

文章编号:1007-2780(2025)09-1381-09

基于 PATN 和自注意力机制的多分辨率 人体姿态迁移方法

黄庆东^{1,2*}, 苏宇辉¹, 刘依华¹, 陈梓煌¹, 姚咏琪¹

(1. 西安邮电大学 通信与信息工程学院, 陕西 西安 710121;

2. 陕西省信息通信网络及安全重点实验室, 陕西 西安 710121)

摘要:针对现有人体姿态迁移方法因编码阶段特征处理不当导致图像变形失真的问题,提出基于 Pose-Attentional Transfer Network(PATN)和自注意力机制的多分辨率人体姿态迁移方法。首先,设计了姿态引导自注意力模块,通过多头注意力机制增强关键身体区域特征通道的权重,并减小背景无关特征的影响,自适应地探索两条支路特征之间的关联性;其次,在解码阶段加入多尺度注意力模块,增强不同尺度的姿态信息表达,有效提升局部细节和整体纹理的保真度;最后,引入三元像素损失对生成图像进行约束,提高了图像的特征一致性和结构一致性。在 DeepFashion 和 Market-1501 数据集上进行验证,实验结果表明,本文方法在结构相似性(SSIM)、初始评分(IS)、感知相似度(LPIPS)指标上均优于现有的 PATN 方法,并且在视觉感观、边缘纹理方面均有所提升,在行人重识别的下游任务中具有重要的潜力。

关键词:图像处理;姿态迁移;CGAN;自注意力机制;多分辨率

中图分类号:TP391 **文献标识码:**A **doi:**10.37188/CJLCD.2025-0100 **CSTR:**32172.14.CJLCD.2025-0100

Multi-resolution human posture transfer algorithm based on PATN and self-attention mechanism

HUANG Qingdong^{1,2*}, SU Yuhui¹, LIU Yihua¹, CHEN Zihuang¹, YAO Yongqi¹

(1. School of Communications and Information Engineering, Xi'an University of Posts and
Telecommunications, Xi'an 710121, China;

2. Shaanxi Key Laboratory of Information Communication Network and Security, Xi'an 710121, China)

Abstract: Aiming at the problem that the existing human pose transfer methods suffer from image deformation and distortion due to improper feature processing in the encoding stage, a multi-resolution human pose transfer method based on pose-attentional transfer network (PATN) and self-attention mechanism is proposed. Firstly, a pose-guided self-attention module is designed to enhance the weight of the feature channel of the key body region through the multi-head attention mechanism, reduce the influence of background-irrelevant features, and adaptively explore the correlation between the two branch features. Secondly, a multi-scale attention module is added in the decoding stage to enhance the expression of pose information

收稿日期:2025-05-15;修订日期:2025-07-06.

基金项目:国家自然科学基金(No.62201456)

Supported by National Natural Science Foundation of China (No.62201456)

*通信联系人, E-mail:huangqingdong@xupt.edu.cn

at different scales, effectively improve the fidelity of local details and overall texture. Finally, the ternary pixel loss is introduced to constrain the generated image, which improves the feature consistency and structural consistency of the image. The experimental results are verified on the DeepFashion and Market-1501 datasets. They show that the proposed method is superior to the existing PATN method in terms of structural similarity (SSIM), initial score (IS) and perceptual similarity (LPIPS), and has improved visual perception and edge texture, showing important potential in the downstream task of person re-identification.

Key words: image processing; pose transfer; CGAN; self-attention mechanism; multi-resolution

1 引言

随着深度生成模型的不断演进,人体姿态迁移(Human Posture Transfer, HPT)技术近年来在智能照片编辑、行人重识别等领域得到广泛应用,成为当前计算机视觉领域的研究热点。

主流的 HPT 方法主要分为基于变分自编码器(Variational Autoencoder, VAE)^[1]和生成对抗网络(Generative Adversarial Network, GAN)^[2]的两种深度生成模型。VAE通过捕捉数据不同维度间的关系生成新数据,而GAN通过生成器和判别器的对抗博弈生成图像。由于GAN的对抗损失优于VAE的逐点损失,能生成更清晰的图像,因此得到广泛应用。

为了增强生成图像的可控性,Mirza等人^[3]设计了条件生成对抗网络(Conditional Generative Adversarial Network, CGAN)。在人物图像生成中,姿态图通常被用作约束条件。Ma等人^[4]提出了姿态引导的人体图像生成模型PG²,该模型通过两阶段策略,首先生成模糊图像,再优化纹理和色彩,虽然提升了质量,但仍未能完全捕捉图像分布变化,导致纹理细节缺失。Essner等人^[5]提出了将人体图像外观和姿态信息解耦的方法,但基于U-Net的跳跃连接机制导致了姿态差异带来的特征偏移问题。Siarohin等人^[6]引入了可变形跳跃连接机制,将姿势迁移变形分解为多个仿射变换,从而有效缓解了特征偏移。Zhu等人^[7]设计了一种渐进式姿态注意力转换网络(Pose Attentional Transfer Network, PATN),通过级联9个姿态注意力单元逐步实现外观和姿态特征转换。尽管现有方法取得了显著进展,但仍面临遮挡导致关键点丢失与衣物纹理、装饰细节

难以完整迁移,以及处理姿态和外观两种模态信息出现肢体扭曲或风格不一致问题,需进一步优化。

本文提出了一种融合姿态引导自注意力机制的HPT算法,利用PATN^[7]生成器架构获得深层特征并增强特征关联。通过(Pose Guide Attention Mechanism, PGAM)模块,利用多头注意力机制增强关键身体区域特征,减小背景干扰,自适应探索特征间关联,提升图像风格一致性。在编码器和解码器间加入多尺度注意力模块(Multi-Scale Attention Module, MSAM)^[8],增强姿态信息表达并保留纹理细节。此外采用PatchGAN^[9]判别器调整感受野,注重高频信息和纹理构建,并引入三元像素损失,使用L1损失、对抗损失和感知损失加权结合来训练模型,灵活平衡图像质量与结构一致。在DeepFashion和Market-1501公开数据集上验证了所提方法的有效性。

2 网络结构

PATN算法是一种基于CGAN的编码器-解码器框架,整体框架如图1所示。编码器分为图像编码 E_i 和姿态编码 E_p 两个分支,每个分支由4个 3×3 卷积层的下采样模块组成。通过这两个分支分别获取初始的图像 F_0^p 和姿态特征 F_0^s ,并将其输入到PATN网络中,经过9个级联的姿态迁移模块。假设在第 t 个模块,姿态特征 F_{t-1}^s 经过两个卷积层(包含归一化层和ReLU激活函数)后,通过逐元素激活函数映射到 $[0, 1]$ 区间:

$$M_t = \sigma(\text{Conv}_s(F_{t-1}^s)), \quad (1)$$

式中: M_t 为注意力掩码图; σ 为归一化激活函数;

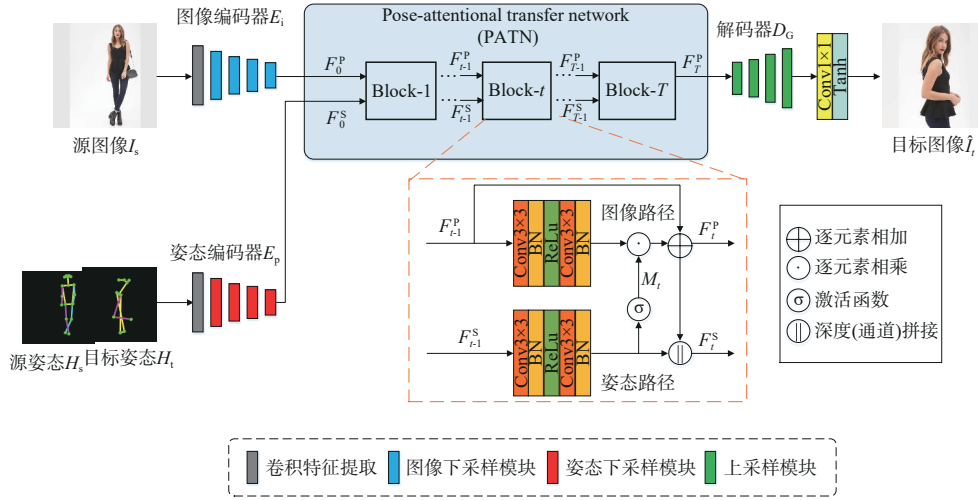


图1 PATN的生成框架

Fig. 1 PATN generation framework

Conv_s 为姿态卷积操作,并与图像特征和注意力掩码相乘,生成更新后的图像特征 F_t^p :

$$F_t^p = M_t \odot \text{Conv}_p(F_{t-1}^p) + F_{t-1}^p, \quad (2)$$

其中: $\odot$ 为逐元素相乘, Conv_p 为图像卷积操作。图像特征与姿态特征经过拼接和 Conv_s :

$$F_t^s = \text{Conv}_s(F_{t-1}^s) \parallel F_t^p, \quad (3)$$

式中, $\parallel$ 表示通道拼接。最终,得到更新后的图像 F_T^p 。在解码阶段,通过4个上采样模块恢复图像尺度后,使用卷积和激活函数生成图像 $\hat{I}_t$ 。

3 改进的PATN算法

为解决PATN算法中图像信息和姿态信息的模态异构问题,本文在其基础上提出结合PGAM和MSAM的HPT算法,算法结构如图2所示。PGAM模块通过自注意力机制,引导姿态特征对图像特征进行调制,实现跨模态特征融合。MSAM通过多尺度卷积促进特征交互,有效减少因下采样提取深层特征而导致的空间几何信息丢失。

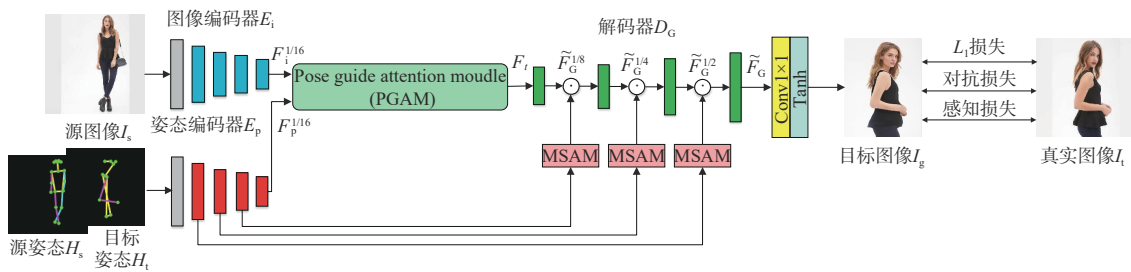


图2 生成网络

Fig. 2 Generative network

判别器采用基于CNN的PatchGAN结构,以L1损失、对抗损失和感知损失联合优化,通过判断图像中各局部图像块(Patch)的真伪实现区域级监督。假设各图像块间像素独立,该方法能在降低计算成本、提升训练效率的同时,有效捕捉纹理细节。本实验中图像块大小设置为70,参考了pix2pix等经典方法的经验,该设置已在图像

生成任务中得到广泛验证^[9]。

3.1 PGAM模块

传统卷积生成网络依赖局部特征合成图像,受限于固定卷积核,难以建模全局信息,且堆叠多层卷积会降低效率。SAGAN^[10]较好地解决了这一问题,它将不同空间位置的相关信息融合到生成图像中,使模型能够学习到全局空间信息。

具体而言, SAGAN 通过将一个空间点处的响应计算为所有空间点特征的加权和, 以最小的计算成本来获得权重。

针对源姿态与目标姿态差异较大时的空间变形问题, 本文设计 PGAM 模块以增强姿态与图像特征的关联性, 缓解边界模糊现象。PGAM 的结构如图 3 所示。

为有效捕捉图像特征的全局关联性, 将 $F_i^{1/16}$ 与 $F_p^{1/16}$ 拼接后通过自注意力块进行特征变换, 生成融合特征 F_f 。此模块通过注意力权重自适应地选择显著性的区域, 以整合全局特征图信息和姿态信息。随后, 分别输入两个由 3×3 卷积、实例归一化和 ReLU 组成的卷积块, 压缩通道数至原来的 $1/2$, 以降低相似度计算的运算量。

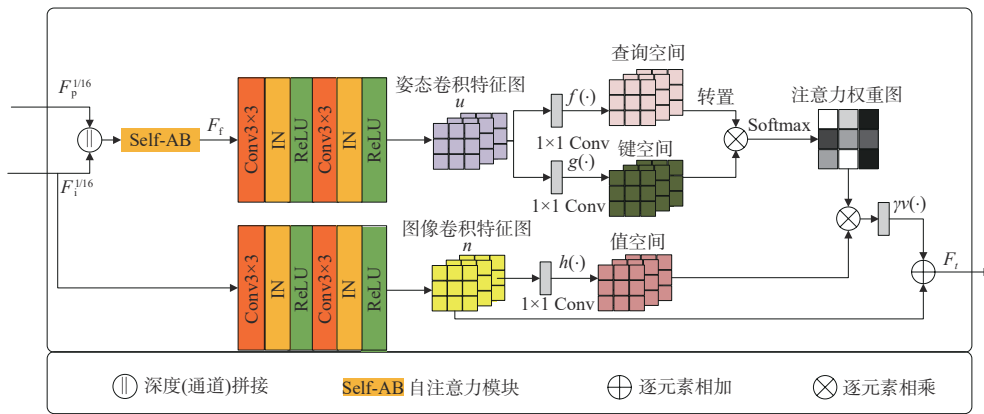


图 3 PGAM 模块结构

Fig. 3 PGAM module structure

对 u 进行卷积得到查询空间 $f(u)$ 和键空间 $g(u)$, 对 d 进行卷积得到值空间 $h(d)$:

$$f(u) = W_f u, g(u) = W_g u, h(d) = W_h d, \quad (4)$$

式中, W_f 、 W_g 和 W_h 为可学习的权重矩阵。

引入查询-键空间的相似性建模, 使网络能够根据姿态图中每个关节位置有选择地关注图像中对应区域的特征表达, u 可以将焦点聚集于关节区域, 有效地对背景区域的信息进行抑制。

注意力机制通过计算查询与键的相似度, 并应用 Softmax 得到归一化的注意力权重图。相似度矩阵 S 定义为:

$$S = \text{Softmax} \left[f(u)^T g(u) \right] = \text{Softmax} \left[(W_f u)^T W_g u \right], \quad (5)$$

$$o_n = v(S h(d)), v(x) = W_v x, \quad (6)$$

其中: o_n 表示第 n 个注意力头的输出, $v(\cdot)$ 是第 n 个注意力头的输出经过卷积得到的特征空间; W_v 表示可学习的参数矩阵。

为了增强特征表达能力并激励网络从不同子空间提取互补信息, PGAM 通过多头注意力机制联合关注不同子空间的信息表示, 并聚合各头的注意力输出, 得到完整的注意力层输出:

$$o = \text{Concat}(o_1, o_2, \dots, o_n, \dots, o_N), \quad (7)$$

式中: $\text{Concat}(\cdot)$ 表示拼接操作, N 为子空间数量。

为优化特征融合, 引入可训练参数 γ , 将注意力层输出与输入图像编码加权组合, 形成残差结构, γ 初始化为 0。因此, 最终输出为:

$$F_f = \gamma o + \text{Conv}(F_i^{1/16}). \quad (8)$$

3.2 MSAM 模块

在卷积神经网络中, 浅层特征包含丰富的空间几何信息, 深层特征则具有更强的语义表达能力。为充分融合多尺度信息, 引入 MSAM 模块应用于编码器和解码器中, 可以增强网络的信息传输能力, 自适应地融合模型中不同层次的特征信息, 同时避免因增加深度特征图的维度而导致计算复杂度过高。MSAM 结构如图 4 所示。

为了根据不同分辨率自适应地选择特征, 多尺度注意力模块引入了 3 种具有不同感受野的卷积方法来提取特征, 分别为逐点卷积(Pointwise Convolution)、普通卷积(Ordinary Convolution)和扩张卷积(Dilated Convolution)。每种卷积操作后都接一个批量归一化层, 以稳定训练过程。

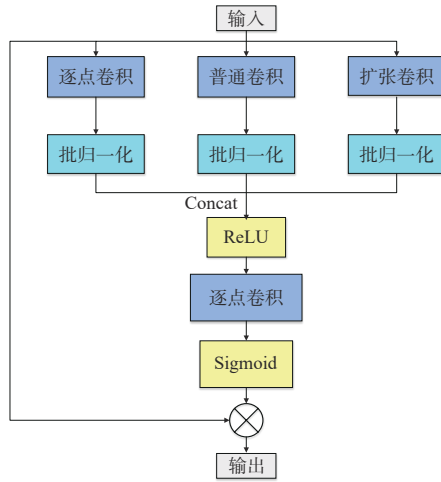


图4 MSAM模块结构

Fig. 4 MSAM module structure

逐点卷积作用于每个像素位置。通过对通道维度的信息进行点乘操作,有助于捕捉输入特征的全局上下文信息。普通卷积使用核大小为 3×3 的卷积核,步长为1,填充为1,主要用于提取浅层局部特征,建模细粒度的纹理结构。扩张卷积则在保持卷积核大小为 3×3 的基础上,通过设置步长为2、填充为2以及膨胀率为2,在卷积核中引入空洞,从而扩大感受野,捕捉更广泛的深层特征上下文信息。这3种卷积操作分别生成尺寸一致的特征图,并在经过ReLU激活之前将这些特征图连接起来。接着,模块通过另一个逐点卷积层和Sigmoid激活函数生成权重图,用于衡量不同位置特征的重要性。最终,通过将该权重图应用于原始输入特征,增强特征表示。整个过程公式定义为:

$$F_{\text{concat}} = \text{ReLU} \left(\text{Concat} \left\{ \text{BN} \{ \text{Pointwise Conv} (F_p) \}, \right. \right. \\ \left. \left. \text{BN} \{ \text{Ordinarywise Conv} (F_p) \}, \text{BN} \{ \text{Dilationwise} \right. \right. \\ \left. \left. \text{Conv} (F_p) \} \right\} \right), \quad (9)$$

$$F_s = F_p \times \text{Sigmoid} \{ \text{Pointwise Conv} (F_p) \}, \quad (10)$$

式中: F_p 表示输入的特征图, F_{concat} 表示拼接3种卷积结构, F_s 表示经过多尺度注意力模块的输出特征图。

3.3 三元像素损失

三元损失函数用于引导模型进行像素重构,包含对抗损失、重构损失和感知损失。为评估生成图像的高频正确性,如边缘结构,使用PatchGAN的二进制交叉熵损失函数惩罚图像分布差异。

条件对抗网络的目标函数为:

$$L_{\text{CGAN}}(G, D) = E_{(x, y)} [\log(D(x, y))] + \\ E_x [\log(1 - D(x, G(x)))], \quad (11)$$

式中: G 为生成器, D 为判别器, $E(\cdot)$ 为期望, x 为源图像, y 为真实图像, $G(x)$ 为生成的目标图像, $D(x, y)$ 表示输入 x 与 y 判断为真的概率, $D(x, G(x))$ 表示输入 x 与 $G(x)$ 判断为真的概率。

为确保生成图像的低频正确性并防止模糊,引入 L_1 范数约束,公式如式(12)所示:

$$L_{l_1}(G) = \|G(x) - y\|_1. \quad (12)$$

感知损失通过预训练的VGG19模型提取特征,监督像素间微小差异,提高生成图像的视觉保真度,保留整体视觉内容:

$$L_{\text{per}}^{\phi, \rho}(G) = \frac{1}{C_{\rho} H_{\rho} W_{\rho}} \|\phi_{\rho}(G(x)) - \phi_{\rho}(y)\|_2^2, \quad (13)$$

式中: ϕ 表示预训练的VGG19模型, ρ 表示层数, C_{ρ} 、 H_{ρ} 、 W_{ρ} 表示该层特征图的通道数、高度和宽度乘积, $\phi_{\rho}(y)$ 和 $\phi_{\rho}(G(x))$ 分别表示真实图像和目标图像在该层提取的特征, $\|\cdot\|_2$ 为2范数。本文提出的最终优化目标函数为:

$$L_{\text{total}} = \lambda_1 L_{l_1}(G) + \lambda_2 L_{\text{CGAN}}(G, D) + \lambda_3 L_{\text{per}}, \quad (14)$$

其中, λ_1 、 λ_2 、 λ_3 分别表示预设的各项加权系数。

4 实验结果与分析

实验在24 GB显存的NVIDIA GeForce RTX 4090 GPU服务器上基于Pytorch框架实现。初始化 $\eta = 0.001$, $\beta_1 = 0.5$, $\beta_2 = 0.999$ 。系数 λ_1 、 λ_2 、 λ_3 设置为(5, 1, 5),批大小(Batch Size)设置为16。使用Adam优化器进行280k次迭代训练,学习率在250k次后线性衰减至0。

4.1 数据集

本研究在DeepFashion和Market-1501数据集上进行性能评估。DeepFashion数据集包含 176×256 尺寸的人像图像,调整为 256×256 ;Market-1501数据集包含32 668张不同光照和视角的人物图像,且具有复杂的背景信息。通过预训练的姿势估计器^[11]生成18个关键点的姿态热力图。为了公平比较,采用Zhu等人^[7]提供的精确训练测试划分,在DeepFashion随机选取

101 966 对图像用于训练,8 570 对用于测试;在 Market-1501 中使用 206 628 个训练对和 13 000 个测试对。

4.2 定性与定量分析

本文参考以往的姿态迁移算法,采用 SSIM、IS、DS、PCKh 和 LPIPS 作为评估指标。SSIM 衡量结构相似性,IS 评价图像质量与多样性,DS 基于检测置信度反映感知质量,PCKh 评估关键点定位精度,LPIPS 则反映深层特征差异。

将本文模型与几个先进的方法进行定性比较,包括 PG^[4]、PISE^[12]、PATN^[7]、DPTN^[13]、PIDM^[14]、DPRM^[15]和文献[16]方法。为确保公正

性,所有方法均使用与文献[7]相同的图像(I_s, I_t)对作为输入,定性结果如图 5 所示。PISE、PG²容易产生模糊扭曲的结果;DPTN 借助辅助人物改进了 HPT 模型,但模型上下文表示能力限制了图像的生成质量;本文通过 PGAM 和 MASM 结合的方式,能更好地捕捉上下文特征,生成的图像在清晰度、色彩和细节表现上优于其他方法。

表 1 展示了在 DeepFashion 和 Market-1501 数据集上 7 种最新姿势迁移方法与本文方法的定量比较,其中最佳结果用粗体标出,次佳结果用下划线标出。从表 1 数据可以看出,对于 DeepFashion 数据集,本文方法在多项关键性能指标

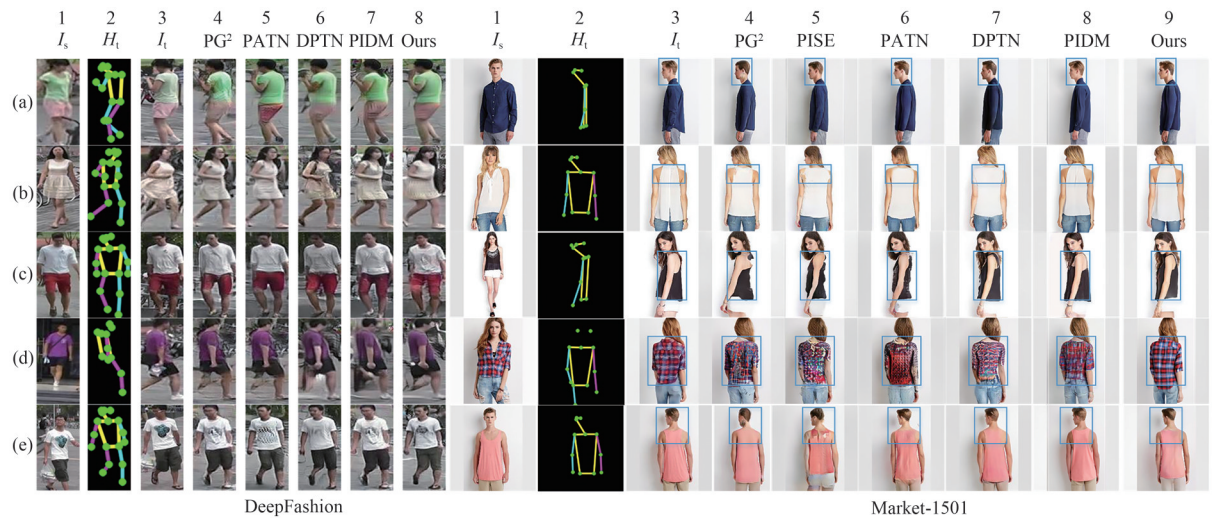


图 5 不同模型在 DeepFashion 和 Market-1501 数据集上的定性结果,(a~e)分别表示不同的测试样例,每一行展示同一样例在各模型下的生成结果。

Fig. 5 Qualitative results of different models on DeepFashion and Market-1501 datasets, (a~e) denote different testing cases, each row shows the results of the same case across models.

表 1 在 DeepFashion 和 Market-1501 数据集上的定量比较

Tab. 1 Quantitative comparison on DeepFashion and Market-1501 datasets

Methods	DeepFashion					Market-1501			
	SSIM ↑	IS ↑	DS ↑	PCKh ↑	LPIPS ↓	SSIM ↑	IS ↑	DS ↑	LPIPS ↓
PG ²	0.756	3.163	0.960	0.89	0.523	0.261	<u>3.495</u>	0.39	0.292
PISE	0.768	<u>3.362</u>	0.967	0.94	0.208	—	—	—	—
PATN	0.773	3.209	0.976	<u>0.96</u>	0.299	0.311	3.323	0.74	0.254
DPTN	<u>0.778</u>	3.351	0.970	0.94	0.196	<u>0.314</u>	3.095	<u>0.68</u>	0.271
PIDM	0.731	—	—	—	<u>0.167</u>	0.305	—	—	<u>0.241</u>
DRDM	0.740	—	—	—	0.165	—	—	—	—
文献[16]	0.713	—	—	—	0.210	—	—	—	—
Ours	0.784	3.692	<u>0.971</u>	0.98	0.191	0.318	3.778	0.66	0.224
Real-Data	1.000	3.796	0.980	1.000	0.000	1.000	3.890	0.74	0.000

均显著超越PATN模型。具体来说,本文方法在SSIM(0.784)、IS(3.692)和PCKh(0.98)上相比次佳结果分别提高了0.011、0.483和0.01,证明本文生成图像的结构相似性、图像质量和多样性、人体关键点准确率以及感知相似度都有显著提升。对于Market-1501数据集,本文方法的IS指标达到3.778,提升尤为明显,显示其生成的图像更符合视觉感知,目标图像的相似性也更高。然而,尽管本文方法已取得较好的表现,真实数据(Real Data)作为对比标准,仍展现出最佳的各项指标,表明生成方法在某些方面还需进一步优化。

4.3 消融实验

为探索模块设计对模型性能的影响,进行了4种模型的消融实验:模型A:不添加新模块的基准模型。模型B:w/o PGAM,从模型中删除PGAM模块来探索姿态特征对图像特征的影响。模型C:w/o MSAM,删除MSAM模块以评估其重要性。模型D:完整模型,即包括PGAM与MSAM。

消融实验的定性结果如图6所示。在去掉PATN的基准模型中,缺乏姿态引导,人物姿态发生了明显的严重形变,衣服纹理缺失,生成的图像在细节还原方面表现不佳,说明在无姿态对齐机制的情况下,网络难以准确捕捉和迁移人物的姿态与外观特征。在添加PGAM后,人物姿态得到改善,但面部特征和衣服细节方面仍然存在缺失,且与原图相差较大。这是由于人物纹理较为复杂,在网络缺乏相应约束的情况下,生成图像中的人物纹理难以与目标人物保持一致。

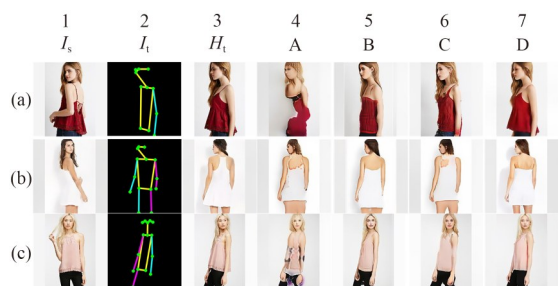


图6 消融实验的定性比较,(a~c)分别表示不同的测试样例,每一行展示同一样例在各消融实验变体下的生成结果。

Fig.6 Qualitative comparison of ablation experiments, (a~c) denote different testing cases, each row shows the results of the same case across different ablation variants.

加入MSAM后,生成的衣物纹理更清晰。该实验结果表明,模型中加入PGAM和MSAM具有显著的效果,从全局对齐和局部细化两个维度提升了生成质量。

消融实验的定量分析结果如表2所示。从表2可得,两个方面的改进对网络在各个指标上的表现均有不同程度的提升。当网络缺少PGAM并且未添加MSAM时,定量结果与最终结果相差较大,说明这两部分的改进对网络性能有显著提升。PGAM通过姿态引导的注意力掩码图完成图像特征的转换,而MSAM则通过充分提取多尺度信息,直接提升人物衣服纹理细节效果。

表2 消融实验在DeepFashion数据集下定量结果

Tab.2 Quantitative results of ablation experiments on the DeepFashion dataset

Model	SSIM ↑	IS ↑	DS ↑	PCKh ↑	LPIPS ↓
A	0.729	3.261	0.897	0.76	0.268
B	<u>0.770</u>	<u>3.459</u>	<u>0.965</u>	<u>0.95</u>	0.284
C	0.755	2.859	0.953	0.90	<u>0.243</u>
D	0.784	3.692	0.971	0.98	0.191
Real Data	1.000	3.796	0.974	1.00	0.000

注:粗体为最佳结果,下划线为次佳结果。

4.4 PATN与本文算法的特征图可视化比较

图7对比了PATN与本文方法在姿态迁移任务中的特征图,展示了二者在特征提取与图像生成方面的差异,其中Inconv对应图1的卷积特征提取部分,Dc1~Dc4表示姿态编码器下采样

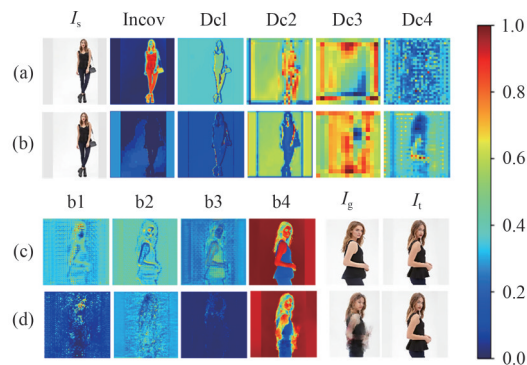


图7 PATN(a,c)与本文方法(b,d)的特征可视化

Fig.7 Feature visualization of PATN (a, c) and the proposed method (b, d)

过程, b1~b4 表示解码器的上采样过程。图 7(c) 和图 7(d) 中的 b1 特征图显示, PATN 存在特征分布不均、边界模糊的问题, 难以准确定位人体结构。相比之下, 本文引入的 PGAM 模块通过引导姿态信息建模, 有效增强了关键区域响应, 使注意力更集中, 结构表达更清晰, 显著提升了特征层的建模能力。同时, 图 7(c) 中的 PATN 特征图 b1~b4 在不同尺度下响应强度不一致, 层间缺乏有效协调, 造成纹理信息损失和细节模糊。而本文在解码阶段引入 MSAM 模块, 通过多种卷积提取不同感受野下的语义特征, 并生成注意力图实现多尺度调节。图 7(d) 中特征图 b1~b4 显示, 在加入 MSAM 后, 特征图在细节层面更加清晰, 边缘结构轮廓更锐利, 纹理细节得以更完整地保留。PGAM 与 MSAM 协同优化了特征融合, 弥补了 PATN 在空间结构建模与细节表达上的不足, 显著增强了姿态迁移任务的生成质量。

5 结 论

本文提出了一种基于 PATN 和自注意力机制的多分辨率人体姿态迁移方法。为了更好地引导图像生成过程, 设计了 PGAM 模块, 通过多头注意力机制增强姿态特征的表示, 作为引导条件控制人体部位, 从而提升网络对全局信息的控制。此外, 添加 MSAM 模块捕捉不同尺度的特征信息, 减少下采样过程中深层空间几何信息的丢失。引入三元损失函数在生成阶段从整体结构分布、低频准确性、视觉保真度等多维度对生成图像进行约束优化, 实现对图像质量的全局建模。本文在 DeepFashion 和 Market-1501 数据集上进行了实验验证, 结果表明, 本文方法在多个指标上均优于现有方法或与现有最先进方法相当, 特别是在图像质量和关键点准确率方面表现突出, 验证了其有效性。未来可探索本文方法在其他场景的迁移应用。

参 考 文 献:

- [1] JIANG Y M, YANG S, KOH T L, *et al.* Text2Performer: text-driven human video generation [C]//*Proceedings of 2023 IEEE/CVF International Conference on Computer Vision*. Paris: IEEE, 2023: 22747-22757.
- [2] KULAL S, BROOKS T, AIKEN A, *et al.* Putting people in their place: affordance-aware human insertion into scenes [C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver: IEEE 2023: 17089-17099.
- [3] MIRZA M, OSINDERO S. Conditional generative adversarial nets [J/OL]. *arXiv*, 2014: 1411.1784.
- [4] MA L Q, JIA X, SUN Q R, *et al.* Pose guided person image generation [C]//*Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach: Curran Associates Inc., 2017: 405-415.
- [5] ESSER P, SUTTER E. A variational U-net for conditional appearance and shape generation [C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018: 8857-8866.
- [6] SIAROHIN A, SANGINETO E, LATHUILIÈRE S, *et al.* Deformable GANs for pose-based human image generation [C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018: 3408-3416.
- [7] ZHU Z, HUANG T T, SHI B G, *et al.* Progressive pose attention transfer for person image generation [C]//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 2347-2356.
- [8] 杨大为, 刘志权, 王红霞. 结合混合卷积和多尺度注意力的视频异常检测算法[J]. 液晶与显示, 2024, 39(8): 1128-1137.
YANG D W, LIU Z Q, WANG H X. Video anomaly detection algorithm combining mixed convolution and multi-scale attention [J]. *Chinese Journal of Liquid Crystals and Displays*, 2024, 39(8): 1128-1137. (in Chinese)
- [9] ISOLA P, ZHU J Y, ZHOU T H, *et al.* Image-to-image translation with conditional adversarial networks [C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017: 1125-1134.
- [10] ZHANG H, GOODFELLOW I J, METAXAS D N, *et al.* Self-attention generative adversarial networks [C]//

- Proceedings of the 36th International Conference on Machine Learning*. Long Beach: PMLR, 2019: 7354-7363.
- [11] ZHANG F, SHI Q X, MA Y L. Combining self-attention and depth-wise convolution for human pose estimation [J]. *Signal, Image and Video Processing*, 2024, 18(8): 5647-5661.
- [12] ZHANG J S, LI K, LAI Y K, *et al.* PISE: person image synthesis and editing with decoupled GAN [C]// *Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville: IEEE, 2021: 7982-7990.
- [13] ZHANG P Z, YANG L X, LAI J H, *et al.* Exploring dual-task correlation for pose guided person image generation [C]// *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans: IEEE, 2022: 7713-7722.
- [14] BHUNIA A K, KHAN S, CHOLAKKAL H, *et al.* Person image synthesis via denoising diffusion model [C]// *Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver: IEEE, 2023: 5968-5976.
- [15] HUANG E B, ZHANG Y, HUANG F L, *et al.* DRDM: a disentangled representations diffusion model for synthesizing realistic person images [C]// *Proceedings of 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*. Hyderabad: IEEE, 2025: 1-5.
- [16] LIU Y X, JIANG T, SI P J, *et al.* Unpaired semantic neural person image synthesis [J]. *The Visual Computer*, 2025, 41(1): 383-397.

作者简介:



黄庆东,男,博士,副教授,2011年于西安电子科技大学获得博士学位,主要从事图像处理、计算机视觉、信号与信息处理方面的研究。E-mail: huangqingdong@xupt.edu.cn



苏宇辉,男,硕士研究生,2023年于西安邮电大学获得学士学位,主要从事图像处理、人体姿态迁移方面的研究。E-mail: 2631461497@qq.com