浙江理工大学学报,第51卷,第5期,2024年9月
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673-3851(n).2024.05.012

基于三维人体姿态信息的协作式扩散试衣生成网络

丁　焕[a],侯　珏[a,b],杨　阳[a,b],刘　正[b,c]
(浙江理工大学,a.服装学院;b.丝绸文化传承与产品设计数字化技术
文旅部重点实验室;c.国际时装技术学院,杭州 311199)

摘　要:为解决人体关节遮挡产生的试衣错误、试衣生成结果保真度低以及试衣前后人体形态保持不一致等问题,提出了一种基于三维人体姿态信息的协作式扩散试衣生成网络—3DM-VTON网络。首先,通过三维人体姿态估计模型重构二维人体图像对应的三维人体姿态信息,获取人体各部位的空间顺序。其次,利用多部件服装变形网络,根据服装部件的外观流场对目标服装各部位进行变形,并在人体部位空间顺序的指导下对各服装部件特征进行融合。最后,将人体全局特征与融合后的服装特征引入扩散网络,通过人体姿态控制模块与区域度判别器限制其生成自由度,保证试衣前后人体形态一致性,从而提升图像局部细粒度,获得高质量的虚拟试衣结果。将3DM-VTON网络与目前常用的二维虚拟试衣网络进行实验对比,结果表明:3DM-VTON网络在人体关节遮挡处的服装变形准确度更高,试衣结果更加清晰、自然和逼真,试衣前后人体形态保持更一致。3DM-VTON网络能够生成高质、多姿势的试衣结果,满足虚拟试衣用户更多的试穿需求,可以提升消费者在线购物体验度,助推服装电商行业发展。

关键词:虚拟试衣;生成对抗网络;扩散网络;外观流场;服装部件分割
中图分类号:TS942.8　　**文献标志码**:A　　**文章编号**:1673-3851(2024)09-0691-14
引文格式:丁焕,侯珏,杨阳,等. 基于三维人体姿态信息的协作式扩散试衣生成网络[J]. 浙江理工大学学报(自然科学),2024,51(5):691-704.
Reference Format: DING Huan, HOU Jue, YANG Yang, et al. A collaborative diffusion-based network for virtual try-on based on 3D human pose information[J]. Journal of Zhejiang Sci-Tech University,2024,51(5):691-704.

A collaborative diffusion-based network for virtual try-on based on 3D human pose information

DING Huan[a], *HOU Jue*[a,b], *YANG Yang*[a,b], *LIU Zheng*[b,c]
(a. School of Fashion Design & Engineering; b. Key Laboratory of Silk Culture Inheritance and Products Design Digital Technology, Ministry of Culture and Tourism; c. International Institute of Fashion Technology, Zhejiang Sci-Tech University, Hangzhou 311199, China)

Abstract: To address the issues of clothing errors caused by joint occlusion, low fidelity of try-on results, and poor consistency in human body shape before and after try-on, a collaborative diffusion-based virtual try-on network using 3D human body pose information, termed 3DM-VTON, was proposed. Firstly, a 3D human body pose estimation network was used to reconstruct the 3D pose information corresponding to the 2D human images, and the spatial order of different body parts was acquired. Secondly, a multi-part garment deformation network was used to deform each part of the target garment based on local appearance flow fields and to integrate the features of each garment component under the guidance of the spatial relationship of body parts. Finally, the garment features integrating global human features with local garment features were introduced into the diffusion-based network, and the generation

收稿日期:2024-04-15　　网络出版日期:2024-07-22
基金项目:浙江理工大学科研启动基金项目(21072325-Y);嘉兴市重点研究计划项目(2023BZ10009)
作者简介:丁　焕(2000—　),男,浙江绍兴人,硕士研究生,主要从事虚拟试衣方面的研究。
通信作者:刘　正,E-mail:koala@zstu.edu.cn

freedom of the diffusion network was constrained by a human body pose control network and a regional fidelity discriminator, ensuring consistency in human body shape before and after try-on, enhancing local granularity of the image, and achieving high-quality virtual try-on results. Comparative experiments with 3DM-VTON and the currently commonly used 2D virtual try-on network showed that 3DM-VTON achieved higher accuracy in garment deformation at joint occlusions, provided clearer, more natural, and realistic overall try-on results, and maintained better consistency in human body shape before and after try-on. The 3DM-VTON can generate high-quality, multi-posture try-on results to meet the needs of virtual try-on users, which can improve consumers' online shopping experience and boost the development of the clothing e-commerce industry.

Key words: virtual try-on; generative adversarial network; diffusion network; appearance flow; garment part segmentation

0 引 言

随着移动互联网的快速发展与普及,线上购物正逐渐成为消费者购买服装的主要选择[1]。但不同于线下选购服装,消费者无法在线上根据服装款式图进行个性化的风格搭配,不仅降低了消费者的购物满意度与服装商品成交率,还增加了服装企业的运营成本[2]。为了构建高度逼真的虚拟试衣环境,提升消费者购物体验,推动服装电商行业发展,基于人工智能技术的虚拟试衣应运而生[3]。

目前基于人工智能技术的虚拟试衣主要分为基于图像的二维试衣与基于三维测量建模的三维试衣两大类。其中,基于三维测量建模的三维试衣通过三维扫描方法获取人体数据并构建人体模型和服装模型,利用虚拟仿真技术模拟服装面料的物理特性,再对三维人体模型和服装模型进行渲染,得到最终的试衣结果[4]。虽然三维试衣能够呈现立体的着装效果,高度还原服装细节,但需要庞大的技术投入与计算成本,这在一定程度上限制了三维试衣的应用与发展[5]。与之相比,基于图像的二维试衣对硬件性能与数据样本的要求低,运行速度快,利用较少的算力便能实现良好的试衣效果,降低了应用门槛,提升了试衣效率,正逐渐成为研究热点[6-7]。

基于图像的二维试衣主要分为两个阶段:a)服装变形阶段,通过二维人体解析信息对目标服装进行非线性变换,获得与目标人物姿态贴合的服装变形图像。b)试衣生成阶段,利用生成式网络融合人体图像与变形后的服装图像,获得逼真的试衣结果。根据生成网络的不同,二维试衣可分为基于生成对抗网络(Generative adversarial network, GAN)[8]的二维试衣和基于扩散网络[9]的二维试衣。在服装变形阶段,当试衣人体出现关节遮挡以及非传统试穿姿势(如举手、叉腰等)时,相应的试衣部位会出现服装变形错误以及纹理失真问题。为解决这一问题,Lee 等[10]构建了高分辨率试衣网络(High-resolution virtual try-on, HR-VTON),在网络中引入条件生成器,对试穿人体进行二维解析信息生成,并通过一个针对性判别器对服装变形过程进行信息监督与对齐,以解决关节遮挡部位所产生的变形问题。Yang 等[11]构建了去遮挡试衣网络(De-occlusion try-on model, DOC-VTON),通过预测关节遮挡区域掩码,提升试衣生成网络对遮挡区域的捕捉能力与局部修复能力。Xie 等[12]构建了通用目标试衣网络(General purpose virtual try-on model, GP-VTON),对不同人体部位(左手、右手、上身躯干)进行局部流场估计,并融合全局信息,以更有效地应对服装错位现象。但上述方法仅通过二维图像信息来实现服装变形,构建的试衣网络缺乏试穿人体与服装之间的空间结构关系信息,因而无法精准、有效地解决该问题。

另外,在试衣生成阶段,基于生成对抗网络的二维试衣会受到生成对抗网络中固有的模式坍塌、数据拟合困难等问题,导致生成的试衣结果保真度低、图像质量整体较差,无法有效保留人体其余部位的细节[13]。扩散网络是一种通过加噪、去噪的迭代过程来学习目标分布的生成网络。与生成对抗网络相比,扩散网络拥有更高的训练稳定性,能够生成具有更高细粒度的图像,并且可以在网络生成过程中引入不同的条件特征以增强其可控生成能力,从而保持服装纹理与细节,更好地模拟实际试衣效果[14]。Zhu 等[15]通过引入平行 Unet 扩散网络架构与隐式服装变形方法,在不同尺度的隐式特征中利用交叉注意力机制来匹配目标服装与试穿人体之间的相关性,最后利用超分辨率扩散网络生成高分辨率试穿结果。Morelli 等[16]在扩散网络的基础上引入了一

个自动编码器模块,利用跳跃连接的方式将未变形的目标服装与提示文本嵌入隐式特征中,从而增强扩散网络的生成能力并控制其生成自由度。Gou 等[17]在条件填充扩散试衣生成网络(Diffusion-based conditional inpainting virtual try-on model,DCI-VTON)中引入服装变形模块,将其作为局部控制条件以增加扩散网络的可控性,并通过建立粗糙重建分支与细化重建分支,最大限度地保留生成结果中的服装纹理细节与人体细节。然而上述方法未直接将人体姿态作为控制因素对试衣生成过程进行限制,导致试衣前后人体形态保持不一致,人体其余部位细节的保真度低。

针对上述问题,本文提出了一种基于三维人体姿态信息的协作式扩散试衣生成网络—3DM-VTON 网络。该网络引入三维人体姿态估计模型以获取人体各部位的空间顺序,通过多部件服装变形方法,对变形后的服装部件进行基于人体各部位空间顺序的叠加与融合,从而缓解关节遮挡所造成的服装变形失真问题;通过服装轮廓、品类编码模块提取服装整体信息,强化网络低维感知能力,提升服装整体变形准确度;利用扩散网络对人体图像与变形后的服装图像进行特征融合,引入人体姿态控制模块提升扩散网络的整体可控性,保证试衣前后人体形态一致性,获得逼真、自然的虚拟试衣结果。将 3DM-VTON 网络与其他网络进行定量和定性实验对比,以评估该网络的性能。

1　3DM-VTON 网络设计

3DM-VTON 网络由两部分组成,分别为基于三维人体姿态信息的多部件服装变形网络(Multi-part clothing warping model, MPWM)和基于人体姿态控制模块、区域度判别器的扩散网络(Controlled virtual try-on diffusion model,CVDM)。

1.1　基于三维人体姿态信息的多部件服装变形网络

为提升外观流场的局部估计精度,解决关节遮挡处服装变形精度低的问题,本文引入三维人体姿态信息,提出了多部件服装变形方法,构建了基于三维人体姿态信息的多部件服装变形网络,其网络结构示意图如图 1 所示。

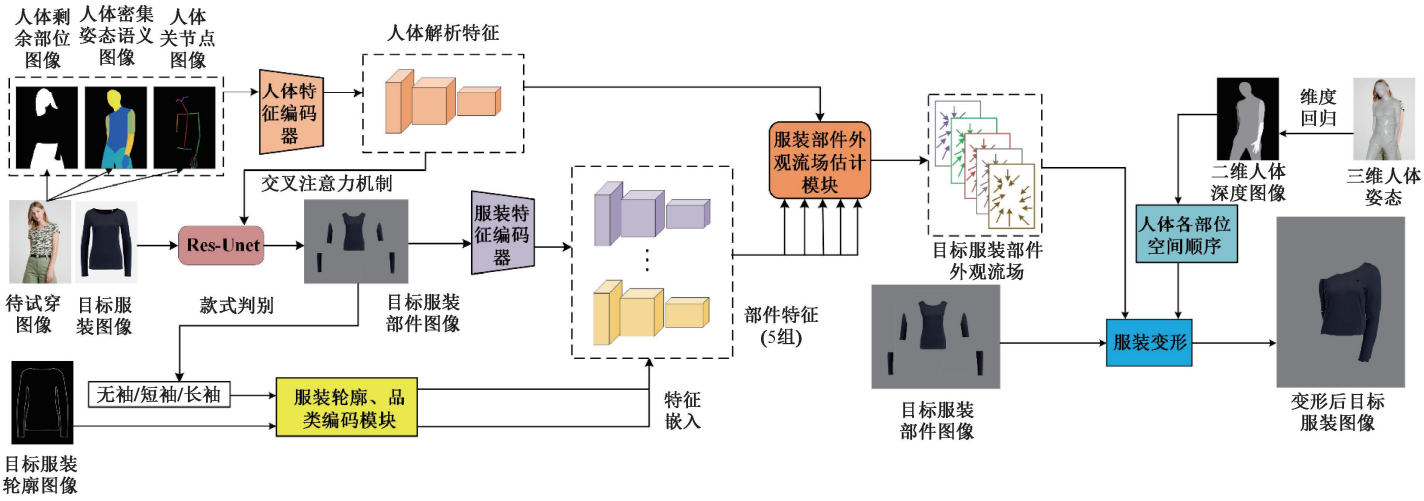


图 1　多部件服装变形网络结构示意图

首先,该网络通过人体特征编码器将一个用于表征二维人体信息的语义解析组(包括人体剩余部位图像、人体密集姿态语义图像和人体关节点图像)编码为不同尺度的人体解析特征,利用 Res-Unet 网络[18]将目标服装分割为 5 个服装部件,包括左上袖、左下袖、服装前片、右上袖、右下袖。该 Res-Unet 网络通过交叉注意力机制[19]在网络编码解码部分嵌入人体解析特征,以实现相应的语义信息匹配,提升服装分割准确度。目标服装图像分割完成后,服装特征编码器将目标服装部件图像编码为 5 组部件特征,服装部件外观流场估计模块则分别捕捉人体解析特征与各部件特征之间的非线性相关性,获得目标服装部件外观流场。

其次,由于目标服装部件图像缺少完整服装信息,多部件服装变形网络对服装轮廓、品类的感知能力较弱,导致目标服装部件外观流场的估计准确度下降,因此本文通过服装轮廓、品类编码模块补充缺失的服装信息,以增强网络的低维感知能力,其模块结构示意图如图 2 所示。该模块根据目标服装部件图像中的服装部件数量判断其款式类别并进行品类特征编码;利用边缘检测算法提取目标服装图像的整体轮廓,等距筛选 63 个轮廓坐标点并对其进行正弦、余弦编码,编码公式为:

$$(P'_x)_i=\sin\left(\frac{(P_x)_i}{W}\right) \tag{1}$$

$$(P'_y)_i=\cos\left(\frac{(P_y)_i}{H}\right) \tag{2}$$

其中:$i\in\{1,2,3,\cdots,63\}$为轮廓点索引,$(P_x)_i$、$(P_y)_i$为第 i 个服装轮廓点的横、纵坐标,$(P'_x)_i$、$(P'_y)_i$为经过正弦、余弦编码后第 i 个轮廓点的横、纵坐标,W、H 为目标服装服装轮廓图像的宽和高。正弦、余弦编码完成后,将轮廓编码特征块与品类编码特征块拼接,通过一个线性层进行特征展开获得包含服装轮廓、品类信息的条件特征块。该条件特征块通过特征嵌入的方式分别加至 5 组服装部件特征中,以提升多部件服装变形网络对服装轮廓、品类信息的感知能力。

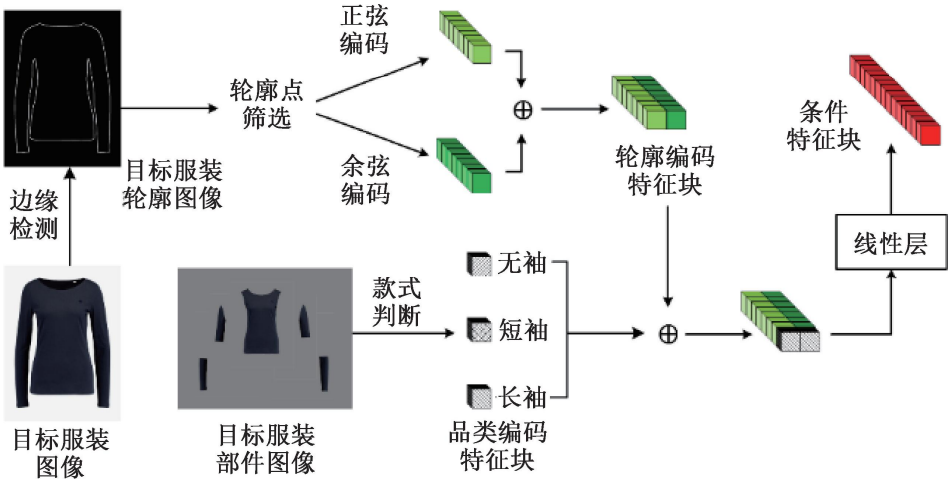


图 2 服装轮廓、品类编码模块结构示意图

然后,本文通过三维人体姿态估计模型获取待试穿图像对应的三维人体姿态信息,提取人体各部位的空间顺序,以此为基础对变形后的各服装部件图像进行融合。利用 Li 等[20]提出的混合神经分析逆运动网络(Hybrid analytical-neural inverse kinematics network,HybrIK)重建单张待试穿图像的三维人体姿态点云模型,通过维度回归的方法,将三维点云坐标映射至二维平面以获得初步的二维人体深度图像,公式可以表示为:

$$C=F(R(\boldsymbol{I})) \tag{3}$$

$$\boldsymbol{I}_{\mathrm{D}}(x,y)=(C_i)_z \tag{4}$$

其中:$\boldsymbol{I}$ 为待试穿图像;R 为单张图像恢复三维人体点云坐标操作;F 为点云过滤操作,点云过滤操作将所有点云回归至二维平面,并过滤重叠的点云,获得单面人体点云信息;C 为过滤后的点云集合;$i\in\{1,2,3,\cdots,6890\}$为点云索引;$(C_i)_z$ 为第 i 个点云的 z 轴坐标值;$\boldsymbol{I}_{\mathrm{D}}(x,y)$为二维人体深度图像中对应像素点的像素值。

为获取人体各部位空间顺序,本文通过二维人体深度图像与人体密集姿态语义图像对人体各部位进行深度赋值。但由于维度回归会导致二维人体深度图像中的人体轮廓与人体密集姿态语义图像中的人体轮廓存在一定差异,因此本文通过服装变形方法的原理,构建人体体型变形网络,降低轮廓差异性,提升深度赋值准确度。该网络利用人体关节点图像、人体密集姿态语义图像对初步获得的二维人体深度图像进行变形,直至其轮廓符合人体密集姿态语义图像。网络训练流程如图 3 所示,其中人体体型变形网络结构与多部件服装变形网络结构相同,该网络的训练目标为 L_1^{D} 最小。L_1^{D} 可用式(5)计算:

$$L_1^{\mathrm{D}}=|\boldsymbol{I}_{\mathrm{Dm}}-\boldsymbol{I}'_{\mathrm{Dm}}| \tag{5}$$

其中:L_1^{D} 为针对变形后二维人体深度图像掩码的 L_1 范数损失,$\boldsymbol{I}_{\mathrm{Dm}}$ 为人体密集姿态语义图像的掩码,$\boldsymbol{I}'_{\mathrm{Dm}}$ 为人体体型变形网络所输出的变形后二维人体深度图像的掩码。

本文根据人体密集姿态语义图像和变形后二维人体深度图像计算各人体部位(左肩、左臂、上肢躯干、右肩、右臂)的像素级平均深度并进行排序。像素级平均深度的计算公式为:

$$\sigma_i=\frac{\boldsymbol{I}'_{\mathrm{D}}\cdot\boldsymbol{I}_i^{\mathrm{Dm}}}{s_i^{\mathrm{Dm}}} \tag{6}$$

其中:$\boldsymbol{I}'_{\mathrm{D}}$ 为变形后二维人体深度图像;$\boldsymbol{I}_i^{\mathrm{Dm}}$ 为人体密集姿态语义图像中第 i 个人体部位的掩码图像;s_i^{Dm} 为该掩码图像的面积;σ_i 为对应人体部位的平均像素深度值;$i\in\{1,2,3,\cdots,5\}$,分别对应人体的左肩、左臂、上肢躯干、右肩、右臂。

获得人体各部位的三维空间顺序后,对已变形的服装部件进行叠加融合,其人体部位与服装部件的对应关系分别为:左肩—左上袖、左臂—左下袖、上肢躯干—服装前片、右肩—右上袖、右臂—右下袖。具体的叠加公式为:

$$w_i^c=\begin{cases}w_i^p, i=1;\\ w_{i-1}^c+(w_{i-1}^c\cap w_i^p), i>1\end{cases}\tag{7}$$

其中:w_i^p 为将服装部件空间顺序由远至近排序后的第 i 个变形后服装部件,w_i^c、w_{i-1}^c 为叠加 i、$i-1$ 个服装部件的部分服装图像。当 i 为 5 时,服装叠加流程结束,获得最终的变形后目标服装图像。

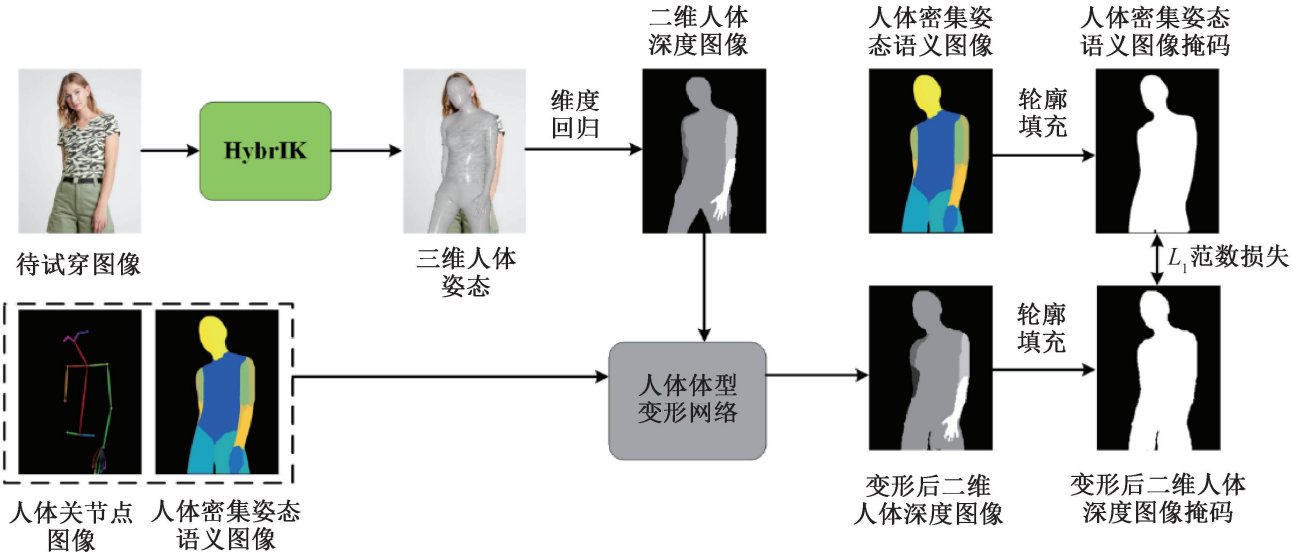


图 3　人体体型变形网络训练流程图

为训练本文所提出的多部件服装变形网络,本文对最终的变形后目标服装图像及其掩码使用 L_1 范数损失 L_1^c、L_1^{cm} 和图像感知损失 L_{VGG}^c,以提升变形前后目标服装纹理细节与形状分布的一致性。L_1^c、L_1^{cm}、L_{VGG}^c 的计算公式为:

$$L_1^c=|w_c\cdot w_{cm}-w'_c\cdot w'_{cm}|\tag{8}$$

$$L_1^{cm}=|w_{cm}-w'_{cm}|\tag{9}$$

$$L_{VGG}^c=\sum|\boldsymbol{\Phi}_i(w_c\cdot w_{cm})-\boldsymbol{\Phi}_i(w'_c\cdot w'_{cm})|\tag{10}$$

其中:w_c、w_{cm} 为真实人体着装图像及其掩码,w'_c、w'_{cm} 为变形后服装图像及其掩码,$\boldsymbol{\Phi}_i$ 为经预训练的 VGG 网络的第 i 层。同时,为提升外观流场的平滑程度,缓解变形后服装出现伪影、模糊和不平滑现象,本文对预估的外观流场应用流场平滑损失 L_S,该损失为广义上的 Charbonnier 损失函数:

$$L_S=\|\nabla f_N\|\tag{11}$$

其中:f_N 为不同尺度下输出的外观流场,$N\in\{1,2,3,4\}$。

1.2　基于人体姿态控制模块和区域度判别器的扩散网络

本文的试衣生成阶段基于扩散网络训练实现,其训练目标函数为:

$$E_{x_0,t,\varepsilon}(\|\varepsilon-\varepsilon_\theta(\sqrt{\bar{\alpha}_t}x_0+\sqrt{1-\bar{\alpha}_t}\varepsilon,t)\|)\tag{12}$$

其中:x_0 为真实的数据分布,时间 $t\in\{1,2,3,\cdots,T\}$ 遵循均匀分布,ε 为随机生成的高斯噪声,$\bar{\alpha}_t$ 为训练方差,$\sqrt{\bar{\alpha}_t}x_0+\sqrt{1-\bar{\alpha}_t}\varepsilon$ 为 t 时刻下对 x_0 添加随机高斯噪声 ε 的加噪图像,ε_θ 为扩散网络。该网络通过预测输入噪声 ε 以实现训练目的。为限制扩散网络在试衣生成阶段的生成自由度,提升其特征保持能力,本文在扩散网络的基础上提出了基于人体姿态控制模块和区域度判别器的扩散网络,其结构图如图 4 所示。

首先,该网络根据二维人体解析信息将待试穿图像处理为只包含头部、下肢、变形后目标服装、部分背景的待填充图像,并与真实加噪图像、待填充掩码共同作为扩散网络的输入。与其他基于扩散网络的二维试衣不同,本文并未选择基于人体轮廓的掩码填充方法[10],而是在其基础上利用矩形区域将人物着装及部分身体信息完全删除,以防止生成过程受填充掩码形状及人体残留着装的影响,其有效性验证见后文实验部分。

其次,由于待填充图像缺少手臂、关节等人体姿态信息,导致扩散网络在生成过程中无法捕捉人体与变形后服装之间的相关性。因此,本文通过所提出的人体姿态控制模块对人体密集姿态语义图像进行下采样编码,获得不同尺度大小的特征块,将其与扩散网络上采样解码阶段输出的各特征块分别相加,以实现特征信息之间的密集匹配,从而保证试衣前后人体形态的一致性。同时,为获取目标服装的全局信息(纹理结构、色彩),本文通过语言图像对比预训练网络 CLIP[21] 对目标服装图像进行条件特征编码,并采用交叉注意力机制将该编码特征加入主干网络,以控制服装纹理与颜色的生成;该 CLIP 网络与扩散网络的编码器部分在训练过程中不更新网络参数,以提升扩散网络的收敛速度与训练稳定性。

以上步骤的最终训练目标函数为：

$$L_2^{\mathrm{DM}}=\mathrm{E}_{\varepsilon(x),y,\varepsilon,t}(\|\varepsilon-\varepsilon_\theta(z_t,t,\tau_\theta(y),\zeta(\boldsymbol{I}_{\mathrm{D}}))\|_2^2) \tag{13}$$

其中：L_2^{DM} 为预测噪声和输入噪声之间的均方误差，$\varepsilon(x)$ 为利用向量量化变分自动编码器将输入图像编码至隐式特征空间中的特征编码操作，τ_θ 为 CLIP 网络，y 为目标服装图像，ζ 为人体姿态控制模块，$\boldsymbol{I}_{\mathrm{D}}$ 为人体密集姿态语义图像。

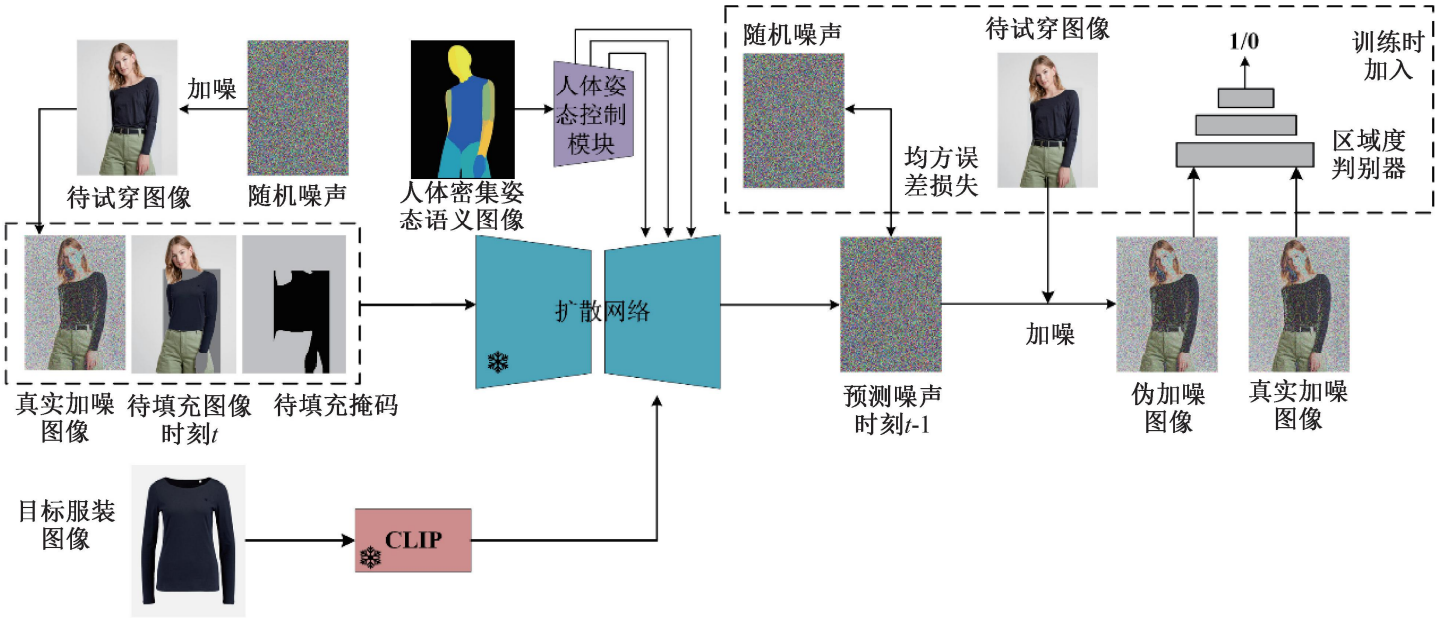


图 4 基于人体姿态控制模块和区域度判别器的扩散网络结构图

最后，为使扩散网络能准确还原图像的局部细节，增强网络对局部特征的捕捉、保持能力，本文引入 PatchGAN[22] 中的区域度判别器与扩散网络进行对抗性训练，从而强化网络学习能力，其训练目标函数为：

$$L_{\mathrm{P}}^{\mathrm{D}}=\mathrm{E}(\log(1-D(\boldsymbol{I}'_{\mathrm{n}})))+\mathrm{E}(\log(D(\boldsymbol{I}_{\mathrm{n}}))) \tag{14}$$

其中：$L_{\mathrm{P}}^{\mathrm{D}}$ 为区域度判别器损失，$\boldsymbol{I}'_{\mathrm{n}}$ 为将预测噪声加噪至待试穿图像的伪加噪图像，$\boldsymbol{I}_{\mathrm{n}}$ 为将输入的随机噪声加噪至待试穿图像的真实加噪图像，D 为区域度判别器。此外，该区域度判别器能与人体姿态控制模块协同工作，保证扩散网络在局部人体区域(手臂皮肤等)生成结果的真实性与完整性。

本文对上述训练损失进行数值加权，获得基于三维人体姿态信息的多部件服装变形网络、基于人体姿态控制模块和区域度判别器的扩散网络的训练总损失 L_{W}、L_{G}：

$$L_{\mathrm{W}}=\lambda_{\mathrm{c}}\cdot L_1^{\mathrm{c}}+\lambda_{\mathrm{cm}}\cdot L_1^{\mathrm{cm}}+\lambda_{\mathrm{VGG}}\cdot L_{\mathrm{VGG}}^{\mathrm{c}}+\lambda_{\mathrm{S}}\cdot L_{\mathrm{S}} \tag{15}$$

$$L_{\mathrm{G}}=\lambda_\varepsilon\cdot L_2^{\mathrm{DM}}+\lambda_{\mathrm{D}}\cdot L_{\mathrm{P}}^{\mathrm{D}} \tag{16}$$

其中：λ_{c}、λ_{cm}、λ_{VGG}、λ_{S}、λ_ε、λ_{D} 为训练超参数，其值分别为 2.0、5.0、1.0、2.0、1.0、0.01。

2 实验结果与分析

2.1 VITON-HD 数据集

本文在公开的 VITON-HD 数据集[23] 上对所提出的网络进行训练与实验。该数据集涵盖了短袖、衬衫、夹克、抹胸、吊带等多种不同的服装品类，待试衣人体姿势覆盖范围广，包括直立、叉腰、举手等多种试穿姿势，并且拥有丰富的人体拍摄视角，因此在基于图像的二维虚拟试衣任务中被广泛使用。VITON-HD 数据集共包括 11647 组分辨率为 1024×768 像素的配对人物训练图像，以及 2032 组相同分辨率的配对测试图像，每组人物图像包含配对的目标服装图像及其掩码、待试穿人体图像和人体解析图像组。

2.2 实验环境与评价指标

本文实验通过深度学习框架 Pytorch1.12.0 进行实现，并将其部署于两块显存为 40 GiB 的 Nvidia A100 显卡上进行训练与测试。此外，本文利用 Adam 优化器对多部件服装变形网络、人体体型变形网络、基于人体姿态控制模块和区域度判别器的扩散网络(以下简称“改进后的扩散网络”)进行训练与优化，优化器初始值分别设置为 0.0001、0.0001、0.00001，其数值随着训练时间的增加而线性衰减，并且各网络的训练时长分别为 50、30、40 轮。在训练改进后的扩散网络时，本文不更新其向量量化变分自动编码器和 CLIP 网络的权重参数，通过网络微调的方式稳定其训练过程，增强网络的学习能力。

为评估所提出网络的实际性能，本文选择弗雷歇距离(Frechet inception distance，FID)[24]、感知图像块

相似度(Learned perceptual image patch similarity, LPIPS)[25]、内核初始距离(Kernel inception distance, KID)[26]、Inception 得分(Inception score, IS)[27]作为试衣结果图像的定量评价指标。弗雷歇距离通过特征空间中生成图像与真实图像之间的特征距离来衡量图像之间的差异,其数值越小,表示两图像之间的相似度越高。感知图像块相似度通过计算图像特征之间的均值方差来衡量两张图像之间的差异,其数值越小,表示两张图片在人类视觉感知上的相似度越高。内核初始距离作为一种无偏估计在较少样本数量的测试中表现要优于弗雷歇距离,更符合人类的视觉感知,其数值越小,表示两图像之间的差异越小。Inception 得分则是一种用于评估生成网络生成图像质量和多样性的图像评价指标,其数值越大,表示图像之间的相似度越大,生成图片的质量越高。

2.3　定量对比分析

为验证本文所提出网络的有效性,分别将 3DM-VTON 网络与目前二维虚拟试衣任务中表现最优异的试衣网络—HR-VTON 网络[10]、GP-VTON 网络[12]、DCI-VTON 网络[17]在 VITON-HD 测试集中进行对比分析,其中:GP-VTON 网络、HR-VTON 网络基于生成对抗网络进行试衣生成,DCI-VTON 网络则在 HR-VTON 网络的服装变形方法基础上引入扩散网络进行试衣生成。各试衣网络在 VITON-HD 测试集上的实验结果如表 1 所示,其中内核初始距离在原基础数值上乘以 100。

表 1　各试衣网络在 VITON-HD 测试集上的评价指标

试衣网络	弗雷歇距离	内核初始距离×10^2	感知图像块相似度	Inception 得分
HR-VTON 网络	9.76	0.42	0.0611	3.21
GP-VTON 网络	8.55	0.20	0.0491	3.46
DCI-VTON 网络	8.40	0.30	0.0475	3.61
3DM-VTON 网络	8.28	0.21	0.0456	3.68

根据表 1 各试衣网络在 VITON-HD 测试集上的评价指标可知,3DM-VTON 网络在弗雷歇距离、感知图像块相似度以及 Inception 得分上的表现均优于其他对照组。其中:3DM-VTON 网络较 HR-VTON 网络、GP-VTON 网络、DCI-VTON 网络在弗雷歇距离得分上的降低,表明该网络能够生成更高质量的试衣结果;感知图像块相似度得分的提升表明该网络生成的试衣结果在低级特征中的表征能力要优于其他网络,并且更加符合人类的视觉感知;而 Inception 得分的提升表明本文所提出网络能够生成更加真实且高质量的试衣结果。

2.4　定性对比分析

不同试衣网络在 VITON-HD 测试集上生成的试衣结果如图 5 所示。当试穿人体面临关节遮挡时,试衣结果示例图像如图 5(a)—(d)所示,从图中可以看出,本文网络能够准确判断人体各关节与目标服装之间的空间结构关系,并保留完整的服装细节特征,生成清晰的试衣结果。当出现非传统的试穿位姿时,试衣结果示例图像如图 5(e)—(h)所示,从图中可以看出,GP-VTON 网络与 DCI-VTON 网络无法在特殊人体部位(弯曲手臂处、背部)还原真实的服装细节,而本文网络在精准还原服装形状分布的同时,还能够生成服装褶皱以模拟对应人体姿态下服装的具体形态。当试穿人体中出现宽松着装时,试衣结果示例图像如图 5(j)所示,从图中可以看出,由于其他网络在生成试衣阶段缺乏人体姿态信息,导致试衣前后试穿人体形态保持不一致,而本文网络则通过人体姿态控制模块引入人体姿态解析信息,增强扩散网络对全局信息的动态捕捉能力,真实还原符合人类感知的人体形态。DCI-VTON 网络的试衣结果示例图像如图 5(b)、图 5(f)和图 5(j)所示,该网络受到初始着装形态的干扰,导致其生成的服装与人体细节特征保真度低,因此本文在基于人体轮廓的填充掩码基础上增加矩形区域,删除原有人体着装与人体其余部位细节,以消除生成试衣结果受填充掩码形态的干扰。

2.5　消融实验

为验证本文所提出网络的有效性,本文分别对多部件服装变形网络以及改进后的扩散网络展开相应的消融实验。

首先,为验证多部件服装变形网络中不同模块对网络整体性能提升的有效性,分别将服装轮廓、品类编码模块和人体各部位空间顺序嵌入模块作为消融实验的控制变量,对不同的多部件服装变形变体网络进行测试与分析。多部件服装变形网络各配置下 3DM-VTON 网络的试衣结果评价指标值如表 2 所示,多部件服装变形网络不同配置下的变形后服装结果图像如图 6 所示。

从表 2 中可以发现:当多部件服装变形网络屏蔽人体各部位空间顺序嵌入模块时(配置 2),该变体网络无法捕捉试穿人体与目标服装之间的空间结构关系,导致弗雷歇距离、内核初始距离、Inception 距离与感知图像块相似度的得分较完整体多部件服装变形网络(配置 3)均有所下降,且下降幅度较大,

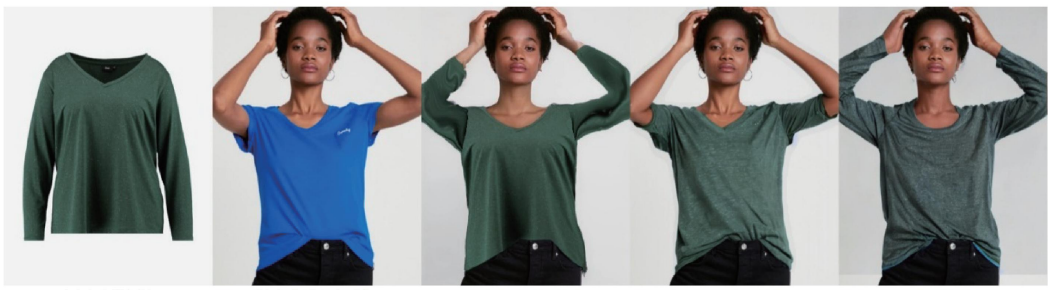
这表明人体各部位空间顺序嵌入模块有助于增强多部件服装变形网络的空间感知能力,提升服装的变形精度。当多部件服装变形网络屏蔽服装轮廓、品类编码模块时,网络整体性能也有所下降,表明该方法能够提升网络对服装品类、轮廓的敏感程度,提升其广泛适用性能。

(a) 服装款式1

(b) 服装款式2

(c) 服装款式3

(d) 服装款式4



(e) 服装款式5

(f) 服装款式6

(g) 服装款式7

(h) 服装款式8

(i) 服装款式9

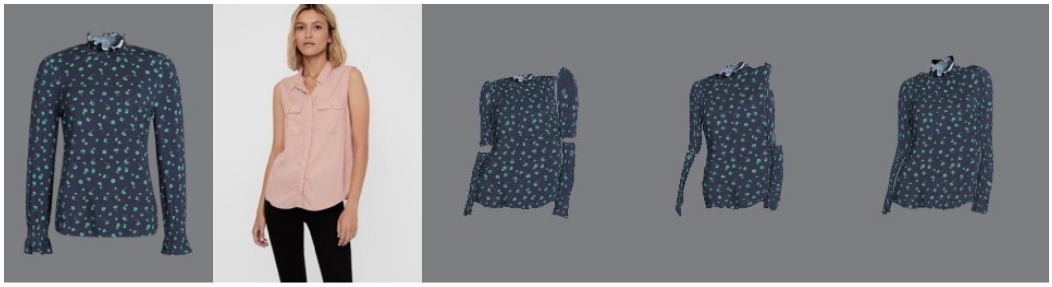
(j) 服装款式10

图 5　不同网络在 VITON-HD 测试集上生成的试衣结果示例图像

表 2　多部件服装变形网络各配置下 3DM-VTON 网络的试衣结果评价指标

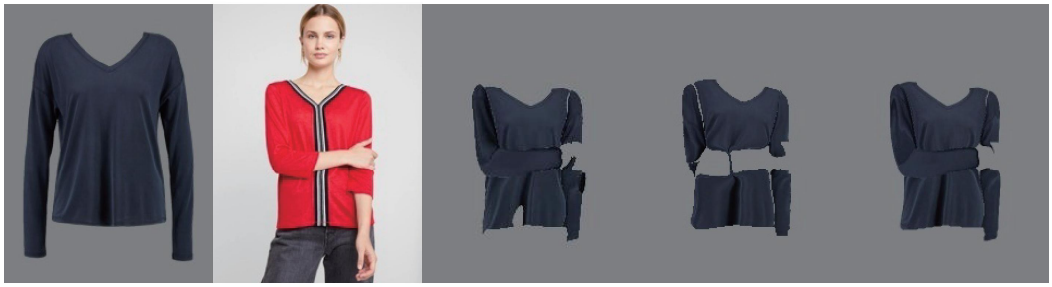
网络配置	服装轮廓、品类编码模块	人体各部位空间顺序嵌入模块	弗雷歇距离	内核初始距离$\times 10^2$	Inception得分	感知图像块相似度
配置 1	×	√	8.32	0.21	3.66	0.0458
配置 2	√	×	8.50	0.23	3.56	0.0463
配置 3	√	√	8.28	0.21	3.68	0.0456

注:×表示屏蔽相应网络模块;√表示加入相应网络模块。



(a) 服装款式11

(b) 服装款式12



(c) 服装款式13

图6 多部件服装变形网络不同配置下的服装变形结果图像

从图6可以得出:当多部件服装变形网络屏蔽人体各部位空间顺序嵌入模块时,该变体网络的空间感知能力下降,无法有效处理非传统试穿位姿与关节遮挡,导致服装整体变形质量低下;当加入该模块时,多部件服装变形网络能够捕捉服装与试穿人体之间的空间关系,实现服装部件与人体关节之间的结构匹配,减少关节遮挡处的服装变形错误现象,从而提升试衣结果的最终质量。与此类似,服装轮廓、品类编码模块也有助于增强多部件服装变形网络对不同数据分布的敏感程度,从而提升服装变形的精确程度。

其次,为验证改进后扩散网络中所引入模块的各自有效性,本文对不同变量设置不同的网络变体,改进后扩散网络不同配置下3DM-VTON网络的试衣结果评价指标如表3所示,改进后扩散网络不同配置下3DM-VTON网络生成的试衣结果图像如图7所示,其中:配置4是用于验证区域度判别器的有效性;配置5是用于验证人体姿态控制模块的有效性。

表3 改进后扩散网络不同配置下3DM-VTON网络的试衣结果评价指标

网络配置	区域度判别器	人体姿态控制模块	弗雷歇距离	内核初始距离$\times 10^2$	Inception得分	感知图像块相似度
配置4	×	√	8.31	0.22	3.61	0.0459
配置5	√	×	8.48	0.24	3.36	0.0523
配置6	√	√	8.28	0.21	3.68	0.0456

注:×表示屏蔽相应网络模块;√表示加入相应网络模块。

从表3和图7中分析得出:当屏蔽人体姿态控制模块时(配置5),各图像评价指标得分均有所下降,且下降幅度较大,试衣前后人体姿势一致性较差,并出现了人体细节丢失、失真现象,整体试衣质量下降严重;当加入该模块时,人体姿态控制模块能够提取用于表征人体姿势形态的

全局特征信息,并与主干扩散网络实现特征信息交互和匹配,约束扩散网络的生成自由度,使得试衣结果中的身体部位细节(手臂皮肤等部位)完整清晰,保证试穿前后人体形态的一致性;当网络训练过程中未加入区域度判别器时,该变体网络在各图像评价指标得分上均有所降低,其局部服装细节的保持能力较差,在试衣结果中出现与目标服装无关的面料褶皱形态,因此验证了区域度判别器能够提升网络的局部特征捕捉能力和生成稳定性。

图 7　改进后扩散网络不同配置下的 3DM-VTON 网络生成的试衣结果图像

最后,本文将 MPWM、CVDM 与其他二维试衣方法进行交叉组合,以进一步验证各自有效性。选择 3 组不同的服装变形方法,分别为 HR-VTON、GP-VTON 网络服装变形方法与本文提出的多部件服装变形方法 MPWM,将其分别与 DCI-VTON 网络中的扩散网络部分(以下简称“DCI-VTON 扩散网络”)与本文提出的改进后扩散网络 CVDM 进行交叉组合测试。其中,DCI-VTON 网络沿用了 HR-VTON 的服装变形方法,并以此为基础进行扩散网络的训练。该消融实验共有 6 组测试组,各交叉测试组的定量比较表如表 4 所示,各交叉测试组生成的试衣结果图像如图 8 所示,图 8 组合分别对应表 4 列出的组合配置。

表 4　各交叉测试组的评价指标

组合	配置详情	弗雷歇距离	内核初始距离$\times 10^2$	Inception 得分	感知图像块相似度
组合 1	DCI+HR	8.40	0.30	3.61	0.0475
组合 2	DCI+GP	8.38	0.22	3.59	0.0473
组合 3	DCI+MPWM	8.32	0.26	3.65	0.0468
组合 4	CVDM+HR	8.35	0.28	3.62	0.0473
组合 5	CVDM+GP	8.36	0.22	3.68	0.0459
组合 6	CVDM+MPWM	8.28	0.21	3.68	0.0456

(a) 服装款式17

(b) 服装款式18

(c) 服装款式19

图8 各交叉测试组生成的试衣结果图像

从表4可以发现:MPWN(本文服装变形方法)与DCI-VTON扩散网络的组合测试表现较与CVDM(本文改进后的扩散网络)的组合测试表现均有所下降,而HR-VTON、GP-VTON与CVDM的组合测试得分较与DCI-VTON扩散网络的组合测试得分有所上升,这表明CVDM网络性能具有一定的优越性。此外,当控制扩散网络变量不变时,本文的服装变形方法MPWN组合较HR-VTON、GP-VTON网络在弗雷歇距离、Inception得分以及感知图像块相似度得分上均有所提升,这表明本文提出的多部件服装变形方法更准确有效。

从图8各交叉测试组生成的试衣结果图像中分析可以发现:组合4—6的试衣结果明显优于组合1—3,当宽松服装对人体形态造成一定的信息干扰时,本文所引入的人体姿态控制模块能够增加扩散网络对全局人体信息的表达能力,生成更符合真实人体形态的试衣结果。

综上所述,通过定性、定量对比分析与消融实验结果可知,本文提出的3DM-VTON网络可以较好地解决关节遮挡以及非传统位姿所导致的服装变形错误现象,保证试衣前后人体形态的一致性,并获得高质的试衣结果。

3 结 语

本文提出了一种基于三维人体姿态信息的协作式扩散试衣生成网络—3DM-VTON网络,该网络通过三维人体姿态估计模型估计三维人体姿态,获得人体各部位空间顺序,对多部件服装变形过程进行约束与指导,提升服装变形网络的空间感知能力,并加入服装轮廓、品类编码模块,增强了网络的数据敏感程度与低维感知能力。此外,本文在扩散网络的基础上通过加入人体姿态控制模块与区域度判别器,提升了扩散网络的全局、局部信息捕捉能力,限制网络生成自由度,并获得高质量的虚拟试衣结果。本文在公开数据集VITON-HD中进行网络训练与测试,结果表明本文所提出网络在弗雷歇距离、感知图像块相似度、Inception距离的得分均优于目前常用的二维虚拟试衣网络,能够缓解关节遮挡处服装变形错误以及试衣前后人体形态保持不一致等问题,生成清晰、自然的试衣结果,满足虚拟试衣用户的多种试穿需求,有助于提升消费者在线购买服装满意度,推动服装电商行业发展。

本文提出的试衣网络仍存在一些不足,例如:在服装变形阶段中,对于带有复杂纹理的服装,变形后存在纹理畸变、失真问题;在试衣生成阶段中,服装纹理无法保持完整。针对这些问题,须进一步提升外观流场的估计精度,限制试衣生成网络的生成自由度,以满足用户对带有复杂纹理服装的虚拟试穿需求。

参考文献:

[1] 赵娟, 魏雪霞, 徐增波. 基于深度学习的2D虚拟试衣技术研究进展[J]. 丝绸, 2021, 58(9): 48-52.

[2] 谭泽霖，白静. 二维图像虚拟试衣技术综述[J]. 计算机工程与应用，2023，59(15)：17-26.
[3] 施倩，罗戎蕾. 基于生成对抗网络的服装图像生成研究进展[J]. 现代纺织技术，2023，31(2)：36-46.
[4] 薛萧昱，何佳臻，王敏. 三维虚拟试衣技术在服装设计与性能评价中的应用进展[J]. 现代纺织技术，2023，31(2)：12-22.
[5] 崔萌，陈素英，殷文，等. 基于虚拟试衣技术的服装设计与开发[J]. 毛纺科技，2020，48(6)：58-61.
[6] 郭宇轩，孙林. 基于扩散模型的 ControlNet 网络虚拟试衣研究[J]. 现代纺织技术，2024，32(3)：118-128.
[7] 花爱玲，余锋，陈子宜，等. 深度学习在二维虚拟试衣技术的应用与进展[J]. 计算机工程与应用，2023，59(11)：37-45.
[8] Goodfellow I J，Pouget-Abadie J，Mirza M，et al. Generative adversarial nets[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal，Canada. ACM，2014：2672-2680.
[9] Ho J，Jain A，Abbeel P. Denoising diffusion probabilistic models [C] // Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver，BC，Canada. ACM，2020：6840-6851.
[10] Lee S Y，Gu G，Park S，et al. High-resolution virtual try-on with misalignment and occlusion-handled conditions[C]// Avidan S，Brostow G，Cissé M，et al. European Conference on Computer Vision. Tel-Aviv，Israel. Cham：Springer，2022：204-219.
[11] Yang Z J，Chen J Y，Shi Y K，et al. OccluMix：Towards de-occlusion virtual try-on by semantically-guided mixup[J]. IEEE Transactions on Multimedia，2023，25：1477-1488.
[12] Xie Z Y，Huang Z Y，Dong X，et al. GP-VTON：Towards general purpose virtual try-on via collaborative local-flow global-parsing learning[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver，BC，Canada. IEEE，2023：23550-23559.
[13] 张颖，刘成霞. 生成对抗网络在虚拟试衣中的应用研究进展[J]. 丝绸，2021，58(12)：63-72.
[14] 祖雅妮，张毅. 基于大规模预训练文本图像模型的虚拟试穿方法[J]. 丝绸，2023，60(8)：99-106.
[15] Zhu L Y，Yang D W，Zhu T，et al. TryOnDiffusion：a tale of two UNets[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver，BC，Canada. IEEE，2023：4606-4615.
[16] Morelli D，Baldrati A，Cartella G，et al. LaDI-VTON：latent diffusion textual-inversion enhanced virtual try-on[C]// Proceedings of the 31st ACM International Conference on Multimedia. Ottawa ON，Canada. ACM，2023：8580-8589.
[17] Gou J H，Sun S Y，Zhang J F，et al. Taming the power of diffusion models for high-quality virtual try-on with appearance flow[C]// Proceedings of the 31st ACM International Conference on Multimedia. Ottawa ON，Canada. ACM，2023：7599-7607.
[18] Ronneberger O，Fischer P，Brox T. U-net：convolutional networks for biomedical image segmentation [C] // International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich，Germany. Cham：Springer，2015：234-241.
[19] Huang Z L，Wang X G，Huang L C，et al. CCNet：criss-cross attention for semantic segmentation[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul，Korea (South). IEEE，2019：603-612.
[20] Li J F，Xu C，Chen Z C，et al. HybrIK：a hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation [C] // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville，TN，USA. IEEE，2021：3382-3392.
[21] Radford A，Kim J W，Hallacy C，et al. Learning transferable visual models from natural language supervision [C]// Proceedings of the 38th International Conference on Machine Learning. New York：PMLR，2021：8748-8763.
[22] Isola P，Zhu J Y，Zhou T H，et al. Image-to-image translation with conditional adversarial networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu，HI，USA. IEEE，2017：5967-5976.
[23] Choi S，Park S，Lee M，et al. VITON-HD：high-resolution virtual try-on via misalignment-aware normalization[C] // 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville，TN，USA. IEEE，2021：14126-14135.
[24] 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. Long Beach，California，USA. ACM，2017：6629-6640.
[25] Zhang R，Isola P，Efros A A，et al. The unreasonable effectiveness of deep features as a perceptual metric[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City，UT，USA. IEEE，

2018: 586-595.

[26] Bińkowski M, Sutherland D J, Arbel M, et al. Demystifying MMD GANs[EB/OL]. (2018-01-04)[2024-05-29]. https://arxiv.org/abs/1801.01401.

[27] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training GANs[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain. ACM, 2016: 2234-2242.

(责任编辑:康 锋)