37]、图像字幕^[38]和视觉问答^[39,40]等都针对语言图像联合表达进行了相应的学习。随着 BERT^[41]在各种语言任务中的成功,许多方法转向使用 Transformer^[42]来学习联合表达。而对比语言图像预训练(CLIP)的出现开拓了一个全新的多模态潜在编码空间,该空间可直接用于估计文本和图像之间的语义相似性。CLIP 模型由 OpenAI 提出,其训练数据集涵盖了超过 4 亿的图像-文本对,可同时对图像和文本的进行编码,实现在同一潜在空间下的语言图像联合表达。CLIP 在多模态任务上取得了非常好的效果,例如图像检索^[37],地理定位^[43],视频动作识别^[44]等,而且在很多任务上仅仅通过无监督学习就可以得到和主流的有监督算法接近的效果。CLIP 的思想非常简单,但它仅通过如此简单的算法也达到了非常好的效果,这也证明了多模态模型强大的发展潜力。

第 2 类中,一些研究转向学习显著对象的对应关系上,这些方法学习了基于对象共现的对应关系,并在语言图

像匹配方面取得了很大进展。但是,这类方法仅学习粗略的对应关系,因为它们主要依赖于显著对象的对应关系,而忽略其余非显著对象在过程中的作用及状态,尤其是场景信息的丢失,导致无法实现细粒度的语言图像联合表达。

2.2 StyleGAN2 模型反演

随着生成对抗网络的快速发展,许多方法都尝试去理解和控制生成模型中的潜在空间,因此,生成对抗网络的模型反演吸引了大量的关注。模型反演旨在获取给定目标的潜在空间编码,使得预训练的生成对抗网络能够重建给定目标,从而试图理解和控制潜在编码。考虑到 StyleGAN2 模型具有极高的图像生成质量和语义丰富的潜在空间,许多工作开始试图对 StyleGAN2 模型进行反演操作^[35,36,45,46],其目的在于获取给定图像的潜在编码,试图理解潜在编码是否与人脸的语义属性相关从而通过修改潜在编码实现不同程度的人脸操作,增强 StyleGAN2 模型的可解释性以实现更多复杂的人脸生成和编辑任务。因此,高质量的模型反演能力对于人脸图像生成和编辑技术起着至关重要的作用,主要体现在其在重建性和可编辑性上的优势:首先,生成器可使用从模型反演中获得的潜在编码正确地重建输入图像,验证了潜在编码在保留语义属性方面的优势,为之后的属性解耦操作提供了基础;其次,借助 StyleGAN2 潜在空间中语义属性丰富的特点,可通过修改潜在编码实现对于输入图像的编辑目的,以实现不同程度的人脸图像操作。

通常情况下,反演模型的训练方式分为以下 3 种:(1)基于大量数据直接优化预测的潜在编码,使其与输入图像潜在编码的误差尽可能小;(2)基于大量数据训练一个编码器从而学习一个映射方法,直接将图像转变为对应生成器潜在空间下的编码,使得生成图像与输入图像尽可能相似;(3)结合上述两种方法。上述方法在尽可能减少失真细节方面优势明显,但都需要相当长的时间训练优化器或编码器,并且可编辑的能力较差。

2.3 StyleGAN2 潜在编码操作

StyleGAN2 中包含了多个类型的潜在空间,比如 Z 、 W 和 $W+$ 空间。具体而言,让一个生成器 G 充当映射函数 $G: Z \rightarrow X$,其中 X 是目标图像域。潜在编码 $z \in Z$ 来自先验分布 $p(z)$,通常选择为高斯分布。使用由 8 个全连接层组成的映射函数,将 z 向量变换到一个中间的潜在空间 W 。 $W+$ 空间是 W 空间的一个扩展版本,它在合成网络的每一层上使用不同的潜在向量。目前已经证明, $W+$ 空间比 W 空间具有更加明显的解耦属性,因此更常用于模型反演中。

借助每个潜在空间的不同特点,实现可解释的生成对抗网络结构设计且便于对生成图像的可控操作成为可能,因此如何合理利用预训练生成器的潜在编码引起了相关研究者的强烈关注^[35,47,36]。有些方法^[48-52]以学习端到端的方式对图像进行处理,通过训练一个网络,将给定图像编码为目标图像的潜在表示。其他方法是找到潜在空间中的操作方向并进行遍历找到所需的操作方向。这些方法可以分为:使用图像注释来寻找有意义的潜在路径的方法^[53,54],和在没有监督的情况下找到有意义的方向,并要求对每个方向进行注释^[32,34,55,56]。虽然大多数作品在 W 空间中进行图像处理,但 Wu 等人^[48]提出使用样式空间,并表明基于 $W+$ 空间的操作比 W 空间更易实现面部属性解耦。我们的潜在优化器 $W+$ 空间中工作,操作直接来自文本输入,我们唯一的监督来源是预训练的 CLIP 模型。由于 CLIP 是在数以亿计的文本-图像对上训练的,可以在许多领域中进行使用,使得我们的方法不会被现有图像-文本对数据集中的预设操作所局限。

3 基于多级残差映射器的文本驱动人脸图像生成和编辑方法

受 StyleCLIP 的启发,我们利用预训练 CLIP 模型的语言图像联合表达能力和预训练 StyleGAN2 模型的图像生成能力,配合 3 个处理不同级别语义信息的残差映射器来获取潜在编码的残差量,从而实现基于文本内容的人脸图像生成和编辑。更具体地说,给定要编辑的真实图像或基于高斯分布的随机向量,我们首先使用 StyleGAN2 反演方法在 $W+$ 空间中获得其潜在编码 w ,接下来利用 CLIP 模型获取文本描述中对于人脸属性 e_s^i 和属性状态 e_c^i 的要求以及输入图像中人脸属性 e_s^i 和属性状态 e_c^i 的现状,其中 e_s^i 和 e_s^i 输入高、中级语义信息残差映射器 M_t^c 、 M_t^m , e_c^i 和 e_c^i 输入低级语义信息残差映射器 M_t^f ,进而预测潜在编码的残差 Δw 。最后,将修改后的潜在编码 $w + \Delta w$ 输入预训练 StyleGAN2 模型获得最终结果。整个流程如图 1 所示,每个模块将在下文中进行详细说明。

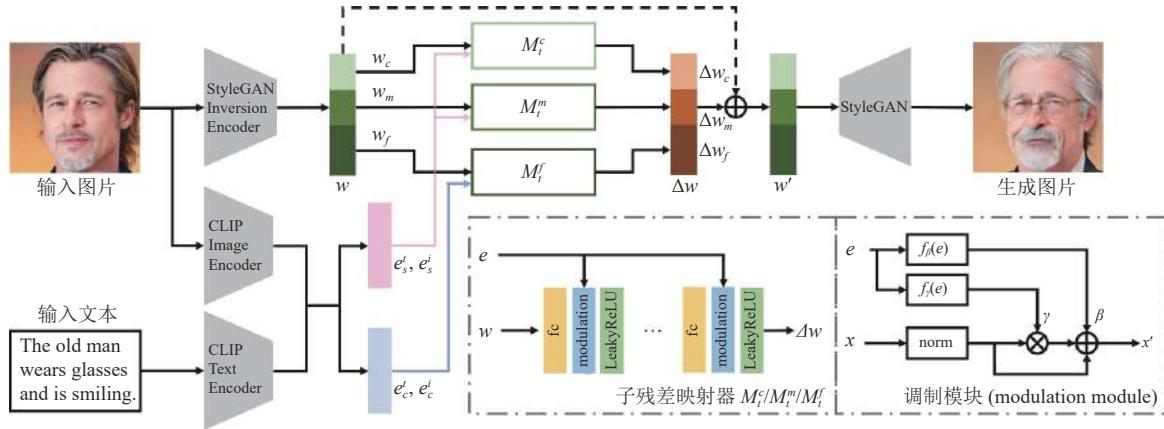


图 1 网络结构图

3.1 基于 CLIP 模型的共享条件编码学习

为了将文本域和图像域的特征统一到同一框架下,首先基于给定要编辑的真实图像或满足高斯分布的随机向量,我们使用 StyleGAN2 反演方法在 $W+$ 空间中获得潜在编码 w . 此后,对于用户提供的有关人脸属性和属性状态的文本描述,我们使用 CLIP 的文本编码器将其编码为 512 维的潜在编码 e^t ,由 e_s^t 和 e_c^t 组成. 类似地,输入图片由 CLIP 的图像编码器编码 e^i ,由 e_s^i 和 e_c^i 组成.

一旦获取到 CLIP 共享潜在空间内的图像和文本编码结果,接下来的目标是将其转换到 StyleGAN 的 $W+$ 空间中,同输入图片的反演编码一起输入残差映射器网络预测潜在编码的残差. 因此,当前的困难是如何将 CLIP 模型潜在空间下的图像和文本编码转换到另一个公共的潜在空间中,即 $W+$ 空间. 转换后的潜在编码是 L 个不同的 C 维 w 向量的连接,分别对应 StyleGAN 的不同输入层. 该潜在变量转换模块的训练可以公式化为:

$$\min_{\theta_T} L_T = \left\| \sum_{m=1}^L p_m (e_m^i - e_m^t) \right\|_2^2 \quad (1)$$

其中, θ_T 表示转换模块 $T(\cdot)$ 的参数, e^i 和 e^t 是获得的 CLIP 潜在空间下的图像编码和文本编码. e^i 和 e^t 具有相同的形状 $L \times C$, 意味着有 L 层, 其中每层都有一个 C 维潜在编码; p_m 是潜在编码中第 m 层的权重. 当 CLIP 的潜在编码被转换至 StyleGAN 的 $W+$ 空间后,所有的潜在编码均处于同一框架内,满足了输入残差映射器网络预测潜在编码残差的条件.

3.2 基于 StyleGAN2 模型的多级语义信息分离

正如 StyleCLIP 所验证的, StyleGAN2 生成器中的不同层对应生成图像中不同级别的语义信息,同一维度下层数越多其对应的语义信息级别越高. 因此,为了实现更加细粒度的人脸图像生成和编辑操作,本文中提出了多级残差映射器 M_t 分别对不同级别的脸语义属性参考文本描述的目标计算残差量,该残差量可直接用来修改已有的图像潜在编码,从而实现对人脸属性的生成或编辑操作. 与 StyleCLIP 中提出的映射器不同,本文提出的子残差映射器可根据全新定义的语义信息级别(人脸属性和属性状态)分别将图像和文本编码映射到 StyleGAN2 的潜在空间下,在实现属性解耦的同时,预测对应潜在编码的残差量.

该映射器包含 3 个子残差映射器 M_t^c 、 M_t^m 和 M_t^f , 分别对应高、中、低级别的语义信息,首先利用 StyleGAN2 反演模型对输入图像进行编码操作,根据语义信息级别将反演潜在编码 $w = (w_c, w_m, w_f)$ 的不同语义信息分别输入到相应的子残差映射器 M_t^c 、 M_t^m 和 M_t^f 中,其中, w_c 、 w_m 和 w_f 分别对应于输入图像反演编码的高级语义信息、中级语义信息和低级语义信息. 然后将通过 CLIP 图像和文本编码器获取的来自图像和文本内容的人脸属性潜在信息 $e_s \in \{e_s^i, e_s^t, 0\}$ 作为条件输入对应高、中级别语义信息的子残差映射器 M_t^c 和 M_t^m , 将来自图像和文本内容的属性状态 $e_c = \{e_c^i, e_c^t, 0\}$ 作为条件输入对应低级别语义信息的子残差映射器 M_t^f , 此时多级残差映射器的内部组成可

以表示为:

$$M_t(w, e_s, e_c) = (M_t^c(w_c, e_s), M_t^m(w_m, e_s), M_t^f(w_f, e_c)) \quad (2)$$

某些情况下, 保留某种语义级别并固定相应级别的语义信息同样可以完成多种操作, 因此本文提出的方法同样支持只训练3个映射器中任意1个或2个子映射器以实现不同粒度的人脸操作。训练过程中将不断优化多级映射器的残差估计能力, 使结果更加符合文本描述内容的同时保留无关人脸属性。

3.3 基于条件调制模块的特征融合

为了合理融合来自StyleGAN2反演模型和CLIP模型的潜在编码, 本文设计了一个全新的调制模块, 其功能在于当生成过程中图像分辨率较低时, 更多地保持随机生成图像或输入图像本身的身份信息; 在分辨率较高时, 填补更多地细节信息。如图1所示, 每个子残差映射器由5个子块组成, 每个子块由1个全连接层、1个调制模块和1个非线性激活层组成。其中全新设计的调制模块不是简单地将转换后的CLIP潜在编码与反演图像潜在编码叠加起来进行特征融合, 而是使用条件潜在编码来调制前一层的中间输出 x 。因此, 调制模块可以公式化为:

$$x' = (1 + f_\gamma(e)) \frac{x - \mu_x}{\sigma_x} + f_\beta(e) \quad (3)$$

其中, μ_x 和 σ_x 分别为 x 的平均值和标准偏差。 f_γ 和 f_β 为全连接网络(包含两个全连接层, 一个是中间归一化层和一个ReLU激活层)。受到条件图像翻译工作^[22,25,52]的启发, 如果没有为人脸属性或属性状态提供条件输入, 则将这种情况表示为 $e_s = 0$ 或 $e_c = 0$ 。通过这种方式, 我们灵活地支持用户只编辑人脸属性、只编辑属性状态, 或者同时编辑人脸属性和属性状态。

3.4 损失函数

我们的目标是在给定文本描述的基础上, 生成或编辑文本描述中想要实现的效果, 同时要求其他无关属性(例如背景、身份)得到良好保存。因此, 我们专门设计了两种类型的损失函数来训练残差映射器网络: 文本操作丢失和属性保留丢失。

3.4.1 文本操作损失

为了根据文本描述执行相应的人脸操作, 我们借助CLIP模型设计了文本操作损失 L_t , 公式如下:

$$L_t = L_{st}^{\text{CLIP}} + L_{ct}^{\text{CLIP}} \quad (4)$$

首先, 我们在CLIP联合表达空间下测量生成或编辑图像与给定文本之间在人脸属性上的余弦距离:

$$L_{st}^{\text{CLIP}} = 1 - \cos(E_i(G(w + M_t(w, e_s^t, e_c))), e_s^t) \quad (5)$$

其中, $\cos(\cdot)$ 表示余弦距离, E_i 表示CLIP图像编码器, G 表示预训练的StyleGAN生成器, M_t 表示残差映射器网络, e_s^t 表示由CLIP文本编码器提取的人脸属性信息, $e_c = \{e_c^i, e_c^t, 0\}$ 。同样地, 属性状态操作损失为:

$$L_{ct}^{\text{CLIP}} = 1 - \cos(E_i(G(w + M_t(w, e_s, e_c^t))), e_c^t) \quad (6)$$

其中, e_c^t 表示由CLIP图像编码器提取的人脸属性状态信息, $e_s \in \{e_s^i, e_s^t, 0\}$ 。

3.4.2 属性保留损失

为了确保人脸图像编辑前后的身份一致性, 身份损失如下:

$$L_{id} = 1 - \cos(R(G(w + M_t(w, e_s, e_c))), R(G(w))) \quad (7)$$

其中, $e_c = \{e_c^i, e_c^t, 0\}$, $e_s \in \{e_s^i, e_s^t, 0\}$, R 表示带有人脸识别功能的ArcFace网络^[57], $G(w)$ 表示重建的原始图像。此外, 为了保证在修改人脸属性时属性状态不丢失, 我们设计了一个损失函数 L_{sc} :

$$L_{sc} = \left\| \text{avg}(G(w + M_t(w + M_t(w, e_s, e_c)) \cdot P_h(G(w + M_t(w, e_s, e_c)))) - \text{avg}(G(w) \cdot P_h(G(w))) \right\|_1 \quad (8)$$

其中, $P_h(\cdot)$ 表示人脸区域的遮罩。根据经验, 如果我们只改变属性状态, 人脸属性可以很好地保存下来, 所以我们不会增加相应的保留损失。

此外, 我们在人脸分析网络的帮助下引入了背景损失:

$$L_{bg} = \|G(w + M_t(w, e_s, e_c)) - G(w) \cdot (P_{nh}(G(w + M_t(w, e_s, e_c))) \cap P_{nh}(G(w)))\|_2 \quad (9)$$

其中, $P_{nh}(\cdot)$ 表示非人脸区域的遮罩。通过这种方式, 我们可以确保与文本内容不相关的属性保持不变。出于同样的

目的, 我们引入了潜在空间中的操作步骤的 L_2 正则化损失:

$$L_{\text{norm}} = \|M_t(w, e_s, e_c)\|_2 \quad (10)$$

整体属性保留损失为:

$$L_{\text{ap}} = \lambda_{\text{id}} L_{\text{id}} + \lambda_{\text{sc}} L_{\text{sc}} + \lambda_{\text{bg}} L_{\text{bg}} + \lambda_{\text{norm}} L_{\text{norm}} \quad (11)$$

其中, λ_{id} , λ_{sc} , λ_{bg} , λ_{norm} 为权重系数.

综上, 总损失函数定义为:

$$L = \lambda_t L_t + \lambda_{\text{ap}} L_{\text{ap}} \quad (12)$$

其中, λ_t , λ_{ap} 为权重系数平衡损失函数.

4 实验分析

4.1 实验数据

本文中我们使用了两个图像数据集, 分别是 CelebA-HQ^[56] 和 FFHQ^[2] 数据集. CelebA-HQ 数据集的主要作用是训练和评估映射器网络. 首先针对数据集中频繁出现的人脸属性种类和人脸属性描述, 我们收集了 60 个通用性的人脸属性文本描述和 72 个通用性的属性状态文本描述, 紧接着在符合文本描述的图像集合中随机挑选两张人脸图像, 一张作为输入图片, 另一张图像的文本描述送入 CLIP 模型提取潜在编码, 经映射器、StyleGAN2 生成器获得图像与提供文本的图像计算损失, 不断优化映射器. 我们还使用其他文本引导的人脸编辑方法生成了一部分编辑过的图像, 以增加图像的多样性. FFHQ 数据集的主要作用是训练 StyleGAN2 模型反演, 该过程参考 Richardson 等人^[50]提出的方法及数据集分割方式, 保证模型反演的编码结果可由 StyleGAN2 生成器重新生成具有与输入图像极高相似度的人脸图像, 便于后续的潜在编码操作.

CelebA-HQ 数据集是 CelebA 的高质量版本, 即高质量名人人脸属性数据集, 包含 10177 个名人身份的 202599 张人脸图片, 每张图片都进行了语义标注, 包含人脸标注框、5 个人脸特征点坐标以及 40 个属性标记, 被广泛用于人脸相关的计算机视觉任务, 包括人脸属性检测、人脸属性编辑以及人脸关键点检测等.

FFHQ 数据集的全称为 Flickr-Faces-HQ Dataset, 最初作为生成对抗网络的训练数据集创建, 后期被用作 StyleGAN2 的训练数据集. FFHQ 是一个高质量的人脸数据集, 包含 70000 张 1024×1024 分辨率的高清人脸图像, 其在年龄、种族和图像背景上丰富多样且差异明显, 在人脸属性上也拥有非常多的变化, 拥有不同的年龄、性别、种族、肤色、表情、脸型、发型、人脸姿态等, 涵盖普通眼镜、太阳镜、帽子、发饰及围巾等多种人脸周边配件, 因此该数据集也是可以用于开发一些人脸属性分类或者人脸属性编辑算法. FFHQ 的图像从 Flickr 上爬取, 且均有许可才会下载, 并使用了 dlib 进行人脸对齐和裁剪, 之后使用算法移除了一些非真实人脸如雕像、画作及照片等图像.

4.2 实验细节

本文中 StyleGAN2 生成模型采用基于 FFHQ 预训练生成器, 反演模型训练过程中的数据集划分、优化方法参考 Richardson 等人^[50]提出的方法. CLIP 模型使用官方提供的 ViT-B/32 预训练分类器. 在训练过程中, 残差映射器被要求根据提供的文本描述, 只编辑人脸属性状态或人脸属性及其状态. 训练策略方面, 初始学习率为 0.0005, batch 大小为 8. 训练迭代次数为 500000 次, 使用了 Adam 优化器^[58], β_1 和 β_2 分别设置为 0.9 和 0.999. 权重分别为 $\lambda_{\text{id}} = 0.5$, $\lambda_{\text{sc}} = 0.04$, $\lambda_{\text{bg}} = 1$, $\lambda_{\text{norm}} = 1.2$, $\lambda_t = 2$, $\lambda_{\text{ap}} = 1.5$.

4.3 评价指标及基准模型

评价指标包含了 4 个比较重要的方面: 图像质量、图像多样性、操作准确性和操作真实性.

我们使用 Fréchet 初始距离 (FID)^[59] 评估生成或编辑图像的质量, 其计算方法为随机选择了 10000 张源图像和 10000 张目标图像, 模型通过文本命令将所有源图像进行相同的目标属性生成或操作, 再使用 10000 张生成的图像与选择的 10000 张目标图像来估计 FID. 使用感知图像块相似性 (LPIPS)^[60] 衡量生成图像的多样性, 其计算方法为利用深度神经网络^[61] 提取两幅图像的特征并计算距离, 我们随机选择 100 张输入图像并将其转换到不同的域.

每进行一次域转换, 我们为每个输入图像生成 10 幅图像, 并计算平均 LPIPS 距离, 所有距离的平均值作为 LPIPS 数值, 更高的数值表示生成的图像之间有更高的多样性.

通过用户研究评估了操作准确性和操作真实性. 我们通过在相同条件下随机抽取 50 幅图像, 并从不同背景的不同人群中收集 20 多份调查, 来测试准确性和真实性. 为了评估操作的准确性, 除了修改合成图像的视觉属性与文本对齐外, 还要求用户判断是否保留了与文本无关的内容. 为了真实感, 用户被要求在上述方法的给定结果中判断哪一个更真实. 生成的准确性通过给定描述和相应生成图像之间的相似性来评估, 用户被要求判断哪个图像与给定文本更一致.

在文本驱动人脸图像生成任务的定性、定量对比实验中, 基准模型包括 AttnGAN^[10]、ControlGAN^[26]、DM-GAN^[27]、DF-GAN^[23]和 TediGAN^[35]. 其中 TediGAN 为使用 CLIP 模型作为文本编码器的版本.

在文本驱动人脸图像编辑任务的定性、定量对比实验中, 基准模型包括 StyleCLIP-Global Direction^[48]、TediGAN^[35]. 考虑到 StyleCLIP-Latent Mapper^[36]会针对某种特殊文字描述进行专门的优化, 无法处理随机语句的情况, 故不将其作为对比试验的方法. 其中 TediGAN 为使用 CLIP 模型作为文本编码器的版本. 对于上述所有被比较的基准模型, 我们使用官方的训练代码或预训练模型.

4.4 实验方法

针对文本驱动人脸图像生成任务, 本实验采取的方案是首先采样满足高斯分布的随机向量, 使用 StyleGAN2 反演方法在 $W+$ 空间中获得潜在编码, 同时使用 CLIP 模型提取输入文本描述的语义向量并转换到同一潜在空间 $W+$ 进行潜在编码残差的计算, 获取到图像编辑方向重新生成符合语义信息的人脸图像, 再与基准模型进行定性、定量对比并制定用户调查的样本集. 针对文本驱动人脸图像编辑任务, StyleGAN2 编码器和 CLIP 编码器将图像潜在编码和语义表示向量放入到一个共享潜在空间中, 实现文本和图像潜在编码的相似性比较从而获取潜在编码残差, 并重新生成符合语义信息的人脸图像, 再与基准模型进行定性、定量对比并制定用户调查的样本集. 为了验证网络结构设计的有效性, 我们针对网络结构中的一些机制进行了消融性实验.

StyleGAN2 反转编码器采用了 e4e 框架及数据集分割方式, 将真实图像转化到 StyleGAN2 潜在空间并实现高保真的图像重建. 我们用于反演的 StyleGAN2 模型基于 FFHQ 数据集的预训练模型. 对于 CLIP 模型, 我们采用官方发布的预训练模型将语义内容转化为表示向量. 修改后的潜在编码我们使用预训练的 StyleGAN2 生成模型进行最终结果的获取.

4.5 实验结果与分析

4.5.1 基于文本驱动的图像生成算法

如后文图 2 所示, 大多数现有的文本到图像的生成方法都可以生成逼真且与文本相关的结果. 然而, 文本中包含的一些属性没有出现在生成的图像中, 生成的图像缺乏细节. 当使用多阶段训练方法生成更高分辨率的图像时, 细节的缺失现象将非常明显且不可修复. 此外, 大多数现有解决方案的输出多样性有限, 即使提供的条件包含不同的含义. 例如, “有胡子”可能意味着山羊胡、短胡子或长胡子, 并且可能有不同的属性状态. 我们的方法不仅可以产生多样性的结果, 而且还可以通过使用控制参数实现期望的改变. 为了产生不同的结果, 在与文本相关的层不变的情况下, 其他层可以被先前分布中的任何值所取代. 例如, 如图 2 第 1 行所示, 关键的视觉属性(男性、眼袋、鼻子和胡子)维持了文本中的描述, 而其他属性, 如发型、年龄、和表情, 则显示出很大程度的改变.

4.5.2 基于文本驱动的图像编辑算法

如图 3 所示, 第 2 行的文本中有增加耳环的描述, 改变女性的脸型和发型, 我们的方法完成了这个复杂的编辑需求, 而无法产生所需的属性. StyleCLIP 产生的结果均存在颜色信息及清晰度的缺失. 在某些情况下, TediGAN 的结果会出现与文本相关的区域不会被修改, 而与文本无关的区域会被更改的情况. 此外, 由于我们使用的 StyleGAN 是在一个非常大的人脸数据集上预先训练的, 潜在空间几乎覆盖了人脸属性的整个空间, 这使得我们的方法对真实图像具有鲁棒性. 最后两行中的图像是来自其他人脸数据集, 这说明我们的方法已准备好在图像产生令人满意的结果.



图 2 基于文本驱动的图像生成算法对比实验

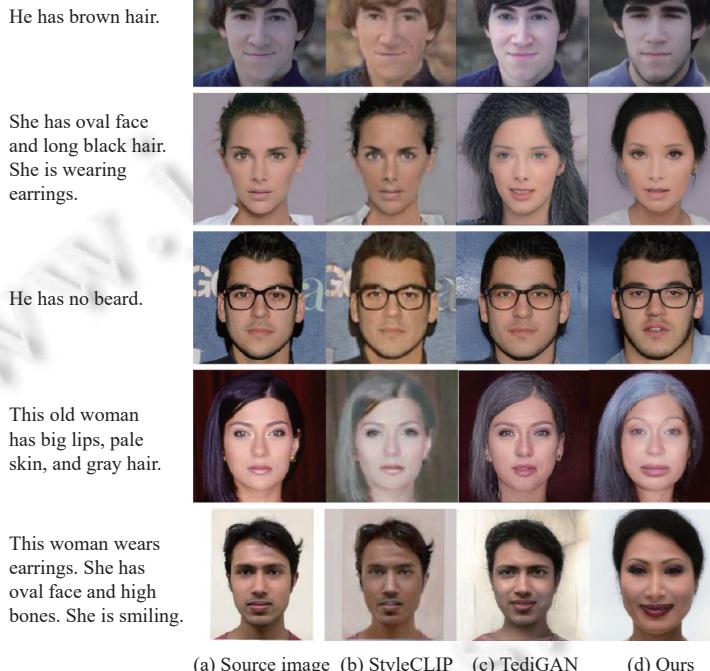


图 3 基于文本驱动的图像编辑算法对比实验

4.6 定量实验结果与分析

4.6.1 基于文本驱动的图像生成算法

在我们的实验中, 我们评估了从随机选择的文本描述生成的大量样本的 FID 和 LPIPS。为了评估准确性和真实性, 我们使用不同的方法从 50 个随机抽样的文本中生成图像。在一項用户研究中, 用户被要求判断哪一个最逼真, 与给定文本最连贯。结果如表 1 所示。与现有技术相比, 我们的方法获得了更好的 FID、LPIPS、准确率和真实感, 这证明我们的方法可以生成具有高质量、多样性、真实性和文本相关性的图像。

4.6.2 基于文本驱动的图像编辑算法

在我们的实验中, 我们评估了 FID, 并使用随机选择的描述对随机选择的 FFHQ 和非 FFHQ 数据集内的图像

进行了用户研究, 结果如表 2 所示。与 StyleCLIP 和 TediGAN 相比, 我们的方法实现了更好的 FID、准确性和真实性。这表明我们的方法可以生成高质量的合成图像, 修改的内容与给定的描述高度一致的同时, 保留了与文本无关的相关属性。

表 1 基于文本驱动的图像生成算法对比实验结果

方法	FID	LPIPS	准确率 (%)	真实感 (%)
AttnGAN	125.98	0.512	14.2	20.3
ControlGAN	116.32	0.522	18.2	22.5
DM-GAN	137.60	0.581	22.8	25.5
DF-GAN	131.05	0.544	19.5	12.8
TediGAN	106.37	0.456	25.3	31.7
Ours	103.34	0.441	33.8	35.9

表 2 基于文本驱动的图像编辑算法对比实验结果

方法	FID	准确率 (%)	真实感 (%)
StyleCLIP	107.25	34.3	42.6
TediGAN	101.27	38.3	46.5
Ours	97.33	40.8	48.9

4.7 消融实验

为了验证我们提出的网络结构和损失函数的有效性, 我们交替地停用其中一个关键组件, 通过保持除所选组件之外的所有组件不变来重新训练我们方法的变体。

4.7.1 属性保留损失

为了验证每种成分在属性保留丢失中的作用, 我们随机选择 4400 张图像进行定性消融研究, 研究的任务是只编辑发型, 文本描述为“slicked back hairstyle”。从图 4 中可以得出一致的结论: L_{bg} 、 L_{id} 和 L_{norm} 有助于保持不相关属性, L_{sc} 有助于仅编辑发型时保持属性状态不变。

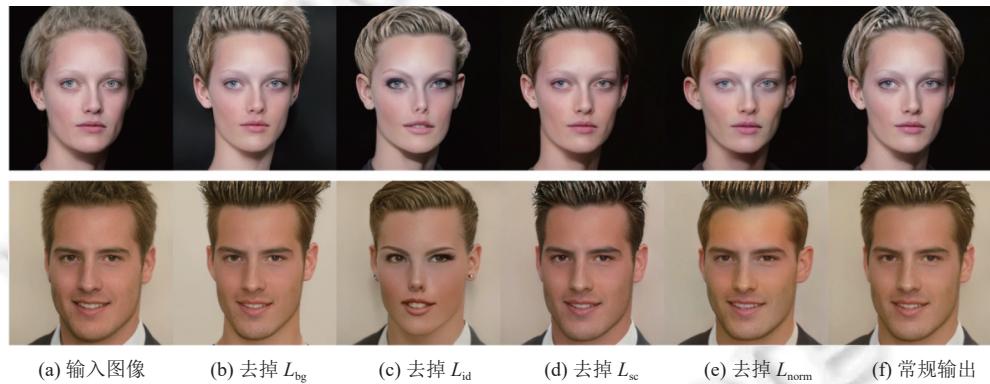


图 4 属性保留损失笑容消融对比实验

4.7.2 网络结构设计

我们将模型与 3 种变体进行了比较, 实验中对应的语句描述为“perm hairstyle and red hair”。① 将调制模块替换为 Vanilla 归一化层, 并将条件输入与潜在编码连接起来送入网络。② 将 M_t^c 和 M_t^m 的条件输入替换为属性状态编码, 将 M_t^m 的条件输入替换为面部属性编码。③ 将 M_t^m 的条件输入替换为属性状态编码, 保持其余不变。如后文图 5 所示, 只有我们的模型完成了语句描述的图像编辑任务。图 5(b) 的结果证明我们的调制模块能够更好地将状态信息融合到潜在空间中, 提高了模型的泛化能力。图 5(c) 和图 5(d) 的结果确认了我们根据不同尺度分离信息实现语义匹配的必要性。

5 总 结

在本文中, 我们提出了一个统一的文本驱动人脸图像生成和编辑框架, 在该框架下用户可以单独提供文本描述实现人脸图像的生成也可以提供文本描述和参考图片实现人脸图像的编辑。这种多模式交互极大地增加了人脸

编辑的灵活性,降低了用户使用成本。通过最大限度地挖掘 CLIP 模型对于文本-图像差异性的捕捉,配合定制化网络结构设计和损失函数,我们的方法借助 StyleGAN2 模型的图像生成能力实现高质量的人脸图像生成和编辑,同时对文本描述以外的部分实现尽可能地保留。额外的定性、定量和消融对比实验以及大量的用户调查证明了我们的方法相比于其他方法在操作能力、无关属性保存和图像真实性方面的优越性。

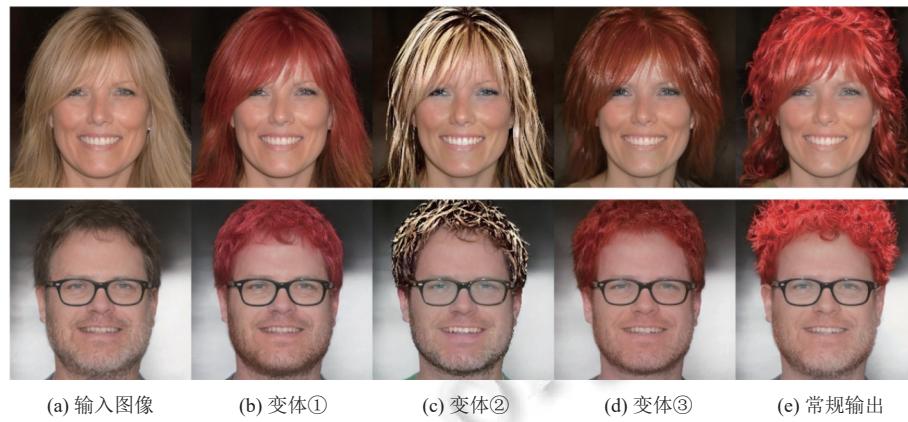


图 5 网络结构设计消融对比实验

References:

- [1] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: ACM, 2014. 2672–2680.
- [2] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4401–4410. [doi: [10.1109/CVPR.2019.00453](https://doi.org/10.1109/CVPR.2019.00453)]
- [3] Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of StyleGAN. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8110–8119. [doi: [10.1109/CVPR42600.2020.00813](https://doi.org/10.1109/CVPR42600.2020.00813)]
- [4] Ghosh A, Zhang R, Dokania P, Wang O, Efros A, Torr P, Shechtman E. Interactive sketch & fill: Multiclass sketch-to-image translation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 1171–1180. [doi: [10.1109/ICCV.2019.00126](https://doi.org/10.1109/ICCV.2019.00126)]
- [5] Xia WH, Yang YJ, Xue JH. Cali-Sketch: Stroke calibration and completion for high-quality face image generation from Human-like sketches. arXiv:1911.00426, 2019.
- [6] Lin JX. Research on image-to-image translation [Ph.D. Thesis]. Hefei: University of Science and Technology of China, 2020 (in Chinese with English abstract). [doi: [10.27517/d.cnki.gzkj.2020.000509](https://doi.org/10.27517/d.cnki.gzkj.2020.000509)]
- [7] Wang TC, Liu MY, Zhu JY, Tao A, Kautz J, Catanzaro B. High-resolution image synthesis and semantic manipulation with conditional GANs. In: Proc. of the 2018 IEEE Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8798–8807. [doi: [10.1109/CVPR.2018.00917](https://doi.org/10.1109/CVPR.2018.00917)]
- [8] Gu GH, Cao YY, Li G, Zhao Y. Image hierarchical classification based on semantic label generation and partial order structure. Ruan Jian Xue Bao/Journal of Software, 2020, 31(2): 531–543 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/5630.htm> [doi: [10.13338/j.cnki.jos.005630](https://doi.org/10.13338/j.cnki.jos.005630)]
- [9] Nam S, Kim Y, Kim SJ. Text-adaptive generative adversarial networks: Manipulating images with natural language. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: ACM, 2018. 42–51.
- [10] Xu T, Zhang PC, Huang QY, Zhang H, Gan Z, Huang XL, He XD. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In: Proc. of the 2018 IEEE Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1316–1324. [doi: [10.1109/CVPR.2018.00143](https://doi.org/10.1109/CVPR.2018.00143)]
- [11] Du PF, Li XY, Gao YL. Survey on multimodal visual language representation learning. Ruan Jian Xue Bao/Journal of Software, 2021, 32(2): 327–348 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/6125.htm> [doi: [10.13338/j.cnki.jos.006125](https://doi.org/10.13338/j.cnki.jos.006125)]
- [12] Song JM, Zhao SJ, Ermon S. Generative adversarial learning of Markov chains. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017. 1–7.

- [13] Ma D. Research on key technology and applications of generative adversarial nets [Ph.D. Thesis]. Chengdu: University of Electronic Science and Technology of China, 2020 (in Chinese with English abstract). [doi: [10.27005/d.cnki.gdzku.2020.004624](https://doi.org/10.27005/d.cnki.gdzku.2020.004624)]
- [14] Wu FX, Cheng J. Configurable text-based image editing by autoencoder-based generative adversarial networks. *Ruan Jian Xue Bao/Journal of Software*, 2022, 33(9): 3139–3151 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/6622.htm> [doi: [10.13328/j.cnki.jos.006622](https://doi.org/10.13328/j.cnki.jos.006622)]
- [15] Lin Y. Cross-domain face synthesis and application based on generative adversarial networks [Ph.D. Thesis]. Chengdu: Sichuan University, 2021 (in Chinese with English abstract). [doi: [10.27342/d.cnki.gscdu.2021.000741](https://doi.org/10.27342/d.cnki.gscdu.2021.000741)]
- [16] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 8748–8763.
- [17] Xiao JS, Zhou JL, Lei JF, Li L, Ding L, Du ZY. Improved generative adversarial network for image scene transformation. *Ruan Jian Xue Bao/Journal of Software*, 2021, 32(9): 2755–2768 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/5986.htm> [doi: [10.13328/j.cnki.jos.005986](https://doi.org/10.13328/j.cnki.jos.005986)]
- [18] Xu XZ, Chang JY, Ding SF. Image style transferring based on StarGAN and class encoder. *Ruan Jian Xue Bao/Journal of Software*, 2022, 33(4): 1516–1526 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/6482.htm> [doi: [10.13328/j.cnki.jos.006482](https://doi.org/10.13328/j.cnki.jos.006482)]
- [19] Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: ACM, 2016. 2180–2188.
- [20] Cheng J, Wu FX, Tian YL, Wang L, Tao DP. RiFeGAN: Rich feature generation for text-to-image synthesis from prior knowledge. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10908–10917. [doi: [10.1109/CVPR42600.2020.01092](https://doi.org/10.1109/CVPR42600.2020.01092)]
- [21] Li BW, Qi XJ, Torr PHS, Lukasiewicz T. Lightweight generative adversarial networks for text-guided image manipulation. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: ACM, 2020. 22020–22031.
- [22] Reed SE, Akata Z, Yan XC, Logeswaran L, Schiele B, Lee H. Generative adversarial text to image synthesis. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York City: PMLR, 2016. 1060–1069.
- [23] Tao M, Tang H, Wu S, Jing XY, Bao BK, Xu CS. DF-GAN: A simple and effective baseline for text-to-image synthesis. arXiv:2008.05865, 2020.
- [24] Zhang H, Xu T, Li HS, Zhang ST, Wang XG, Huang XL, Metaxas D. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 5908–5916. [doi: [10.1109/ICCV.2017.629](https://doi.org/10.1109/ICCV.2017.629)]
- [25] Zhang H, Xu T, Li HS, Zhang ST, Wang XG, Huang XL, Metaxas DN. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2019, 41(8): 1947–1962. [doi: [10.1109/TPAMI.2018.2856256](https://doi.org/10.1109/TPAMI.2018.2856256)]
- [26] Li BW, Qi XJ, Lukasiewicz T, Torr PHS. Controllable text-to-image generation. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: ACM, 2019. 2065–2075.
- [27] Zhu MF, Pan PB, Chen W, Yang Y. DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5802–5810. [doi: [10.1109/CVPR.2019.00595](https://doi.org/10.1109/CVPR.2019.00595)]
- [28] Dong H, Yu SM, Wu C, Guo YK. Semantic image synthesis via adversarial learning. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 5707–5715. [doi: [10.1109/ICCV.2017.608](https://doi.org/10.1109/ICCV.2017.608)]
- [29] Li YT, Gan Z, Shen YL, Liu JJ, Cheng Y, Wu YX, Carin L, Carlson D, Gao JF. StoryGAN: A sequential conditional GAN for story visualization. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 6329–6338. [doi: [10.1109/CVPR.2019.00649](https://doi.org/10.1109/CVPR.2019.00649)]
- [30] Mirza M, Osindero S. Conditional generative adversarial nets. arXiv:1411.1784, 2014.
- [31] Härkönen E, Hertzmann A, Lehtinen J, Paris S. GANSpace: Discovering interpretable GAN controls. In: Proc. of the 34th Conf. on Neural Information Processing Systems. Vancouver: NeurIPS, 2020. 9841–9850.
- [32] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019. 1–35.
- [33] Shen YJ, Zhou BL. Closed-form factorization of latent semantics in GANs. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 1532–1540. [doi: [10.1109/CVPR46437.2021.00158](https://doi.org/10.1109/CVPR46437.2021.00158)]

- [34] Yüksel OK, Simsar E, Er EG, Yanardag P. LatentCLR: A contrastive learning approach for unsupervised discovery of interpretable directions. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 14243–14252. [doi: [10.1109/ICCV48922.2021.01400](https://doi.org/10.1109/ICCV48922.2021.01400)]
- [35] Xia WH, Yang YJ, Xue JH, Wu BY. TediGAN: Text-guided diverse face image generation and manipulation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 2256–2265. [doi: [10.1109/CVPR46437.2021.00229](https://doi.org/10.1109/CVPR46437.2021.00229)]
- [36] Patashnik O, Wu ZZ, Shechtman E, Cohen-Or D, Lischinski D. StyleCLIP: Text-driven manipulation of StyleGAN imagery. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 2065–2074. [doi: [10.1109/ICCV48922.2021.00209](https://doi.org/10.1109/ICCV48922.2021.00209)]
- [37] Chen J, Bai C, Ma Q, Hao PY, Chen SY. Adversarial training triplet network for fine-grained sketch based image retrieval. Ruan Jian Xue Bao/Journal of Software, 2020, 31(7): 1933–1942 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/5934.htm> [doi: [10.13328/j.cnki.jos.005934](https://doi.org/10.13328/j.cnki.jos.005934)]
- [38] Hossain MDZ, Sohel F, Shiratuddin MF, Laga H. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys, 2019, 51(6): 118. [doi: [10.1145/3295748](https://doi.org/10.1145/3295748)]
- [39] Antol S, Agrawal A, Lu JS, Mitchell M, Batra D, Zitnick CL, Parikh D. VQA: Visual question answering. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 2425–2433. [doi: [10.1109/ICCV.2015.279](https://doi.org/10.1109/ICCV.2015.279)]
- [40] Teney D, Anderson P, He XD, van den Hengel A. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In: Proc. of the 2018 IEEE Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4223–4232. [doi: [10.1109/CVPR.2018.00444](https://doi.org/10.1109/CVPR.2018.00444)]
- [41] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423)]
- [42] Vaswani A, Shazeer N, Parmar N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: ACM, 2017. 6000–6010.
- [43] Shen S, Li LH, Tan H, Bansal M, Rohrbach A, Chang KW, Yao ZW, Keutzer K. How much can CLIP benefit vision-and-language tasks? In: Proc. of the 10th Int'l Conf. on Learning Representations. OpenReview.net, 2021. 1–18.
- [44] Wang MM, Xing JZ, Liu Y. ActionCLIP: A new paradigm for video action recognition. arXiv:2109.08472, 2021.
- [45] Tov O, Alaluf Y, Nitzan Y, Patashnik O, Cohen-Or D. Designing an encoder for StyleGAN image manipulation. ACM Trans. on Graphics, 2021, 40(4): 133. [doi: [10.1145/3450626.3459838](https://doi.org/10.1145/3450626.3459838)]
- [46] Xia WH, Zhang YL, Yang YJ, Xue JH, Zhou BL, Yang MH. GAN inversion: A survey. arXiv:2101.05278, 2021.
- [47] Xia WH, Yang YJ, Xue JH, Wu BY. Towards open-world text-guided face image generation and manipulation. arXiv:2104.08910, 2021.
- [48] Wu ZZ, Lischinski D, Shechtman E. StyleSpace analysis: Disentangled controls for StyleGAN image generation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 12863–12872. [doi: [10.1109/CVPR46437.2021.01267](https://doi.org/10.1109/CVPR46437.2021.01267)]
- [49] Nitzan Y, Bermano A, Li YY, Cohen-Or D. Face identity disentanglement via latent space mapping. ACM Trans. on Graphics, 2020, 39(6): 225. [doi: [10.1145/3414685.3417826](https://doi.org/10.1145/3414685.3417826)]
- [50] Richardson E, Alaluf Y, Patashnik O, Nitzan Y, Azar Y, Shapiro S, Cohen-Or D. Encoding in style: A StyleGAN encoder for image-to-image translation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 2287–2296. [doi: [10.1109/CVPR46437.2021.00232](https://doi.org/10.1109/CVPR46437.2021.00232)]
- [51] Alaluf Y, Patashnik O, Cohen-Or D. Only a matter of style: Age transformation using a style-based regression model. ACM Trans. on Graphics, 2021, 40(4): 45. [doi: [10.1145/3450626.3459805](https://doi.org/10.1145/3450626.3459805)]
- [52] Shen YJ, Gu JJ, Tang XO, Zhou BL. Interpreting the latent space of GANs for semantic face editing. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9243–9252. [doi: [10.1109/CVPR42600.2020.00926](https://doi.org/10.1109/CVPR42600.2020.00926)]
- [53] Abdal R, Zhu PH, Mitra NJ, Wonka P. StyleFlow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. ACM Trans. on Graphics, 2021, 40(3): 21. [doi: [10.1145/3447648](https://doi.org/10.1145/3447648)]
- [54] Voynov A, Babenko A. Unsupervised discovery of interpretable directions in the GAN latent space. In: Proc. of the 37th Int'l Conf. on Machine Learning. PMLR, 2020. 9786–9796.
- [55] Wang BX, Ponce CR. The geometry of deep generative image models and its applications. arXiv:2101.06006, 2021.
- [56] Lee C, Liu ZW, Wu LY, Luo P. MaskGAN: Towards diverse and interactive facial image manipulation. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5548–5557. [doi: [10.1109/CVPR42600.2020.00559](https://doi.org/10.1109/CVPR42600.2020.00559)]
- [57] Deng JK, Guo J, Xue NN, Zafeiriou S. ArcFace: Additive angular margin loss for deep face recognition. In: Proc. of the 2019 IEEE/CVF

- Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4690–4699. [doi: [10.1109/CVPR.2019.00482](https://doi.org/10.1109/CVPR.2019.00482)]
- [58] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014.
- [59] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: ACM, 2017. 6629–6640.
- [60] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of the 2018 IEEE Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 586–595. [doi: [10.1109/CVPR.2018.00068](https://doi.org/10.1109/CVPR.2018.00068)]
- [61] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

附中文参考文献:

- [6] 林剑新. 图像到图像翻译的研究 [博士学位论文]. 合肥: 中国科学技术大学, 2020. [doi: [10.27517/d.cnki.gzkju.2020.000509](https://doi.org/10.27517/d.cnki.gzkju.2020.000509)]
- [8] 顾广华, 曹宇尧, 李刚, 赵耀. 基于语义标签生成和偏序结构的图像层级分类. 软件学报, 2020, 31(2): 531–543. <http://www.jos.org.cn/1000-9825/5630.htm> [doi: [10.13328/j.cnki.jos.005630](https://doi.org/10.13328/j.cnki.jos.005630)]
- [11] 杜鹏飞, 李小勇, 高雅丽. 多模态视觉语言表征学习研究综述. 软件学报, 2021, 32(2): 327–348. <http://www.jos.org.cn/1000-9825/6125.htm> [doi: [10.13328/j.cnki.jos.006125](https://doi.org/10.13328/j.cnki.jos.006125)]
- [13] 马丹. 生成对抗网络的关键技术研究与应用 [博士学位论文]. 成都: 电子科技大学, 2020. [doi: [10.27005/d.cnki.gdzku.2020.004624](https://doi.org/10.27005/d.cnki.gdzku.2020.004624)]
- [14] 吴福祥, 程俊. 基于自编码器生成对抗网络的可配置文本图像编辑. 软件学报, 2022, 33(9): 3139–3151. <http://www.jos.org.cn/1000-9825/6622.htm> [doi: [10.13328/j.cnki.jos.006622](https://doi.org/10.13328/j.cnki.jos.006622)]
- [15] 林野. 基于生成对抗网络的跨域人脸识别研究和应用 [博士学位论文]. 成都: 四川大学, 2021. [doi: [10.27342/d.cnki.gscdu.2021.000741](https://doi.org/10.27342/d.cnki.gscdu.2021.000741)]
- [17] 肖进胜, 周景龙, 雷俊峰, 李亮, 丁玲, 杜治一. 面向图像场景转换的改进型生成对抗网络. 软件学报, 2021, 32(9): 2755–2768. <http://www.jos.org.cn/1000-9825/5986.htm> [doi: [10.13328/j.cnki.jos.005986](https://doi.org/10.13328/j.cnki.jos.005986)]
- [18] 许新征, 常建英, 丁世飞. 基于StarGAN和类别编码器的图像风格转换. 软件学报, 2022, 33(4): 1516–1526. <http://www.jos.org.cn/1000-9825/6482.htm> [doi: [10.13328/j.cnki.jos.006482](https://doi.org/10.13328/j.cnki.jos.006482)]
- [37] 陈健, 白琮, 马青, 郝鹏翼, 陈胜勇. 面向细粒度草图检索的对抗训练三元组网络. 软件学报, 2020, 31(7): 1933–1942. <http://www.jos.org.cn/1000-9825/5934.htm> [doi: [10.13328/j.cnki.jos.005934](https://doi.org/10.13328/j.cnki.jos.005934)]



李宗霖(1995—), 男, 博士生, 主要研究领域为计算机视觉, 模式识别, 图像处理, 计算机图形学, 多媒体技术.



张兆心(1979—), 男, 博士, 教授, 博士生导师, 主要研究领域为图像处理, 网络安全态势感知, 网络资源测绘, 金融数据分析.



张盛平(1983—), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为计算机视觉, 目标追踪, 机器学习, 多媒体技术, 图像处理及模式识别, 生物特征识别技术.



张维刚(1980—), 男, 博士, 副教授, CCF 高级会员, 主要研究领域为多媒体技术, 视觉感知分析, 图像视频处理, 模式识别, 机器学习.



刘杨(1978—), 女, 博士, 讲师, 主要研究领域为图像处理, 信息安全.



黄庆明(1965—), 男, 博士, 教授, 博士生导师, CCF 会士, 主要研究领域为多媒体技术, 图像处理, 模式识别, 计算机视觉, 机器学习.