

PSPCLIP: 文本引导的人体着装图像编辑

杨正中, 杜小勤

(武汉纺织大学 计算机与人工智能学院, 湖北 武汉 430200)

摘要: 文本引导的人体着装图像编辑是一个具有挑战性的计算机视觉任务,其目标是对给定人体着装图像的服装区域进行符合文本语义的编辑。现有基于StyleGAN的方法在处理多样化服装、身体形状和姿势时存在困难。为解决这些问题,提出一种新颖的文本条件编辑模型PSPCLIP,其可将人体着装图像与文本提示词直接编码到预训练的无条件生成对抗网络的隐空间中,以实现给定的文本条件编辑。具体而言,PSPCLIP通过在生成所需图像时使用分层风格映射和引入全局自注意力的方法增强对隐编码的编辑过程,同时利用预训练多模态模型CLIP强大的图像—文本关联能力来强制执行目标语义。通过对VITON图像的实验以及与几种先进的基于文本条件的图像编辑方法进行比较,发现PSPCLIP能生成高度逼真的人体着装图像,并且在编辑性能方面优于对照方法,Semantics平均提升0.069,FID平均下降20.8。

关键词: 多模态; 图像编辑; 人体着装; 隐编码; 风格生成对抗网络

DOI: 10.11907/rjtk.241189

开放科学(资源服务)标识码(OSID):

中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2025)004-0164-09



PSPCLIP: Text-guided Human Clothing Image Editing

YANG Zhengzhong, DU Xiaolin

(School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, China)

Abstract: Text guided human clothing image editing is a challenging computer vision task that aims to edit the clothing area of a given human clothing image in accordance with text semantics. The existing StyleGAN based methods face difficulties in handling diverse clothing, body shapes, and postures. To address these issues, a novel text conditional editing model PSPCLIP is proposed, which can directly encode human clothing images and text prompt words into the hidden space of a pre trained unconditional generative adversarial network to achieve given text conditional editing. Specifically, PSPCLIP enhances the editing process of steganography by using hierarchical style mapping and introducing global self attention when generating the desired image; Simultaneously utilizing the powerful image text association capability of the pre trained multimodal model CLIP to enforce target semantics. Through experiments on VITON images and comparisons with several advanced text-based image editing methods, it was found that PSPCLIP can generate highly realistic images of human clothing and outperforms the control method in terms of editing performance, with an average improvement of 0.069 in semantics and an average decrease of 20.8 in FID.

Key Words: multimodal; image editing; human clothing; latent code; style generative adversarial network

0 引言

文本引导的人体着装图像编辑是指通过文本描述来指导图像编辑过程,生成视觉上令人信服的人体着装图像。该方法不修改原始身体特征,如面部、皮肤、姿态等,并且可以通过自然语言来表达广泛的视觉概念,在虚拟试穿、在线购物和时尚设计等领域有着广泛应用,大大提升

了用户购物体验。

文本引导的人体着装图像编辑通常依赖于深度学习模型。自多模态预训练大模型CLIP(Contrastive Language-Image Pre-training)出现以来,基于文本的图像合成控制成为一个活跃的研究领域。CLIP模型通过对比学习的方法进行模型参数优化,从而学习图像与文本之间的跨模态表示^[1]。许多研究基于CLIP模型,根据给定文本提示词操纵隐空间编码来控制生成器生成图像,其编辑流程范式为首

收稿日期: 2024-03-01

扫描二维码阅读全文:



作者简介: 杨正中(1999-),男,武汉纺织大学计算机与人工智能学院硕士研究生,研究方向为生成式人工智能;杜小勤(1968-),男,博士,武汉纺织大学计算机与人工智能学院副教授、硕士生导师,研究方向为强化学习、生成式人工智能。

先利用预训练的生成对抗网络(Generative Adversarial Networks, GAN)反演模型将原始图像转换为对应隐空间的隐编码;然后通过设计的文本条件调制模块将文本的语义信息转化为隐编码的位移;最后融合原始隐编码和隐编码位移,通过GAN得到编辑后的图像^[2-4]。然而,以上方法常常忽视服装上的文字信息或损害人物的身份。究其原因,在作为多模态编辑模型核心的文本条件调制模块设计上,上述方法普遍选择全连接层映射网络作为基础网络结构,在对隐编码级别的编辑优化上仅依赖于损失函数的约束,缺乏图像特征的指导,尽管能生成与输入语义对齐的图像,但在语义信息解耦以及图像细节呈现方面仍面临挑战。

1 相关研究

1.1 风格生成对抗网络

GANs由生成器(Generator)和判别器(Discriminator)从输入图像或随机噪声中生成合成图像^[5]。这两个网络采用激烈的对抗方式同时进行训练和提升,生成器逐渐生成越来越接近真实样本的图像。然而传统GAN在图像生成过程中存在不稳定性、低分辨率及细节不清晰等缺点。风格对抗生成网络(Style Generative Adversarial Networks, StyleGAN)通过两大技术创新解决了这些问题。首先,StyleGAN采用分层的生成器架构,通过在生成器和判别器中嵌入多个转换层,允许对图像的各种特征,如分辨率、颜色和纹理进行精细控制。这种分层架构使生成器能够同时捕捉图像的全局与局部特征,进而创造出更细腻、更真实的图像;其次,StyleGAN引入隐空间中的“风格”概念,通过对隐向量的风格化处理来控制生成图像的风格和特征。这种处理方式能够将全局风格与局部风格区分开来,从而对图像风格和细节的调整与控制更加独立和灵活^[6-7]。

1.2 虚拟试衣

虚拟试衣技术可使消费者在线上试穿心仪的服装,无需亲自前往实体店,节省了时间和精力。例如,VITON(Virtual Try-On Network)^[8]和CP-VTON(Characteristic-Preserving Virtual Try-On Network)^[9]是通过学习服装图像变形和合成以适应目标对象的虚拟试衣方法,但二者生成的人物肢体较为模糊,衣服纹理也容易产生过度形变,无法处理人物肢体与衣服有交叉的情况;VTNFP(Virtual Try-on Network with Body and Clothing Feature Preservation)^[10]和ACGPN(Adaptively Conditional Generative Pyramid Network)^[11]通过引入提取分割图像模块,更好地保留了身体和服装特征,可以生成更为生动的试衣图像;VITON-HD(Visual Inference for Textured Neural Avatars in High Definition)^[12]和HR-VITON(High-Resolution Virtual Try-On)^[13]可基于高分辨率图像进行虚拟试衣,减少了由于传统翘曲和分割图之间不一致而导致的失真现象。

1.3 文本引导的图像编辑

目前已有许多利用CLIP进行文本引导图像操作的研究。例如,文献[2]提出的TediGAN利用视觉语义相似性模块将图像和文本投影到一个共同的嵌入空间中,在将人体着装图像反演到联合隐空间后,通过更改服装的文本嵌入实现类型编辑;文献[3]结合CLIP与StyleGAN的优势,通过隐空间优化、隐空间映射、全局方向优化3种方式提出基于文本驱动生成图像的方法StyleCLIP;文献[4]提出的HairCLIP改进了StyleCLIP的隐空间映射方法,专注于使用任意文本输入编辑发型。以上方法在编辑人脸和动物脸部方面取得了显著效果,但在处理人体方面遇到了困难,原因是人类服装、身体形态和姿势的变化更加丰富^[14-19]。

为获得高度逼真的人体着装图像,本文提出一种新颖的基于分层风格映射与全局自注意力的文本引导人体着装图像编辑网络框架PSPCLIP,使用改进后的层级式卷积编码器pSp(Pixel2style2pixel)替代传统的全连接层语义映射网络,其可将人体着装图像与文本提示词直接编码到预训练的无条件GAN隐空间中,以实现给定的文本条件编辑^[20]。本研究主要贡献为:①使用层级式编码器结构,并添加基于空间自注意力机制的Transformer层,解决了卷积操作难以捕捉远程依赖关系的问题,利用图像—文本混合特征指导隐编码级别的编辑,增强图像内容的连贯性和一致性^[21];②在编码器中引入分层风格映射模块,将不同特征转化为对应层级的风格向量,实现对图像细节的优化;③优化编码器训练时使用的对比损失,消除了原对比损失计算时图像文本之间的语义差,提高了编辑后图像的语义相关性。在VITON数据集上进行实验,定性和定量结果均显示PSPCLIP所编辑的图像在具有高语义相关性的同时保持了非编辑区域的不变性,性能指标相较对照方法明显改进。

2 PSPCLIP模型构建

2.1 模型整体结构

给定一张原始的人体着装图像 I_s 和一个文本提示词 t ,目标是将 I_s 转换为包含文本信息 t 的目标图像 I_t 。为使文本信息指导人体图像编辑,通过结合多模态预训练模型CLIP、经过改进的编码器网络TranspSp以及固定、预训练的无条件图像生成器来实现一个图像到图像的转换架构PSPCLIP。编码器直接将给定的源图像和经过多模态预训练模型处理的文本编码为一组风格向量(Style Vector),这些风格向量捕获和表示期望中的图像转换,生成器利用这些表示转换的风格向量生成目标图像。PSPCLIP整体结构如图1所示。

CLIP是一个多模态模型,包含1个图像编码器和1个文本编码器,分别将图像和文本编码为512维的嵌入向

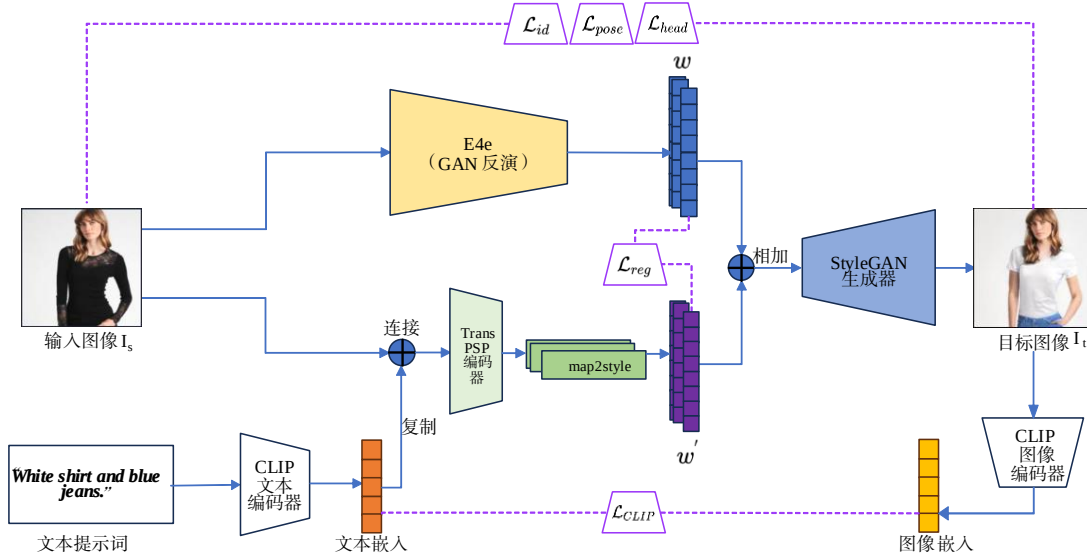


Fig. 1 Overall structure of PSPCLIP

图1 PSPCLIP整体结构

量^[1]。其采用典型的对比学习框架,最小化正确图像文本对编码向量之间的余弦距离并最大化错误对的余弦距离。CLIP通过从互联网上收集的4亿张图像-文本对进行预训练,从而学习到一个共享的图像-文本嵌入空间,很好地衡量图像与文本之间的语义相似性。PSPCLIP使用CLIP进行文本条件调制和对比损失的计算。

若要将文本提示词作为条件信息注入源图像对其进行符合期望的编辑,首先使用基于对比学习的多模态预训练模型CLIP的文本编码器将文本提示词 t 编码为512维的条件嵌入向量 e_t 。表示为:

$$e_t = \text{CLIP}(t) \quad (1)$$

然后将编码后的文本嵌入复制后作为一个通道添加到源图像 I_s 上,得到一个四通道输入张量 x_t 。表示为:

$$x_t = I_s \parallel e_t \quad (2)$$

PSPCLIP使用StyleGAN反演方法E4e获得源图像在 $\mathcal{W}$ +空间中的隐变量编码 w ^[22]。表示为:

$$w = \text{E4e}(I_s) \in \mathbb{R}^{18 \times 512} \quad (3)$$

使用式(2)和式(3)将PSPCLIP模型的输出定义为:

$$\text{PSPCLIP}(x_t) = G(\text{Transpsp}(x_t) + w) \quad (4)$$

式中: $\text{Transpsp}(\cdot)$ 表示PSPCLIP输出隐编码位移的编码器; $G(\cdot)$ 表示StyleGAN生成器。

最终输出表示为:

$$I_t = \text{PSPCLIP}(x_t) \quad (5)$$

PSPCLIP使用E4e($\cdot$)和Transpsp($\cdot$)两个编码器,其中E4e($\cdot$)用于反转源图像以得到初始隐变量编码;Transpsp($\cdot$)用于学习初始隐变量编码与包含文本信息的目标隐变量编码之间的偏差。预训练的E4e编码器在整个训练过程中保持固定,而Transpsp($\cdot$)使用损失函数进行训练。

2.2 TransPSP编码器

在文本条件下的人体着装图像编辑领域,高质量的图

像重建与编辑成果依赖于深度学习模型对图像细节与复杂性的准确捕捉能力。PSPCLIP框架建立在预训练的StyleGAN生成器及其 $\mathcal{W}$ +隐空间的表示能力基础上,目的是通过引入文本描述来指导图像编辑过程,其关键在于设计一个合适的编码器,从而将输入图像与相应的文本提示词融合,映射到 $\mathcal{W}$ +隐空间中一个准确的隐向量上。在最初方案中,编码器网络通过其最终层将输入图像转化为一个单一的512维向量,作为图像的编码表示嵌入至 $\mathcal{W}$ +隐空间中。尽管这种方法为学习StyleGAN中的18个风格向量提供了一种简便路径,但同时也暴露出一个显著问题,即难以充分捕获原始图像中的细节与复杂性,这一限制最终影响了重建图像的质量。该问题的本质在于单个512维向量的表示能力有限,不足以全面捕捉图像的高度复杂性与细腻细节,因此在图像重建与编辑过程中可能会丢失一些关键的原始图像特征,导致编辑后的图像与期望文本描述之间存在差距,从而降低编辑结果的准确度与自然度。为解决这一问题,本研究探讨了更为复杂的编码器架构以及更高维度向量表示的可能性。通过增强编码器的表示能力,可以更有效地捕获和编码输入图像的细节与复杂性,进一步提升图像编辑的精确度和质量。

在深入探究StyleGAN的内部机制时,研究者们已经识别出不同风格输入所对应的细节层次差异,这些层次通常被分为粗糙级、中等级和细致级3个层次^[23]。这种划分方法揭示了StyleGAN在生成过程中如何逐步细化图像特征,从而实现从全局结构到局部细节的渐进式精细化。TransPSP通过在其编码器主干中融入特征金字塔结构设计了一种能够产生三级特征图的方法,不仅能促进不同细节层次信息的有效捕获,而且为后续处理步骤提供了丰富的特征基础。此外,TransPSP在特征金字塔与解码器之间嵌入了基于全局自注意力机制的Transformer层,增强了模型对于全局依赖关系的理解与处理能力,使其在处理图像

不同细节层次时更为精准和高效。

为了精细调控风格向量以实现目标图像编辑, TransPSP 编码器采用本文提出的名为 map2style 的中间网络。结构如图 2 所示。该网络专门用于从特征图中提取风格向量的位移信息。结合全局自注意力机制, map2style 网络能

够准确地从不同层次的特征图中提炼出对应的风格位移向量。这些位移向量随后与原始图像各自层次的风格向量相结合, 并被适配地输入到生成器的相应层, 从而精准驱动图像编辑过程, 保证不同细节层次的准确捕捉与表达。

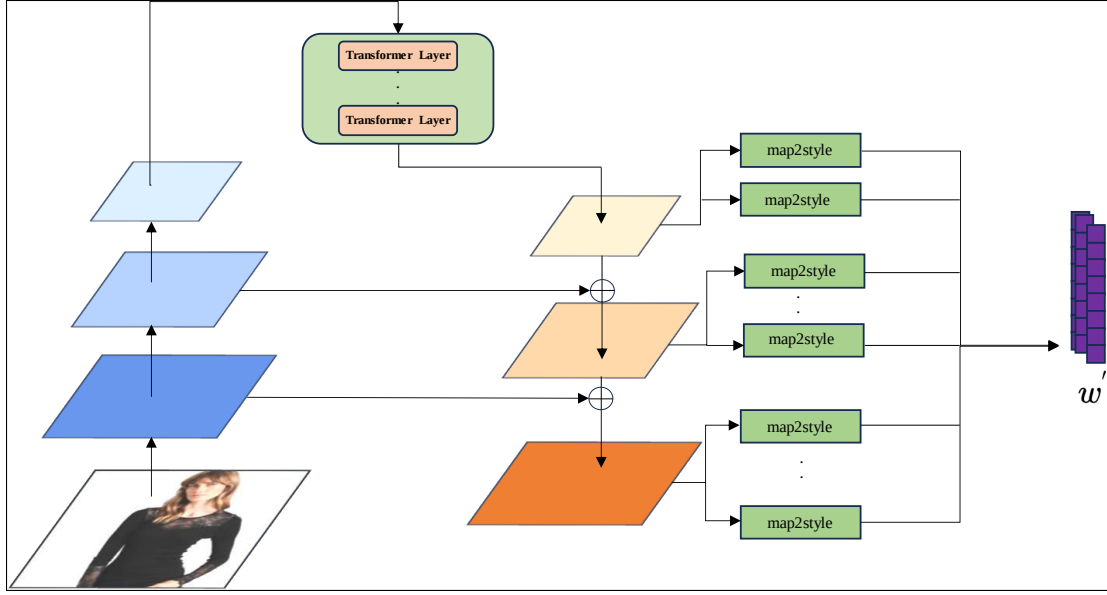


Fig. 2 Structure of TransPSP encoder

图2 TransPSP 编码器结构

2.3 分层风格映射模块

map2style 网络是本文提出的创新性分层风格映射模块, 是一个精巧设计的全卷积网络, 专门用于从多尺度特征图中提取目标风格的向量。该网络采用一系列步长为 2 的卷积层, 配合 LeakyReLU 激活函数, 实现了对特征图空间尺寸的逐步降维, 优化了特征提取过程。在初始阶段, map2style 网络依托于 ResNet 架构的骨干网络, 利用标准的特征金字塔技术进行高效的特征提取, 以确保从低级到高级的特征都能被综合利用^[24]。

对于 StyleGAN 中定义的 18 种目标风格, map2style 网络为每种风格单独训练了一个专门的小型映射网络, 从而确保对每种风格特征的精确捕捉与提取。这种定制化训练方法允许网络根据特征图尺寸和复杂度调整其提取策略, 从而实现更为精细和有针对性的风格向量生成。具体来说, 风格向量的提取根据特征图尺寸进行了细致的分层处理, 其中尺寸最小的特征图(0~2层)对应图像的粗糙级风格特征, 中等尺寸特征图(3~6层)对应中等级风格特征, 最大尺寸特征图(7~18层)对应细致级风格特征。每次提取出的 512 维风格向量位移在与相应层级原始风格向量进行叠加后被输入到 StyleGAN 生成器中, 以生成目标风格图像。这一过程不仅提高了生成图像风格的一致性, 而且改善了其细节和总体质量。

2.4 损失函数

PSPCLIP 的编码器使用数个损失目标的加权组合进行训练。

2.4.1 身份损失

使用预训练的 ArcFace 网络 $\text{Arc}(\cdot)$ 在 ArcFace 嵌入空间中计算原始图像 I_s 与编辑后图像 I_t 之间的余弦距离, 以尽可能维持人脸身份不变^[25]。身份损失 (Identity Loss) 函数公式为:

$$\mathcal{L}_{id} = 1 - \cos(\text{Arc}(I_s), \text{Arc}(I_t)) \quad (6)$$

2.4.2 对比损失

传统的对比损失 (CLIP Loss) 计算过程为将编辑后图像与目标文本描述一起编码到联合嵌入空间中, 然后计算它们之间的余弦距离。然而在本文任务中, 文本描述只指代人体图像的服装部分而非整体, 直接以全局方式计算损失会导致模型忽略一部分细节。因此, 本文提出针对服装编辑的对比损失函数, 利用 CLIP 嵌入支持的线性操作使模型的训练重点尽量只集中在可修改的服装区域^[26]。将原始图像在联合空间中的嵌入减去对应原始训练数据中的服装标签 t_s 嵌入, 从而得到不可编辑区域的联合空间隐编码, 即 $C_{t_m}^i = C_{t_i}^i - C_{t_s}^i$; 然后将不可编辑区域的嵌入与目标文本提示词 t 所对应的嵌入相加, 即可得到目标编辑图像的真实 CLIP 嵌入, 即 $C_t^i = C_{t_m}^i + C_t^i$ 。最终得到更准确的对比损失。表示为:

$$\mathcal{L}_{CLIP} = 1 - \cos(C^i(I_t), C^i(t')) \quad (7)$$

式中: C^i 和 C' 分别表示 CLIP 模型的图像编码器和文本编码器。

2.4.3 正则化损失

对源图像与目标图像的隐编码之差 Δw 应用 $\mathcal{L}_2$ 正则化损失 (Regularization Loss) 来实现稳定的训练, 避免因产生过大的偏移量而导致过度编辑。损失函数公式为:

$$\mathcal{L}_{\text{reg}} = \|\Delta w\|_2 \quad (8)$$

2.4.4 姿态损失

为尽可能保留非编辑区域, 尤其是对输出图像中所有身体部位的位置进行保留, 本文提出姿态损失 (Pose Loss) 来实现详细和准确的姿态保留, 包括各个身体部位也保持其形状和大小不变。该损失函数的计算使用 DensePose 模型, 其以与服装无关的方式预测各个身体部位的位置和形状^[27]。损失函数公式为:

$$\mathcal{L}_{\text{pose}} = \frac{1}{N_D} \sum_{j=1}^{N_D} \|D_j(I_t) - D_j(I_s)\|_2^2 \quad (9)$$

式中: N_D 表示解析的身体部位数量; D 表示姿态解析模型 DensePose; $D_j(\cdot)$ 表示第 j 个身体部位的解析掩码。

最终, 将整体损失函数定义为:

$$\mathcal{L} = \lambda_{\text{CLIP}} \mathcal{L}_{\text{CLIP}} + \lambda_{\text{id}} \mathcal{L}_{\text{id}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}} + \lambda_{\text{pose}} \mathcal{L}_{\text{pose}} \quad (10)$$

式中: $\lambda_{\text{CLIP}}、\lambda_{\text{id}}、\lambda_{\text{reg}}、\lambda_{\text{pose}}$ 分别为对应各个损失函数的平衡权重。

3 实验方法与结果分析

3.1 模型训练与参数设置

全部实验均在 1 张具有 12 G 内存的 NVIDIA Laptop RTX4080 GPU 上进行。GAN 反演编码器 E4e 的编码器在 VITON 数据集上进行训练, 使用预训练的 StyleGAN2 作为目标 GAN 模型进行反演^[7]。CLIP 模型的图像文本编码器使用预训练的 CLIP RN50×4 网络结构。关于训练策略, 在默认情况下, PSPCLIP 训练的编码器学习率设置为 0.5, 在训练进行到总轮数的一半时, 学习率线性减小至 0; 批处理大小设置为 4, 训练迭代次数为 50 000 次; 使用 Adam 优化器, 其中 β_1 和 β_2 分别设置为 0.9 和 0.999^[28]。损失函数权重设置参照文献[4], 即 $\lambda_{\text{CLIP}}、\lambda_{\text{id}}、\lambda_{\text{reg}}、\lambda_{\text{pose}}$ 分别为 1.0、0.1、0.8、0.5。后续参与比较的所有方法使用官方训练的代码或预训练模型。

3.2 实验数据集

PSPCLIP 编码器训练与评估均在 VITON 数据集上进行。该数据集共包括 16 253 组图像, 每组图像包括 1 张女性正面图、1 张服装图像、1 张掩码、1 张原始语义图和 1 组参考姿势点, 每张图像大小为 256×192 像素。在这 16 253 组图像中, 有 14 221 组被用作训练集, 2 032 组用作测试集。在训练前, 数据集中的图像首先被裁剪为 192×192 像素, 去除图像底部后, 使用双线性插值将其调整为 256×256 像素; 然后根据 StyleGAN-Human 中提到的方法, 使用 E4e 模型将预处理过的数据集反演转化为隐编码^[29]。

3.3 评价指标

为定量评估模型生成图像的性能, 定义 4 个评价指标, 涵盖了编辑过程的不同方面。

3.3.1 语义相关性

编辑后的图像应包含与输入文本提示词相关的语义。语义相关性 (Semantic) 度量了编辑图像与目标文本提示词之间的 CLIP 相似度得分^[30]。计算公式为:

$$\text{Semantic} = \cos(C^i(I), C^t(t)) \quad (11)$$

式中: C^i 和 C^t 分别为 CLIP 模型的图像编码器和文本编码器; I 表示编辑后的图像; t 表示目标文本提示词。

3.3.2 交并比

编辑后的图像应保持输入人物的姿势。交并比 (Intersection over Union, IoU) 可用于评估姿势保持能力^[31]。实验使用 DensePose 进行姿势预测, 并计算原始图像与编辑图像之间姿势预测的 IoU 得分。计算公式为:

$$\text{IoU} = \frac{I_s \cap I_t}{I_s \cup I_t} \quad (12)$$

式中: I_s 表示原始图像; I_t 表示编辑后图像。

3.3.3 身份相似性

编辑模型应保留输入图像的身份和面部外貌。使用 ArcFace 提取人脸嵌入向量^[32]。为量化输入图像与编辑图像之间的身份相似性 (Identity Similarity), 计算人脸嵌入向量之间的余弦相似度^[25]。计算公式为:

$$\text{Identity sim.} = 1 - \cos(\text{Arc}(I_s), \text{Arc}(I_t)) \quad (13)$$

3.3.4 图像保真度

编辑模型应合成与输入相媲美的高保真度图像。图像保真度 (Frechet Inception Distance, FID) 可衡量生成图像的质量。计算公式为:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + T_r \left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}} \right) \quad (14)$$

式中: T_r 为伴随矩阵; μ 为平均值; Σ 为协方差。

3.4 实验结果与分析

3.4.1 PSPCLIP 编辑能力评估

图 3 评估了 PSPCLIP 在多个测试图像以及多个文本提示词上表现出的编辑能力, 这些图像具有不同的姿势特征和初始服装风格。可以看出, PSPCLIP 具有出色的编辑能力, 具体表现为: ①能合成较复杂的服装风格。例如, 在第 4 列中展示了迷彩图案, 表明 PSPCLIP 能合成具有复杂设计的服装风格; ②真实地融入文本语义。编辑后的图像能够准确表达文本描述的语义, 编辑结果与文本提示词所表达的意思相符; ③保留被编辑人体的姿势特征和身份。PSPCLIP 能够保持编辑后图像中被编辑主体的姿势和身份特征, 确保其可识别性和一致性; ④添加或删除初始服装的局部区域。PSPCLIP 能够添加或删除初始服装的袖子或裤子, 还能虚构出起初被遮挡的物体, 例如手臂, 使编辑后的图像更加完整和逼真; ⑤无需显式三维建模的贴合。PSPCLIP 能使编辑后的服装与真实服装特征在一定

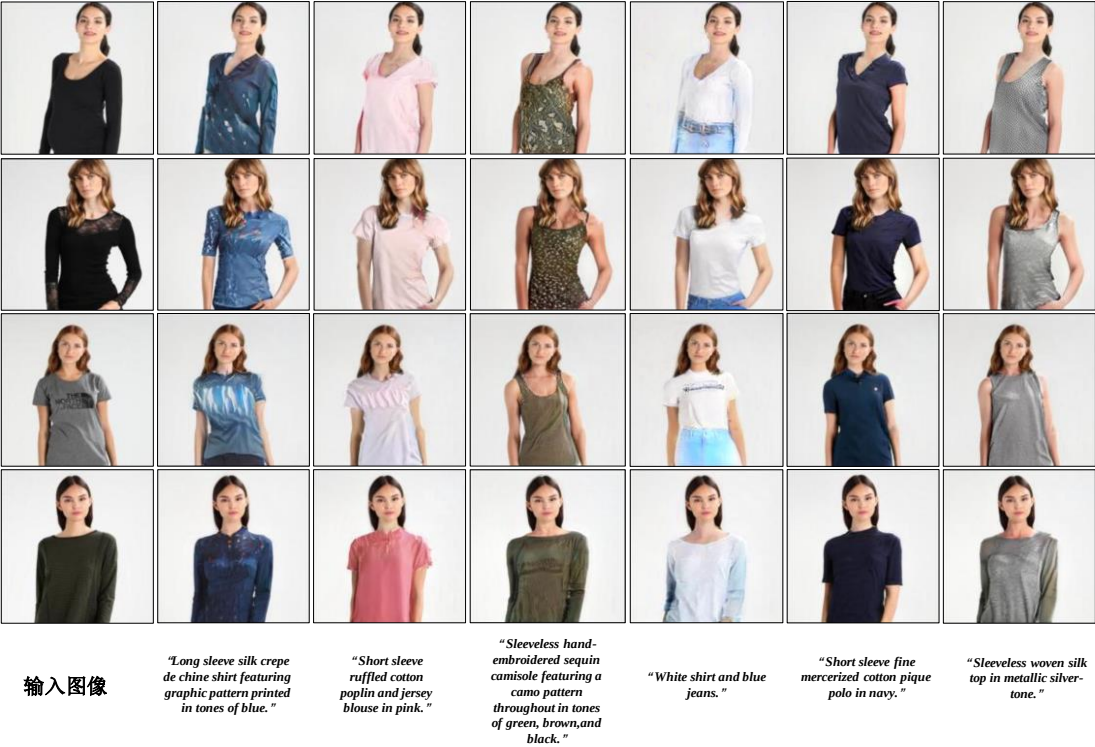


Fig. 3 PSPCLIP editing result example

图3 PSPCLIP编辑结果示例

程度上保持一致,如衣服上的褶皱,从而无需进行显式三维建模。

3.4.2 定性结果比较

将 PSPCLIP 与现有最先进的方法 TediGAN^[2]、StyleCLIP^[3]、HairCLIP^[4]进行可视化比较。考虑两种不同情况,分别为:①相同图像—不同文本。在该情况下,将所有模型应用于相同输入图像,并将其与不同目标文本描述进行配对,以展示模型如何处理不同类型的语义;②相同文本—不同图像。在该情况下,使用不同输入图像,并将其与相同目标文本描述进行配对,以探索模型在不同输入下编辑的一致性。

相同图像—不同文本实验结果如图4所示。可以看出,对照模型在保持身份方面不如PSPCLIP,原因有二:首先,当图像嵌入到GAN隐空间时,生成器的重建不够完美,会丢失一些身份信息;其次,隐编码空间的纠缠导致编辑后的图像发生数个与目标相悖的改变。对照模型的编辑结果展现出一定的纠缠问题,如图4第3行中红色衣服的变化也使得口红颜色加深。尽管HairCLIP整合了一种解纠缠机制,隐编码的变化仍然影响了人的身份。PSPCLIP由于在训练时使用了身份损失函数,保持了编辑图像的身份。此外,PSPCLIP在处理某些文本提示词时表现更好,例如图4第1行和第3行中可以看到更体现编辑精细度的衬衫领口。虽然对照模型编辑的大多数图像可以一定程度地反映目标文本的语义,但PSPCLIP模型的编辑结果在视觉上明显最令人信服。

相同文本—不同图像实验结果如图5所示。可以看

出,对照模型在人物的精确定位方面经常出现问题,与原始图像相比人物经常发生位移。图5中TediGAN模型的结果最能说明这一点,主要原因为TediGAN在训练编码器模型时使用了感知损失,这种损失通常以牺牲精确的物体定位为代价提高图像保真度。虽然这种权衡对于其他图像处理任务是合理的,但是对于需要精确保持输入人物姿势和位置的人体服装编辑任务来说不是最佳选择。PSPCLIP整合了一个保持姿势的损失项和一个隐编码级损失项,以避免该类定位问题。总体来说,所有测试模型均能生成与期望语义比较一致的编辑,但除了PSPCLIP外,对照模型常常引入相当大的视觉伪影。

3.4.3 定量结果比较

表1为StyleCLIP、HairCLIP、TediGAN和PSPCLIP的性能指标比较结果。可以看出,PSPCLIP在各个指标上均获得最佳表现。

3.4.4 损失消融实验

对PSPCLIP的损失函数进行消融实验,以研究不同损失项的贡献。具体来说,将相应的损失权重设置为0来消融3个不同损失项,即姿态保持项 $\lambda_{pose}=0$,身份保持项 $\lambda_{id}=0$ 以及隐编码正则化项 $\lambda_{reg}=0$ 。消融实验保留与语义相关的CLIP损失函数,因为其在基于文本的编辑任务中起着关键作用。损失消融实验结果见表2和图6。可以看出,身份保持项的加入使所有指标产生显著改进。该项通过确保面部构成不会产生不自然的视觉伪像,从而保持了编辑后图像的身份特征,使生成图像更自然。隐代码正则化使语义相关性得分轻微降低,但有助于提高图像保真



Fig. 4 Same image-different text editing effects

图 4 相同图像—不同文本编辑效果

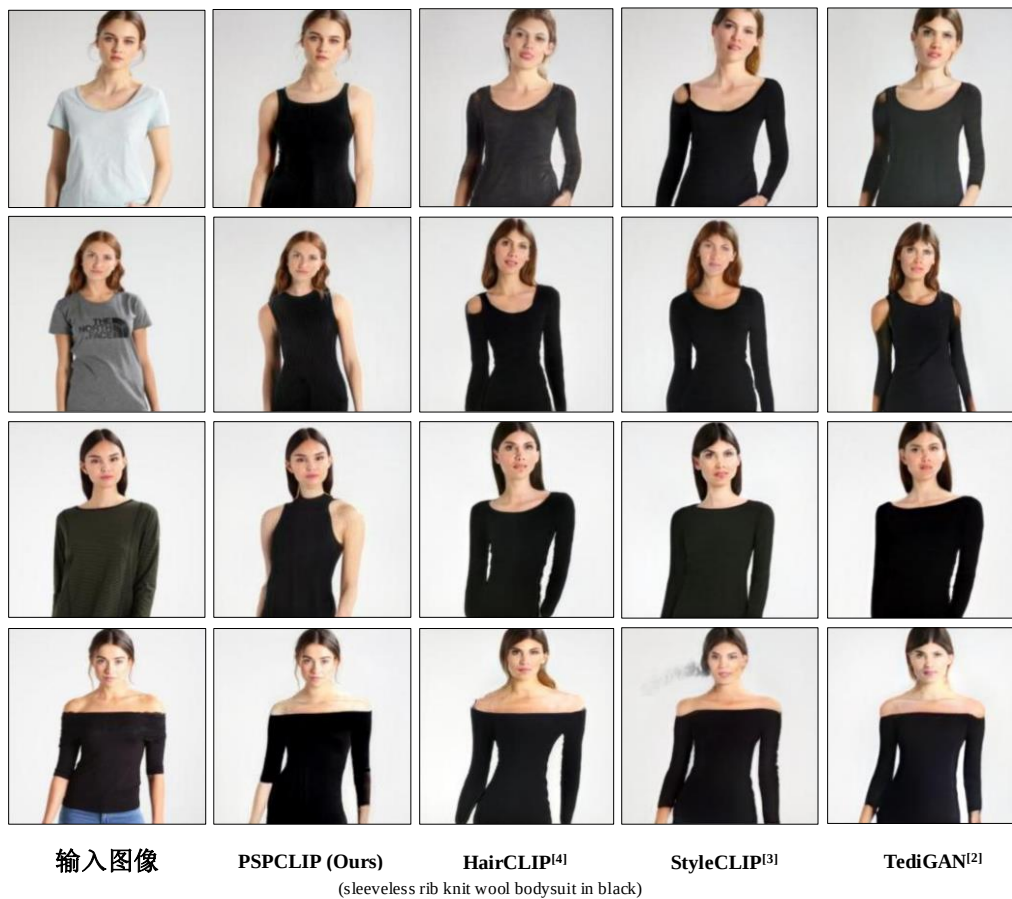


Fig. 5 Same text - different image editing effects

图 5 相同文本—不同图像编辑效果

Table 1 Comparison of performance indicators of different methods

表 1 不同方法性能指标比较

模型	Semantics(↑)	Identity sim.(↑)	IoU(↑)	FID(↓)
HairCLIP ^[4]	0.402	0.61	0.812	85.57
StyleCLIP ^[3]	0.390	0.54	0.834	90.25
TediGAN ^[2]	0.298	0.17	0.717	95.44
PSPCLIP	0.471	0.85	0.940	64.77

度。对隐代码进行正则化能够控制生成器对编辑操作的响应程度,从而减少不必要的变化,提高生成图像的一致性和细节保留程度。姿态保持项亦使语义相关性得分轻

微降低,但显著提高了 IoU 指数,并进一步改善了 FID 得分,说明姿态保持项能够保持编辑后图像的姿态一致性。

Table 2 Loss ablation experiment results

表 2 损失消融实验结果

损失函数	Semantics(↑)	Identity sim.(↑)	IoU(↑)	FID(↓)
$\mathcal{L}_{\text{CLIP}} \#$	0.453	0.82	0.781	82.52
$+\mathcal{L}_{\text{id}}$	0.466	0.92	0.888	68.94
$+\mathcal{L}_{\text{reg}}$	0.450	0.92	0.884	65.66
$+\mathcal{L}_{\text{pose}}$	0.446	0.93	0.939	64.91
全部	0.471	0.85	0.940	64.77

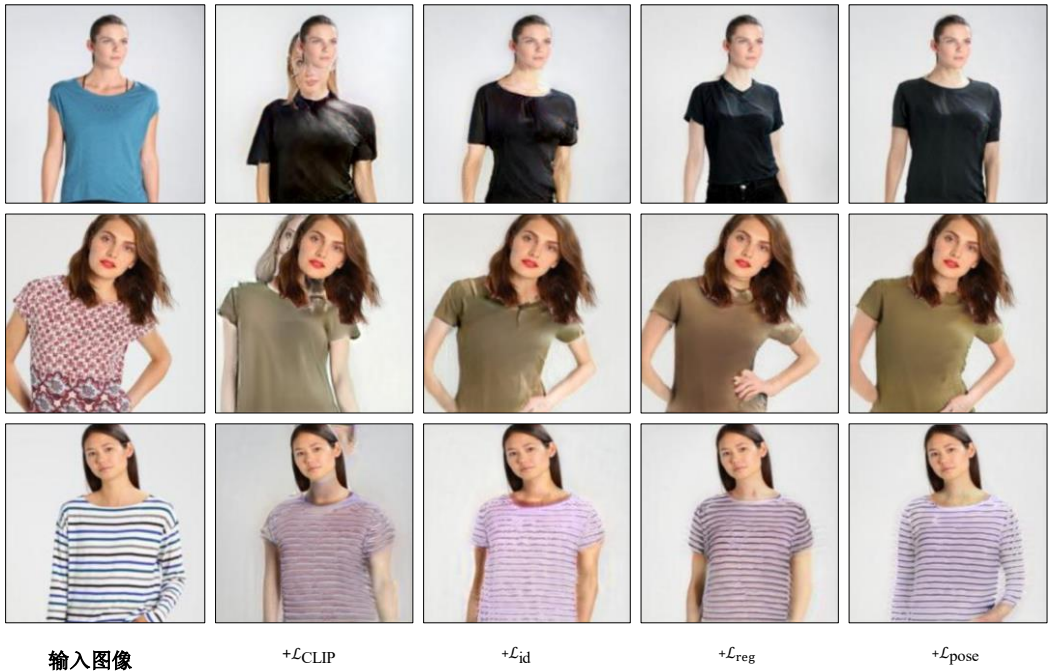


Fig. 6 Visualization results of loss ablation experiment

图 6 损失消融实验可视化结果

3.4.5 模块消融实验

对TransPSP编码器进行模块消融实验,以验证引入的Transformer和分层风格映射模块map2style的有效性。具体来说,分别移除Transformer模块和map2style模块,并与基础pSp编码器和完整TransPSP编码器进行定量比较。实验结果如表3所示。可以看出,在VITON数据集上,两个模块相较基础pSp编码器模块在语义相似度和图像生成质量方面均有提升。主要原因在于引入的Transformer层增强了模型对于图像—文本混合特征全局依赖关系的理解与处理能力;而map2style模块对每一种风格单独训练专用的小型映射网络,对不同层次的风格特征进行精确捕捉与提取,提升了模型生成图像的细节质量,但也小幅度降低了语义准确度。

Table 3 Quantitative results of module ablation study

表 3 模块消融研究定量结果

模块	Semantics(↑)	Identity sim.(↑)	IoU(↑)	FID(↓)
pSp	0.456	0.81	0.939	71.95
w/o transformer	0.453	0.84	0.937	65.59
w/o map2style	0.475	0.81	0.941	69.88
完整TransPSP	0.471	0.85	0.940	64.77

4 结语

本文提出一种新颖的文本引导的人体着装图像编辑模型PSPCLIP。其基于分层风格映射和全局自注意力机制模型,充分利用图像—文本混合特征的全局上下文,提高了编辑后图像的一致性和连贯性。其以一种层次化方式提取风格向量,并输入到预训练的生成器中,丰富了生成结果的外观细节,进一步提高了编辑效果。通过对原始对比损失进行优化,消除了目标图像与文本提示词之间的语义差,使目标图像更符合文本语义。与几种对照方法进行比较后发现,PSPCLIP在定性实验表现和Semantics、Identity sim.、IoU、FID量化指标方面均取得显著改进,能够在文本引导下产生高质量编辑结果。然而该模型继承了CLIP在文本描述长度上的限制(76个token)。此外,PSPCLIP在袖子长度以及服装复杂对象等方面存在表达困难的问题,一是由于StyleGAN生成能力有限,二是CLIP知识有限。解决这个问题的一个可能方案是扩展GAN训练数据集并在大型文本—图像时尚数据集上对CLIP进行

微调。

参考文献:

- [1] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [C]//International Conference on Machine Learning, 2021: 8748–8763.
- [2] XIA W, YANG Y, XUE J H, et al. Tedigan: text-guided diverse face image generation and manipulation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 2256–2265.
- [3] PATASHNIK O, WU Z, SHECHTMAN E, et al. Styleclip: text-driven manipulation of stylegan imagery [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 2085–2094.
- [4] WEI T, CHEN D, ZHOU W, et al. Hairclip: design your hair by text and reference image [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 18072–18081.
- [5] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C]// Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014: 2672–2680.
- [6] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 4401–4410.
- [7] 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, 2020: 8110–8119.
- [8] HAN X, WU Z, WU Z, et al. Viton: an image-based virtual try-on network [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7543–7552.
- [9] WANG B, ZHENG H, LIANG X, et al. Toward characteristic-preserving image-based virtual try-on network [C]//Proceedings of the European Conference on Computer Vision, 2018: 589–604.
- [10] YU R, WANG X, XIE X. Vtnfp: an image-based virtual try-on network with body and clothing feature preservation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 10511–10520.
- [11] SONG D, LI T, MAO Z, et al. SP-VITON: shape-preserving image-based virtual try-on network [J]. Multimedia Tools and Applications, 2020, 79: 33757–33769.
- [12] YANG H, ZHANG R, GUO X, et al. Towards photo-realistic virtual try-on by adaptively generating-preserving image content [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 7850–7859.
- [13] CHOI S, PARK S, LEE M, et al. Viton-hd: high-resolution virtual try-on via misalignment-aware normalization [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 14131–14140.
- [14] ABDAL R, ZHU P, FEMIANI J, et al. Clip2stylegan: unsupervised extraction of stylegan edit directions [C]//ACM SIGGRAPH 2022 Conference Proceedings, 2022: 1–9.
- [15] KIM G, KWON T, YE J C. Diffusionclip: text-guided diffusion models for robust image manipulation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 2426–2435.
- [16] GAL R, PATASHNIK O, MARON H, et al. StyleGAN-NADA: CLIP-guided domain adaptation of image generators [J]. ACM Transactions on Graphics (TOG), 2022, 41(4): 1–13.
- [17] WANG H, LIN G, DEL MOLINO A G, et al. Maniclip: multi-attribute face manipulation from text [DB/OL]. <https://arxiv.org/abs/2210.00445>.
- [18] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with clip latents [DB/OL]. <https://arxiv.org/abs/2204.06125>.
- [19] WU Z, LISCHINSKI D, SHECHTMAN E. Stylespace analysis: disentangled controls for stylegan image generation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 12863–12872.
- [20] RICHARDSON E, ALALUF Y, PATASHNIK O, et al. Encoding in style: a stylegan encoder for image-to-image translation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 2287–2296.
- [21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6000–6010.
- [22] TOV O, ALALUF Y, NITZAN Y, et al. Designing an encoder for stylegan image manipulation [J]. ACM Transactions on Graphics (TOG), 2021, 40(4): 1–14.
- [23] 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, 2020: 8110–8119.
- [24] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770–778.
- [25] DENG J, GUO J, XUE N, et al. Arcface: additive angular margin loss for deep face recognition [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 4690–4699.
- [26] COUAIRON G, GRECHKA A, VERBEEK J, et al. Flexit: towards flexible semantic image translation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 18270–18279.
- [27] GÜLER R A, NEVEROVA N, KOKKINOS I. Densepose: dense human pose estimation in the wild [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7297–7306.
- [28] KINGMA D P, BA J. Adam: a method for stochastic optimization [DB/OL]. <https://arxiv.org/abs/1412.6980>.
- [29] FU J, LI S, JIANG Y, et al. Stylegan-human: a data-centric odyssey of human generation [C]//European Conference on Computer Vision, 2022: 1–19.
- [30] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [C]//International Conference on Machine Learning, 2021: 8748–8763.
- [31] EVERINGHAM M, ESLAMI S M A, VAN GOOL L, et al. The pascal visual object classes challenge: a retrospective [J]. International Journal of Computer Vision, 2015, 111: 98–136.
- [32] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6629–6640.

(责任编辑:尹晨茹)