

中图法分类号: TP392 文献标识码: A 文章编号: 1006-8961(2025)02-0375-16

论文引用格式: Xu Z G, Pu B C, Qin J M, Xiang Y P, Peng Z J and Song C F. 2025. Data benchmark and model framework for high-definition human image generation. *Journal of Image and Graphics*, 30(02):0375-0390(徐正国, 普碧才, 秦建明, 项炎平, 彭振江, 宋纯锋. 2025. 面向高清人体图像生成的数据基准与模型框架. *中国图象图形学报*, 30(02):0375-0390)[DOI:10.11834/jig.240159]

面向高清人体图像生成的数据基准与模型框架

徐正国¹, 普碧才¹, 秦建明¹, 项炎平², 彭振江³, 宋纯锋^{3*}

1. 云南电网有限责任公司怒江供电局, 怒江 673199; 2. 智慧互通科技股份有限公司, 北京 100081;
3. 中国科学院自动化研究所智能感知与计算研究中心, 北京 100190

摘要: 目的 姿态引导下的人物图像生成具有广泛的应用潜力, 受到了广泛关注。低分辨率场景的姿态引导人物图像生成任务取得了很大成功。然而在高分辨率场景下, 现有的人体姿态迁移数据集存在分辨率低或多样性差等问题, 同时也缺乏相关高分辨率图像生成方法。针对这一问题, 构建了具有多模态辅助数据的大规模高清人物图像数据集 PersonHD。方法 PersonHD 数据集收集了包含 100 个不同人物的 299 817 幅图像。在提出的 PersonHD 基础上, 基于现有数据集的公共设置, 本文进一步构建了两个不同分辨率下的评测基准, 并设计了一个实用的高分辨率人物图像生成框架, 为评估最先进的姿态引导人物图像生成方法提供了一个新的平台。结果 与现有数据集相比, PersonHD 在更高的图像分辨率、更多样化的人物姿态和更大规模的样本方面具有显著的优势。基于 PersonHD 数据集, 实验在两个不同分辨率的评测基准上系统地评估了当前具有代表性的姿态引导人物图像生成方法, 并对本文提出框架各模块的有效性进行了系统验证。实验结果表明, 该框架具有良好的效果。结论 本文提出的高清人物图像生成基准数据集具有高分辨率数据规模大、多样性强等特点, 有助于更为全面地评估姿态引导下的人物图像生成算法。本文的数据集和代码可在 <https://github.com/BraveGroup/PersonHD> 上获得。

关键词: 人物图像合成; 姿态引导迁移; 高清数据集; 低分辨率基准; 高分辨率基准

Data benchmark and model framework for high-definition human image generation

Xu Zhengguo¹, Pu Bical¹, Qin Jianming¹, Xiang Yanping², Peng Zhenjiang³, Song Chunfeng^{3*}

1. Nujiang Power Supply Bureau, Yunnan Power Grid Co., Ltd., Nujiang 673199, China;
2. Intelligent Interconnection Technology Co., Beijing 100181, China; 3. Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract: **Objective** Pose-guided person image generation has attracted considerable attention because of its wide application potential. In the early stages of development, researchers relied mainly on manually designing features and models, matching key points between different characters, and then achieving pose transfer via interpolation or transformation. With the rapid development of deep learning technology, the emergence of generative adversarial networks (GANs) has led to considerable progress in posture transfer. GANs can learn and generate realistic images, and variants of related generative adversarial networks have been widely used in pose transfer tasks. Moreover, deep learning has made progress in key

收稿日期: 2024-03-18; 修回日期: 2024-05-14; 预印本日期: 2024-05-21

* 通信作者: 宋纯锋 chunfeng.song@nlpr.ia.ac.cn

基金项目: 国家自然科学基金项目 (62006231)

Supported by: National Natural Science Foundation of China (62006231)

point detection technology. Advanced key point detection models, such as OpenPose, can more accurately capture human pose information, providing tremendous assistance for the development of algorithms in related fields and the construction of datasets. Recent works have achieved great success in pose-guided person image generation tasks with low-definition scenes. However, in high-resolution scenes, existing human pose transfer datasets suffer from low resolution or poor diversity, and relevant high-resolution image generation methods are lacking. This issue is addressed by constructing a large-scale high-definition human image dataset named PersonHD with multimodal auxiliary data. **Method** This study constructs a large-scale, high-resolution human image dataset called PersonHD. Compared with other datasets, this dataset has several advantages. 1) Higher image resolution: the cropped human images in PersonHD have a resolution of $1\,520 \times 880$ pixels. 2) More diverse pose variations: the actions of the subjects are closer to real-life scenarios, introducing more fine-grained nonrigid deformation of the human body. 3) Larger image size. The PersonHD dataset contains 299 817 images from 100 different people in 4 000 videos. On the basis of the proposed PersonHD, this study further constructs two benchmarks and designs a practical high-resolution human image generation framework. Given that most existing works address human images with a resolution of 256×256 pixels, this study first establishes a low-resolution (256×256 pixels) benchmark for general evaluation, evaluates the performance of these methods on the PersonHD dataset, and further improves the performance of the state-of-the-art methods. This study also constructs a high-definition benchmark (512×512 pixels) to verify the performance of the state-of-the-art methods on the PersonHD dataset. These two benchmarks also enable this study to rigorously evaluate the performance of existing and future human pose transfer methods. In addition, this study proposes a practical framework to generate higher-resolution and higher-quality human images. In particular, this study first designs semantically enhanced partwise augmentation to solve the challenging overfitting problem in human image generation. A conditional upsampling module is then introduced for the generation and further refinement of high-resolution images. **Result** Compared with existing datasets, PersonHD has significant advantages in terms of higher image resolution, more diverse pose variance, and larger sample sizes. On the PersonHD dataset, experiments systematically evaluate the current representative pose-guided character image generation methods on two different resolution evaluation benchmarks and systematically validate the effectiveness of each module of the framework proposed in this study. The experiment used five indicators to quantitatively analyze the performance of the model, including the structure similarity index measure (SSIM), Frechet inception distance (FID), learned perceptual image patch similarity (LPIPS), mask LPIPS, and percentage of correct keypoints with head-normalization (PCKh). For low-resolution benchmarks, most current methods are designed for low-resolution images of size 256×256 pixels. The image size was adjusted to 256×256 pixels during the experiment to evaluate these methods on the PersonHD dataset. Moreover, PersonHD was split into two subsets, namely, a clean subset and a complex subset, to evaluate the processing ability of different models for different backgrounds. This method compares several of the latest methods on two subsets of PersonHD, including pose-attentional transfer network (PATN), Must-GAN, Xing-GAN, pose-guided image synthesis and editing (PISE), and semantic part-based generation network (SPGNet). During the experiment, semantically enhanced partwise augmentation and one-shot-driven personalization were used to improve the performance of SPGNet as the baseline. The semantically enhanced partwise augmentation and one-shot driven personalization proposed in this research improve the performance of SPGNet in terms of multiple indicators, and the relevant modules of the framework significantly improve the model's performance. For high-resolution benchmarks, given the limited work on pose-guided character image generation at high resolution, this study uses a conditional upsampling module to design the most advanced SPGNet model and further improves performance by using semantically enhanced partwise augmentation methods and one-shot-driven personalization. The experimental results indicate that the framework has good performance. **Conclusion** A large-scale high-resolution person image dataset called PersonHD, which contains 299 817 high-quality images, was developed. Compared with existing datasets, PersonHD has significant superiority in terms of higher image resolution, more diverse pose-variance, and larger scale samples. The high-definition character image generation benchmark dataset proposed in this article has the characteristics of a large scale and strong diversity of high-resolution data, which helps to comprehensively evaluate pose-guided character image generation algorithms. Comprehensive benchmarks were established, and extensive experimental evaluations were implemented on the basis of general low-definition protocols and the first proposed high-definition protocols, which could contribute to an important platform for analyzing recent state-of-the-art person

image generation methods. A unified framework for high-definition person image generation, including semantically enhanced partwise augmentation and a conditional upsampling module, was also used. Both modules are flexible and can work separately in a plug-and-play manner. The dataset and code proposed in this work are available at <https://github.com/BraveGroup/PersonHD>.

Key words: human image synthesis; pose-guided transfer; high-definition dataset; low-definition benchmark; high definition benchmark

0 引言

随着计算机视觉的发展,人们不再满足于理解相机拍摄的照片,而是试图创造并生成逼真的图像。人物图像生成在虚拟试戴、虚拟现实和人机交互等领域具有广泛的应用潜力,是一个备受关注的热点研究方向。姿态引导的人物图像生成是人物生成的重要子任务(Qian等,2018;Zhu等,2019;Yang等2020b;Men等,2020),其目的是只通过一个人物图像模板和目标姿态,生成具有目标姿态的逼真的人物图像,这个任务也称为人体姿态迁移。近年来,研究人员提出了一系列的方法和数据集,特别是基于有监督学习的人体姿态迁移的方法发展迅猛(Zhu等,2019;Ma等,2017;Siarohin等,2018;Zhang等,2021)。除了有监督方法外,近期也有工作探索了无监督设置下的人体姿态迁移方法(Pumarola等,2018;Tang等,2019;Song等,2019;Ma等,2021)。这些工作取得了巨大的成功,但在生成人物细节的保留方面仍存在一定的问題,例如在保留衣物的复杂质地与罕见颜色等方面仍有许多不足之处。

1)人物图像生成数据集。为了推动人物图像生成研究,近期研究人员陆续提出了若干人物图像数据集。其中对于人体姿态驱动的人物生成而言,最常用的数据集是Market-1501(Zheng等,2015)和DeepFashion(Liu等,2016)。

Market-1501是一个较低分辨率(128×64 像素)的人物图像数据集,该数据集由校园监控镜头收集,最初用于行人重识别任务。由于人物图像是在校园场景中采集的,该数据集的人物图像的姿态和视角差异很大。如今,高质量的图像需求越来越多,使用Market-1501来开发并评估更高分辨率的人物图像生成方法存在诸多限制。

DeepFashion数据集在一定程度上实现了高分辨率的人像生成需求,其平均分辨率为 $1\,101 \times 750$ 像素,而其最常用的数据集版本则是其低分辨率版

本(256×256 像素)。DeepFashion是一个具有多样化样本的大规模数据集。由于DeepFashion最初是面向为服装识别和检索研究任务提出的,该数据集关注的是时尚服装而不是人体,因此该数据集中全身图像的比例仅为6.8%(Fu等,2022)。此外,DeepFashion数据集是从购物网站收集的,图像中的人物图像大多来自于有限姿态的时尚模特。这些姿态是预定义的艺术照经典拍照姿态,这与日常情况的普通人姿态也存在较大差异。同时该数据集具有相当干净的拍摄背景,也无法泛化至真实场景中。总体而言,DeepFashion缺乏足够多样化的人体姿态和真实背景。

此外,近年来也有一些新颖的人物图像数据集相继提出,其中MPV(Dong等,2019)与DeepFashion类似,但数据集规模相对较小,图像分辨率仅为 256×192 像素。尽管现有数据集为该领域的发展做出了巨大的贡献,但这些数据集相对于人物图像生成任务需求而言并不完善。最近提出的一些高分辨率视频数据集存在ID(identity)数量较少、图像规模较小的问题。iPER(Liu等,2019)是一个新建立的流行数据集,但它只包含30个ID,没有配备如骨架关键点和语义人物分割图等辅助信息。数据集StyleGAN-Human-HQ(Fu等,2022)可能是与本文最相关的数据集,其平均分辨率达到了 $1\,024 \times 512$ 像素。尽管该数据集具有大规模和高分辨率的优点,但该数据集是从互联网上收集的,仍然不接近日常场景。

虽然以上工作在人体图像生成数据集建设方面取得了很大进展,但仍然存在两个关键问题尚未解决:缺乏高分辨率数据和相应的高清人物图像生成方法。综上所述,现有的方法大多侧重于低分辨率的人体图像生成,而高分辨率的人体图像生成方法则是业界迫切需要的。因此,需要建立高分辨率的人体图像生成数据集和对应方法。

2)有监督姿态引导的人物图像生成。需要配对的人物图像来监督模型的训练。图像生成模型

(Isola 等, 2017; Wang 等, 2018; Karras 等, 2020)近年来发展迅速。基于姿态的人物图像生成作为图像生成的一个重要子领域(Roy 等, 2023; Khatun 等, 2023)同样取得了巨大的进步。Ma 等人(2017)首先将 Conditional GAN(conditional generative adversarial network)(Mirza 和 Osindero, 2014)应用于姿态引导的人物图像生成,直接将源图像和目标姿态关联起来生成目标图像,然而该方法很难生成逼真的图像。为了处理源姿态和目标姿态之间的差异,Siarohin 等人(2018)提出了 Deformation GAN,在特征映射层面对人体进行变形。Zhu 等人(2019)将姿态传递分解为局部变形序列,并引入注意力机制对不同图像块分配权重。Li 等人(2020)提出了一种姿态引导的非局部注意力机制,用于选择图像特征中有价值的区域。一些工作(Li 等, 2019; Ren 等, 2020; Ma 等, 2023)使用源姿态和目标姿态之间的光流来更好地扭曲源图像的表观信息。Tang 等人(2020)提出了 Xing-GAN 模型,该模型采用两条分支交叉的方式有效地传递和更新了人体形状和外观特征嵌入。Li 等人(2021)提出了 PoT-GAN,通过操纵相应的多尺度特征图来显式地学习姿态之间的迁移。Liu 等人(2021)用分割掩码和解析图代替姿态。一些工作(Yoon 等, 2021; Ren 等, 2021)介绍了一系列针对人体运动的姿态迁移工作。Ly 等人(2021)预测了目标姿态的语义分割图,然后用预测的目标语义分割图通过区域间归一化来指导图像生成。Zhang 等人(2022)引入了一种辅助任务(source-to-source task),利用双任务相关性增强生成目标图像的纹理细节。为了获得更好的性能,Karmakar 和 Mishra(2019)提出了一些数据增强技术,如 Random Crop、Random Distortion 和 Random Jitter,以实现畸变、光照和尺度的变化。这些方法仍存在细节不足、生成图像不稳定以及不自然等问题。李健等人(2024)引入了变分自编码器网络、对比语言—图像预训练网络与门控循环单元网络,实现基于文本描述特征生成三维运动人体模型。

3)无监督姿态引导的人物图像生成。不需要成对的源—目标图像进行训练。Pumarola 等人(2018)首先提出了与 CycleGAN(Zhu 等, 2017)类似的无监督人物图像生成方法,即给定一幅图像,该方法在给定的姿态下渲染新图像,然后渲染回原始姿态,以生成图像的三元组。Tang 等人(2019)也提出了一种

循环损失更大的 Cycle-in-Cycle 方法来引入更多的自监督。Song 等人(2019)使用 CycleGAN 将人物姿态迁移分解为语义分割转换和外观生成,从而将姿态迁移任务变为两个更容易实现的子任务。Ma 等人(2017)对源人物图像的多层次外观特征进行转移,并与姿态特征合并,重构源图像。

4)姿态引导人物高分辨率图像的生成。近年来有一些针对高分辨率姿态引导人物图像生成的研究工作。Albahar 等人(2021)提出了 Pose-with-Style,并在 348×512 像素分辨率下进行了人体图像生成。Zhou 等人(2021)提出了 CoCosNet v2,在 512×512 像素分辨率下进行人体图像生成。与低分辨率方法相比,高分辨率方法面临着更多的挑战,如模型稳定性和计算开销。

为了弥补当下姿态引导人物高分辨率图像的生成任务在数据与方法上的不足,本文首先建设了 PersonHD 这一大规模、高分辨率的人物图像数据集。PersonHD 数据集包含来自 100 名志愿者的 299 817 幅高分辨率真实人物图像,且数据集中的人物姿态更接近日常生活场景,并引入了更细粒度的人体非刚性动作形变。同时,在提出的 PersonHD 数据集的基础上,本研究进一步构建了两个评测基准,并设计了一个实用的高分辨率人物图像生成框架。具体而言,本研究首先设计了一种语义驱动增强方法来解决人物图像生成中经常由于数据不足导致的过拟合问题,然后引入了条件超分辨率采样单元,用于高分辨率图像的生成和进一步细节化。本研究首先建立一个低分辨率(256×256 像素)基准进行通用评估,评测了当前主流方法在 PersonHD 数据集上的表现,并借助本数据集进一步提升最先进方法的性能。本文还构建了一个高分辨率基准(512×512 像素)来验证面向高分辨率的姿态引导人物图像生成方法的表现。这两个基准可为当前和未来的多种分辨率条件下的人体图像生成的开发与评测提供重要支撑。

1 PersonHD 数据集

在本节中将介绍一个大规模、高清的人物图像数据集——PersonHD 数据集。本文提出的 PersonHD 数据集的贡献包含两部分。首先,本研究建立了一个更接近现实生活场景的数据集,为先进的人

物图像生成方法的评估提供了一个平台。其次,本研究引入了一个全面且具有挑战性的评测基准,用于评估低分辨率与高分辨率的姿态引导人物图像生成方法。

1.1 数据构建

1) 图像采集。本研究使用罗技 C922 网络摄像头采集 PersonHD 图像数据。采集的原始 RGB 视频分辨率为 $1\,920 \times 1\,080$ 像素 (1080P), 帧率为 30 帧/s。本研究邀请了 100 名志愿者参与数据的采集工作。每位受试者穿着自己的日常服装, 完成 20 个指定的动作。指定动作包括: A—摇头, B—点头, C—招手, D—侧勾, E—抓, F—左划, G—右划, H—投掷, I—迈步, J—捏腰, K—单腿站立摆“大”字, L—深蹲, M—出拳, N—向前迈步, O—迈步并伸展手臂, P—鼓掌, Q—击球, R—坐下挥手, S—坐下挠头, T—抬起腿坐着。每位志愿者都被要求朝不同的角度多次重复所有动作。数据采集过程中使用两台相机同时从两个不同的地平线视角捕捉相同的动作样本, 因

此每个动作样本都有两个不同的背景与捕捉视图: 其中一个背景是干净的白色幕布, 另一个背景是较为复杂的会议室场景, 如图 1 所示。

2) 数据清洗和数据处理。为了获取更多图像, 本方法从视频中随机提取图像帧进行进一步研究。考虑到部分视频帧存在运动模糊, 本研究采用运动模糊检测方法来筛选过滤存在运动模糊的帧。为了提高前景主体的有效占比, 本文方法还将所选图像从 $1\,920 \times 1\,080$ 像素分辨率裁剪为 $1\,520 \times 880$ 像素分辨率, 以此减少冗余的背景信息。由于当前人体图像生成方法需要人体关节关键点和人体语义分割图进行驱动, 本研究采用 OpenPose (Cao 等, 2017) 提取 18 个人体关节点, 并利用 JPPNet (Liang 等, 2019) 预测人体分割剪影作为附加中间数据。

1.2 数据统计

数据采集过程中共有 100 个不同体型、身高和性别的实验对象。在整个数据集中, 每个受试者被分配一个特定的 ID 号 (1—100)。受试者年龄在

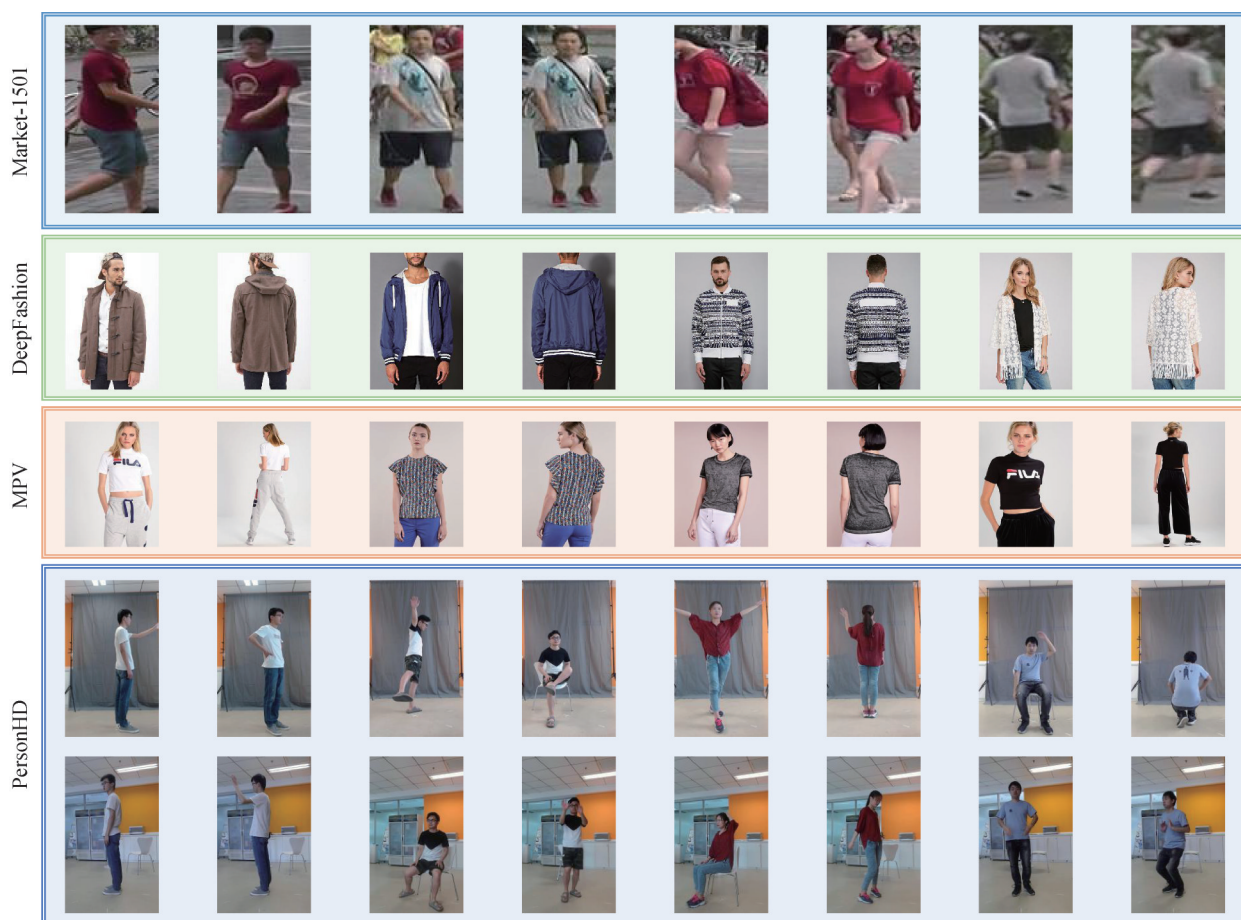


图1 PersonHD与现有数据集的定性比较

Fig. 1 Qualitative comparison of PersonHD with existing datasets

19岁~56岁,身高在1.5 m~1.9 m,体重在42 kg~113 kg。由于该数据集在夏天收集,被采集人物的整体着装比较清凉:很少外套(大衣),大部分是短袖,没有帽子。PersonHD数据集的服装样式和受试者体型类别的详细统计信息如图2所示。

整个数据集包含来自4 000个视频序列的299 817幅图像,其中干净背景的图像为140 817幅,复杂背景的图像为159 000幅。本研究将背景干净的子集命名为Clean-subset,将背景复杂的子集命名为Complex-subset,与其他数据集的对比如表1所示。

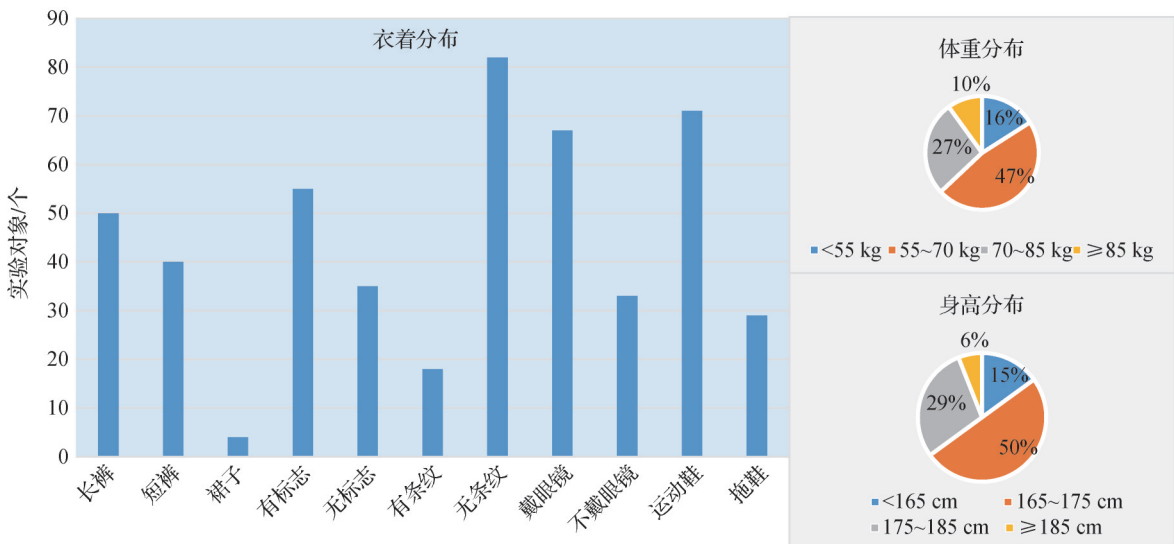


图2 PersonHD数据集的统计信息
Fig. 2 Statistics for the PersonHD dataset

表1 PersonHD与其他数据集的对比情况
Table 1 Comparison of the PersonHD and other datasets

数据集	年份	图像分辨率/像素	背景	数据规模
Market-1501	2015	128 × 64	复杂	1 501 个人物, 32 668 幅图像
DeepFashion	2016	1 101 × 750	干净	11 276 个人物, 52 712 幅店内服装图像
MPV	2019	256 × 192	干净	35 679 个人物, 13 534 幅店内服装图像
iPER	2019	—	复杂	30 个人物, 206 段视频
StyleGAN-Human	2022	1 024 × 512	干净	231 176 幅图像
PersonHD	2024	1 520 × 880	复杂、干净	100 个人物, 299 817 幅图像

注:“—”表示该数据集为视频数据集,视频帧分辨率为256 × 256 像素。

1.3 数据属性

与现有数据集相比,PersonHD具有以下几个显著的优势。

1)更高的图像分辨率。PersonHD 的图像分辨率为1 520 × 880 像素,对比 DeepFashion 的常用图像分辨率为256 × 256 像素,而 Market-1501 的图像分辨率仅为256 × 128 像素。

2)更丰富的姿态多样性。PersonHD 有更多不同人体姿态的图像,通过精心设计的20种动作,描述了更详细和细粒度的人体非刚性变形。如图1所

示,与其他数据集的模型照片姿态相比,图像姿态更接近于日常生活。此外,PersonHD 中的主体通常为全身,而其他数据集中的主体通常为半体。

3)更大的数据规模。PersonHD 中包含来自100个实验对象的299 817幅图像,数据规模超越了现有主流人物图像生成数据集。

1.4 基准评估框架

为了进行全面比较,本研究为提出的 PersonHD 设置了评估方法。如前所述,本文方法根据背景条件将整个数据集分成两个独立的子集,即干净背景

和复杂背景。对于每个子集,本研究根据不同的主题,以8:2的固定比例将所有图像划分为训练/测试集。对于需要成对图像的监督方法,本研究随机选择200 000对用于训练,10 000对用于测试。对于只需要单幅图像的无监督方法,本研究分别从上述训练对中选择相同的33 669幅图像作为训练数据,从上述测试对中选择相同的13 526幅图像作为测试数据。下面对不同分辨率下的具体基准进行具体定义。

1)低分辨率(low-definition, LD)设置方法。考虑到现有的大多数人物图像生成模型都是依据固定的输入图像尺寸 256×256 像素来设计的,显然本文提出的PersonHD基准应兼容默认分辨率。具体来说,本研究直接将图像从 $1\,520 \times 880$ 像素降采样到 256×256 像素。对应地,语义分割图和人体关键点信息都相应地调整了大小,并且这些中间信息是从高清图像中获取或计算得到的,因此其质量可以得到保证。该设置支持当前广泛采用的低分辨率人体图像生成算法的开发与性能评估。

2)高分辨率(high-definition, HD)设置方法。为推动高清人物图像生成研究的发展,本文方法提出一种高分辨率(512×512 像素)设置,为面向实用场景的姿态引导人物图像生成提供新的尝试。事实上,本研究的数据集可以支持更高的分辨率(如1 080 P),

但考虑到目前计算资源的限制,本文暂时没有设置更高分辨率的基准,留待未来继续探索。除了默认分辨率进行调整外,HD方法的其他设置均与LD相同。

2 高清人物图像生成的统一框架

为满足高分辨率和高质量的人物图像生成需求,应基于PersonHD数据集基准建立一个统一的框架。然而,针对姿态引导人物图像生成方法的性能提升,需要解决两个关键问题:1)虽然大多数方法在全局形状一致性上的表现良好,但在保留局部纹理特征方面仍然存在不足;2)现有的工作大多集中在处理低分辨率(如 256×256 像素)的人物图像,针对高分辨率人物图像生成的研究较少。为了解决这些问题,本文提出了一个统一的高分辨率人物图像生成框架。整体框架如图3所示。本文的框架包含多个阶段:1)在数据准备阶段,结合数据集中的源图像与语义分割图标注,基于语义驱动的部件级增强模块获取多样化的姿态迁移训练数据,基于当前最优的姿态迁移生成方法进行模型训练;2)完成模型训练后,选择源图像与目标姿态作为人物图像生成数据对,将数据对送入图像生成模型进行姿态迁移人体图像生成,获取低分辨率的姿态引导人物图像

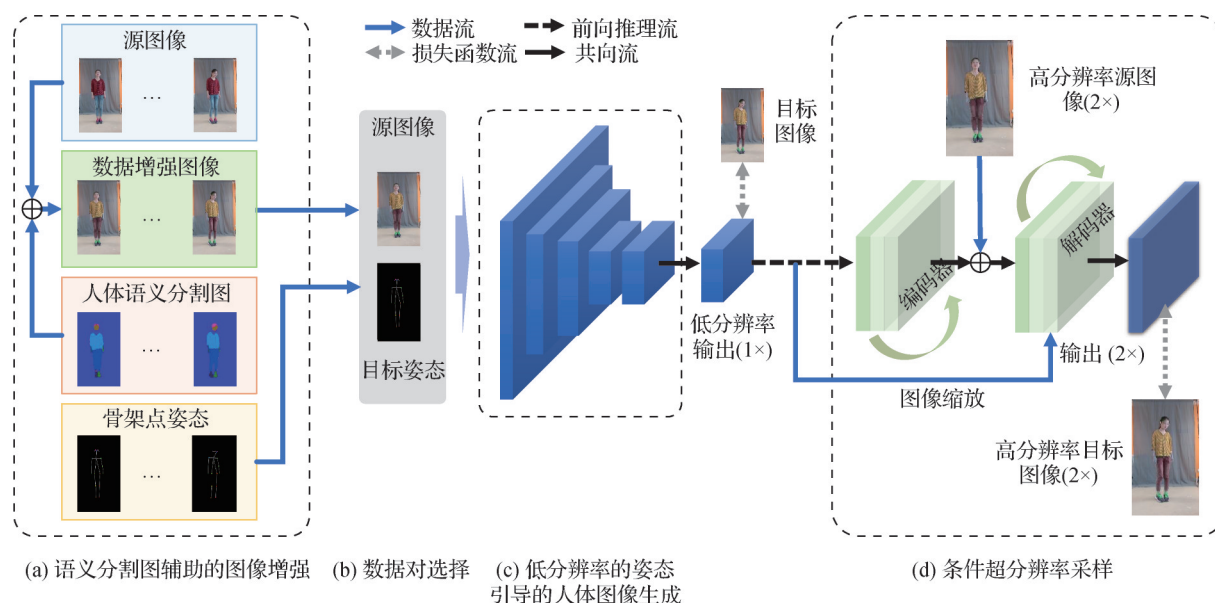


图3 高分辨率人物图像生成框架

Fig. 3 The framework for high-resolution person image generation((a)image enhancement assisted by semantic segmentation graphs;(b)data pair selection;(c)pose guide image generation of the human body based on low resolution image;(d)conditional super-resolution sampling)

生成结果;3)进行二阶段的高清人物图像生成。基于条件超分辨率采样单元,通过输入高分辨率的源图像与一阶段生成的低分辨率姿态迁移生成图像,生成高分辨率姿态迁移人物图像,通过两阶段方法搭建起一个计算量低且性能可靠的高分辨率人物图像生成框架,并且有效避免了现有方法处理高分辨率图像所面临的困难。此外,本研究采用 one-shot 个性化驱动模块,进一步提升了框架性能。

2.1 语义驱动的部件级增强方法

为了增加数据集中训练数据的多样性,本研究提出了语义驱动的部件级增强方法,这是一种通用的技术。PersonHD 中有 100 名实验对象,每名实验对象都身着自己的日常服装。若为了增加数据多样性而要求实验对象多次换装,必然会增加数据收集的难度。因此,本研究提出了语义数据增强方法,可

以让实验对象在不付出额外劳动的情况下实现自动换装。具体来说,本研究根据人类语义分割图对人体部位进行划分,并随机改变每个部位的颜色和光照。本研究使用 JPPNet(Liang 等,2019)将人体划分为 20 个语义部分。由于面部颜色相对稳定,只在面部使用非常轻微的颜色抖动。为了保持左右对称性,本研究在两只鞋子、两只手臂和两条腿上使用相同的数据增强方法。对于其他部分,本研究采用不同的颜色变化和颜色抖动。基于以上数据增强方法,数据集中服装的局部多样性得到了极大的增强,也带来了明显的性能提升。本研究在训练阶段实现了语义驱动的部件级增强方法(SEP-Aug),并对每组数据对进行一致性转换。SEP-Aug 可以在提供人体语义分割图的情况下推广至其他应用程序。SEP-Aug 的可视化如图 4 所示。



图 4 语义驱动的部件级增强方法可视化

Fig. 4 The visualization of semantic enhanced part-wise augmentation

2.2 One-shot 个性化驱动

虽然在大数据集上进行训练可以保证人物图像生成模型在数据域内就具备一定的泛化性,但由于特征域的差距,模型往往不能很好地推广到训练集特征域外的输入。例如,如果模型之前没有学习过某种特定衣服的颜色和纹理,就难以实现对布料信息(如具体的颜色和纹理)的编解码。在实践的测试阶段,每个实验样本都有一个源人物图像可以用于

模型微调。Liu 等人(2022)提出了 one/few shot 的个性化人物图像生成器,首先在大数据集上进行模型预训练,然后通过对预训练模型进行微调,从有限数量的图像中学习每个人物的特定图像生成器。然而,该方法的模型需要对所有个例进行训练,这将在推理阶段引入较高的时间成本。

在实验中,结果显示每个测试集样本的迁移都是不稳定且具有较大计算开销的。由于每个样本都

只有单幅源图像,模型很容易过度拟合。本文认为使用每个穿衣个体进行个性化适应的策略难以得到稳定的训练结果且会带来过高的时间成本。为了解决这些问题,本研究探索同时使用所有测试集中穿衣个体的个性化适应。在模型图像生成测试阶段,本研究基于改进后的 one-shot 个性化驱动使用从测试集获得的穿衣个体的源图像对模型进行整体微调。one-shot 个性化驱动是实践中的一个实用技巧。在后续实验评估中,经过 one-shot 个性化驱动的微调模型在测试集上具有更好的表现,微调模型在所有测试集对象上都具有良好的泛化性。

2.3 基于条件的超分辨率采样单元

重新思考构建 PersonHD 的初衷,本研究面向解决姿态迁移下的高分辨率人物图像生成问题。

本研究进一步提出了条件超分辨率采样单元,在低分辨率输出的基础上生成最终的高清分辨率图像。受近期工作(Zhang 等, 2019; Yang 等, 2020b; Jiang 等, 2021; 熊巍 等, 2023; 孙阳 等, 2024)的启发,本研究借鉴超分辨领域的思想设计了基于条件的超分辨率采样单元。通过输入高分辨率的源图像和低分辨率姿态引导生成图像来合成高分辨生成图像,将低分辨率人物图像生成方法与条件超分辨率采样单元结合为一个统一的框架,以低成本生成高清图像。由于使用高清参考图像来恢复低分辨率图像的细节,本研究推测低分辨率生成图像也可以受益于共享相同身份信息的高清源图像。

基于条件的超分辨率采样单元通过引入额外的高分辨率(HD)源图像,将其用于增强低分辨率(LD)姿态引导生成图像,最终生成高分辨率姿态引导人物图像。为了提升匹配变换的鲁棒性,该模块使用对比对应网络,使用输入图像的增强视图,学习图像匹配变换的不变性特征。为了解决图像重建过程中的分辨率差距,该模块采用师生相关蒸馏方法,从较容易的源图像高分辨率匹配(HD—HD)过程中提取知识,基于KL(Kullback-Leibler)散度将高分辨率图像重建知识嵌入至学生模型的高分辨率图像重建过程,指导较模糊的姿态引导图像的低分辨率至高分辨率(LD—HD)的匹配过程。该模块用于在完成低分辨率姿态引导的图像生成后,将低分辨率图像进行超分辨率上采样生成姿态引导的高分辨率图像。本文提出的超分辨率采样单元具有放大图像分辨率至不同层次(如从2倍放大至4倍)与复现更多

细节的能力,该模块非常灵活,可以通过即插即用方式插入到任何现有的方法中。

3 实验

在 PersonHD 数据集的基础上,本研究在建立的两个基准上系统地评估了具有代表性的姿态引导人物图像生成方法,所有训练方案均遵循原始论文。

在 256×256 像素的低分辨率基准上,本研究评估了现有的方法,并分析了本文的改进方案。在 512×512 像素的高清基准测试上,本研究的主要目标是验证框架中每个模块的有效性。最后,针对未来的研究方向展开了讨论。

3.1 评价指标

本文采用5个定量指标,即SSIM(structure similarity index measure)、FID(Frechet inception distance)、LPIPS(learned perceptual image patch similarity)、Mask-LPIPS和PCKh(percentage of correct keypoints with head-normalization)。SSIM(Wang 等, 2004)通过计算全局方差和均值来评估图像对的结构相似性。FID(Heusel 等, 2017)通过计算生成图像分布与真实图像之间的Frechet距离来衡量生成图像的真实性。LPIPS(Zhang 等, 2018)通过测量生成图像与目标图像在特征空间中的距离来评估相似性,该指标的评估结果与人类感知接近。Mask-LPIPS排除了背景的影响来进行人体测量。值得注意的是,考虑到人体形状的差异,本文使用语义分割图代替人体关键点位置来实现背景滤除。PCKh(Andriluka 等, 2014)通过比较人体关节的定位结果来计算姿态的一致性。除了针对这些定量指标的计算结果进行对比,本文还展示了不同方法的可视化结果。

3.2 256×256 像素低分辨率基准

目前,大多数模型都是针对 256×256 像素的低分辨率图像设计的。为了在 PersonHD 数据集上评估这些模型的性能,本研究将图像降采样为 256×256 像素。本研究将 PersonHD 拆分为干净背景和复杂背景两个部分,并在 PersonHD 的两个子集上比较了几种最新模型的性能,包括PATN(pose-attentional transfer network)(Zhu 等, 2019)、Must-GAN(Ma 等, 2021)、Xing-GAN(Tang 等, 2020)、PISE(pose-guided image synthesis and editing)(Zhang 等, 2021)和SPG-

Net (semantic part-based generation network) (Ly 等, 2021)。

值得注意的是, PATN、Xing-GAN、PISE 和 SPGNet 是有监督方法, 而 Must-GAN 是无监督方法。PISE 和 SPGNet 都是两阶段方法: 首先由语义分割生成器合成与目标姿态对齐的人体语义分割图, 然后由图像生成器生成最终图像。在这个低分辨率基准上, 本文方法使用语义驱动的部件级增强方法和 one-shot 个性化驱动模块, 在基线方法 SPGNet 上进一步优化模型性能。

3.2.1 定性定量对比

图 5 为 PersonHD 的定性对比, 可以观察到以下几点:

1) 无论在干净背景还是复杂背景下, 本文方法进一步提升了生成图像的人体外观效果, 例如本文方法对应的第 1 行女生上衣更接近于源图像中的青色, 第 3 行中的妇女上衣有着更接近源图像的衣物纹理, 第 4 行中的男生上衣更接近源图像中的蓝色。

2) 对比现有的方法可以发现, 大多数现有的方法可以很好地保持生成图像的形状一致性, 但不能

很好地保持服装信息。图 5 第 1 行结果显示, PATN 对保存受试者的衣服和头发效果最差, 通过 PATN 生成的图像不能将上身衣服和裙子区分开, 也无法保留长发等人体的细节特征。在衣服的纹理和颜色方面, SPGNet 优于其他两种方法, 但仍有改进的空间。Must-GAN 生成的人物身着红色上衣, 这可能与 Must-GAN 的粗粒度多级统计有关。由于上装的质地和颜色比较复杂, 仅依靠均值和方差等统计信息可能不足以描述实际情况。

3) 对比不同的背景, 可以看出复杂的背景使得模型难以学习受试者的服装信息。Must-GAN 作为一种无监督的方法, 难以处理较为复杂的背景信息。

PersonHD 的定量结果见表 2。其中, $\uparrow$ ($\downarrow$) 表示更高 (更低) 的数值代表更好的结果。可以观察到, PISE 在干净背景子集的 FID 指标上表现最好, 而 PATN 在复杂背景子集的 FID 指标上表现最好。经本文提出语义驱动的部件级增强方法和 one-shot 个性化驱动模块改善的 SPGNet 在多个指标上都有最好的表现, 框架的相关模块对模型性能的改进效果明显。



图 5 基于低分辨率 PersonHD 基准的图像生成可视化对比
Fig. 5 The visual comparison with the competing methods on low-resolution PersonHD
((a) clean background; (b) complex background)

表 2 低分辨率 PersonHD 上不同方法对比实验结果

Table 2 Experimental results of the quantitative comparison with methods on low-definition PersonHD

方法	提出时间	干净背景					复杂背景				
		SSIM ↑	FID ↓	LPIPS ↓	M-LPIPS ↓	PCK ↑	SSIM ↑	FID ↓	LPIPS ↓	M-LPIPS ↓	PCK ↑
PATN	CVPR19	0.79 4	52.23 1	0.185 2	0.103 3	0.96	0.80 4	40.36 4	0.130 3	0.074 4	0.96
Xing-GAN	CVPR20	0.79 4	65.96 3	0.184 9	0.099 4	0.94	0.78 3	45.22 7	0.145 5	0.077 5	0.96
Must-GAN	CVPR21	0.66 3	49.35 2	0.300 3	0.103 1	0.94	0.44 3	66.88 3	0.465 3	0.086 9	0.93
PISE	CVPR21	0.79 1	47.08 6	0.168 6	0.094 3	0.97	0.64 7	45.00 9	0.206	0.077 7	0.97
SPGNet	CVPR21	0.80 1	58.44 9	0.165 3	0.083 0	0.97	0.60 4	45.24 2	0.270 9	0.069 6	0.97
本文	2024	0.83 6	54.35 4	0.120 1	0.073 6	0.97	0.84 3	43.66 2	0.097 8	0.061 3	0.96

注:加粗字体表示各列最优结果。

表 3 消融实验结果

Table 3 The results of the ablation study

方法	干净背景				复杂背景			
	SSIM ↑	FID ↓	LPIPS ↓	M-LPIPS ↓	SSIM ↑	FID ↓	LPIPS ↓	M-LPIPS ↓
Baseline	0.80 1	58.44 9	0.165 3	0.083 0	0.60 4	45.24 2	0.270 9	0.069 6
Baseline+ Simple Aug	0.79 7	59.49 2	0.161 9	0.082 3	0.61 6	48.18 0	0.267 9	0.068 5
Baseline+ SEP Aug	0.80 3	54.42 0	0.153 4	0.082 6	0.80 1	57.11 3	0.143 0	0.067 2
Baseline+ SEP Aug + ODP	0.83 6	54.35 4	0.120 1	0.073 6	0.84 3	43.66 2	0.097 8	0.061 3

注:加粗字体表示各列最优结果。

3. 2. 2 消融实验

为了探索两种技术的有效性,本文方法在 PersonHD 数据集(256×256 像素)上进行了消融研究,实验基线为 SPGNet。实验结果如表 3 所示,其中 Simple Aug 对应常规数据增强方法(Karmakar 和 Mishra, 2019),SEP Aug 对应本研究提出的语义驱动的部件级增强方法,ODP 对应 one-shot 个性化驱动模块。从实验结果来看,语义驱动部件级增强方法与 one-shot 个性化驱动对姿态迁移人物图像生成的模型性能都有一定程度的提升。

1)与简单的数据增强相比,基于模块化人物图像增强的语义驱动的部件级增强方法具备更好的细粒度与更强的泛化性。细粒度的数据增强而非简单的线性变换式图像增强方法有效提升了数据的多样性,避免了背景对人体语义的干扰,更好地提升了模型的泛化性能。

2)one-shot 个性化驱动不局限于使用姿态迁移的单幅源图像,而是使用获得的所有源图像对模型进行微调。该方法有效避免了模型的过拟合现象,在实验评估中,微调模型在所有测试集对象上都具

有良好的泛化性,在服装的颜色和质感上进一步保持了源图像信息。

3. 2. 3 失败案例研究

失败案例如图 6 所示。源图像和目标图像的真实人体语义分割图是由 JPPNet 预测的。许多方法使用目标语义分割图来指导人物图像的生成。SPGNet 通过输入源图像语义分割图来预测目标姿态的语义分割图。Must-GAN 不预测目标姿态语义分割图,而是使用源图像语义分割图来提取每个人体部位的信息。在实际应用中,语义分割图作为姿态引导的人物图像生成方法的辅助信息存在两个问题:

1)源图像语义分割图可能不准确,甚至具有误导性。在图 6 第 1 行中,源图像语义分割错误地将裙子(红色)映射为裤子(蓝色),因此模型生成的服装是裤子而不是裙子。

2)预测目标的语义分割图较为困难。图 6 第 2 行显示了当预测目标语义分割图存在问题时,将会错误指导人物形象的生成,并对衣服信息的迁移造成较大影响。其中, Must-GAN 和 SPGNet 都没有生成正确形状的裙子。

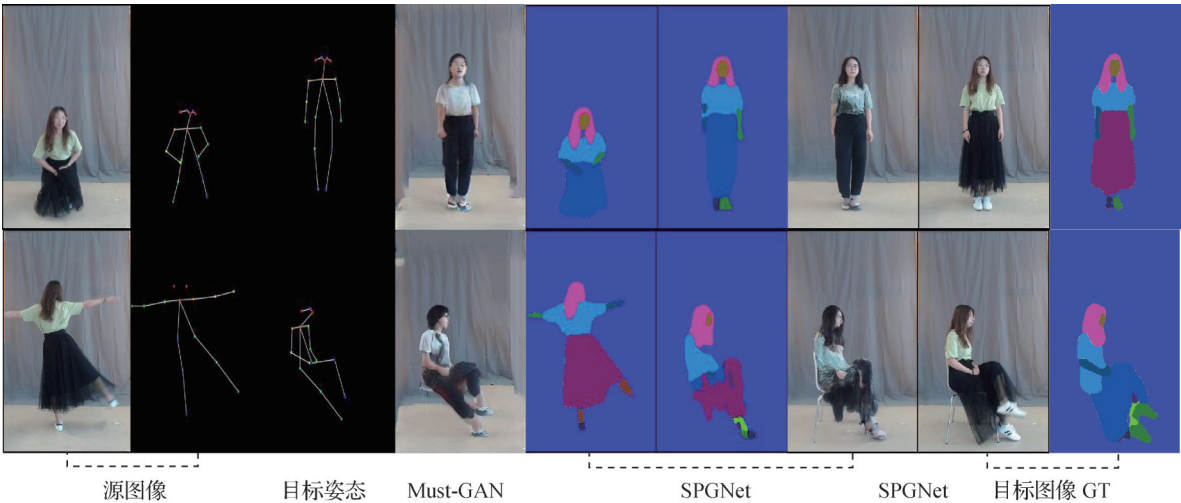


图6 语义分割姿态引导迁移失败案例

Fig. 6 Failure cases of semantic-pose-guided translation

3.3 高分辨率基准

部分针对高分辨率的姿态引导人物图像生成的工作如下所示:Zhou 等人(2021)提出 CoCosNet v2, 在 512×512 像素分辨率下进行人体姿态迁移。因此本研究也在高清基准 (512×512 像素) 上对 CoCosNet v2 的性能进行评价。

可以注意到关注低分辨率 (256×256 像素) 基准的工作比关注高分辨率 (512×512 像素) 基准的工作要多。本研究还想探索将针对低分辨率基准 (256×256 像素) 设计的方法迁移至高分辨率基准 (512×512 像素)。为此,本研究使用了专为低分辨率基准 (256×256 像素) 设计的最先进的 SPGNet 作为基准方法。

当前问题在于如何在高清基准 (512×512 像素) 上评估 SPGNet 方法。第 1 个想法是直接输入分辨率为 512×512 像素的图像并输出相同分辨率图像。但该方法计算成本高,结果不好,如表 4 第 2 行所示。由于生成分辨率为 256×256 像素的图像的相关方法比较成熟,本文方法采用双线性插值来调整低分辨率图像大小。然而该方法结果并不理想,如表 4 第 3 行所示。受到 Zhang 等人(2019)、Yang 等人(2020a)、Jiang 等人(2021)的启发,本文方法最终决定使用基于条件的超分辨率采样单元在低分辨率输出上生成高清图像。本文方法使用语义驱动的部件级增强方法和 one-shot 个性化驱动模块来进一步提升性能。最终框架以较低的计算开销取得了良好的效果。

3.3.1 定性比较

图 7 给出了 PersonHD (512×512 像素) 实验结果的定性对比,CoCosNet 的输出也参与了比较。CoCosNet 的总体输出效果良好,但面对复杂纹理和特定颜色的生成图像仍有不足之处。随着输出维数的增加,图像生成模型的训练变得不稳定,模型难以收敛。当输入分辨率为 512×512 像素的高清图像时,相同分辨率的高清输出有时会出现斑点与伪影。从图 7 的 Base 列的结果可以发现,直接输入高分辨率图像得到的高分辨率输出无法保持手臂和腿部的完整性。而使用条件的超分辨率采样单元(对应 Refined 列)可以得到保留完整人体形态的高质量输出图像。同时,高分辨率人像生成框架中的语义驱动的部件级增强方法(SEP Aug)和 one shot 个性化驱动(ODP) 进一步提升了生成图像的质量。

3.3.2 定量比较

PersonHD 上的定量结果如表 4 所示。在 FID 指标上,CoCosNet 展现出最佳性能,生成的图像更逼真。本文方法提出的高分辨率人体图像生成框架在 SSIM 和 LPIPS 指标上表现得更好,说明在纹理与颜色等特征上提出框架的生成图像与源图像更为接近。由实验结果可知,该框架的各个模块:条件的超分辨率采样单元(Refined)、语义驱动的部件级增强方法(SEP Aug)和 one shot 个性化驱动(ODP)可以有效提升生成图像的质量。

3.4 未来方向和讨论

1)单阶段方法与二阶段方法。针对高分辨率基

表 4 基于高分辨率 PersonHD 不同方法对比实验结果

方法	干净背景				复杂背景			
	SSIM ↑	FID ↓	LPIPS ↓	M-LPIPS ↓	SSIM ↑	FID ↓	LPIPS ↓	M-LPIPS ↓
CoCosNet v2	0.787	46.671	0.244	0.080	0.605	39.284	0.380	0.066
Baseline 512 × 512	0.839	86.083	0.253	0.087	0.745	82.731	0.316	0.072
Baseline + Resize	0.841	60.825	0.280	0.090	0.690	69.828	0.438	0.075
Baseline + Refined	0.848	53.319	0.198	0.084	0.720	61.808	0.336	0.072
SEP Aug + Refined	0.847	54.701	0.195	0.082	0.840	66.623	0.164	0.070
SEP Aug+ ODP + Refined	0.861	58.375	0.178	0.080	0.859	57.702	0.138	0.069

注:加粗字体表示各列最优结果。

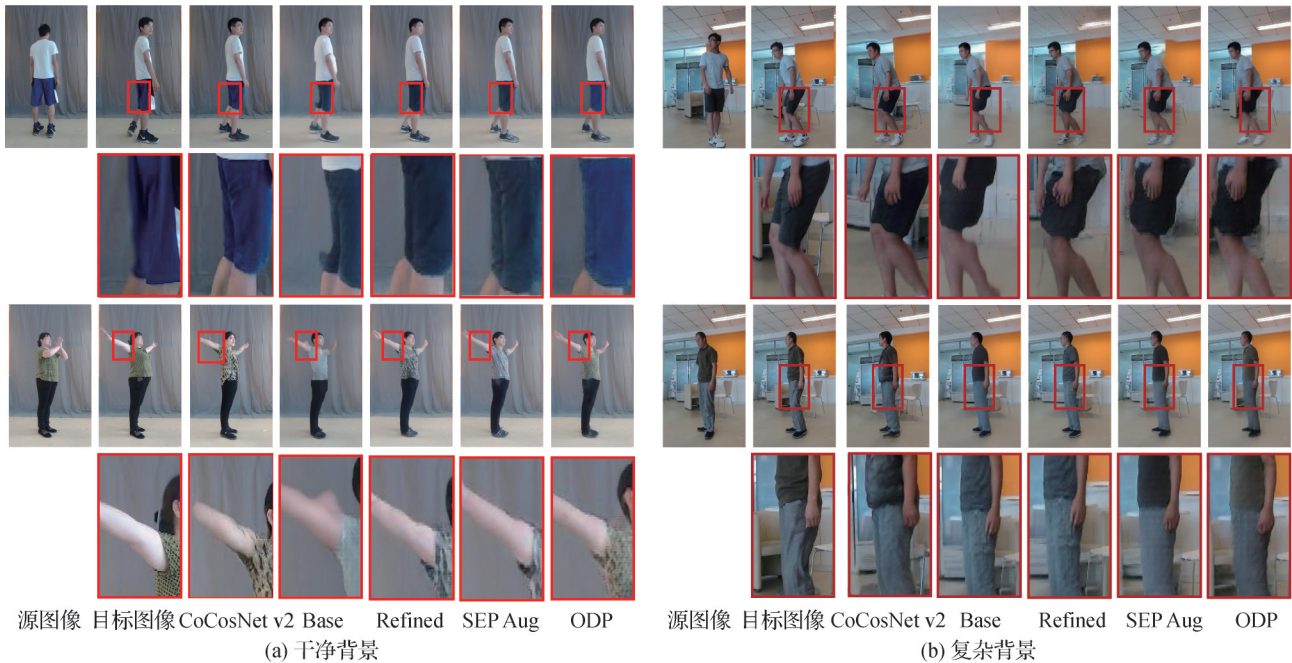


图 7 基于高分辨率 PersonHD 基准的图像生成可视化对比

Fig. 7 Visual comparison of image generation based on high-definition PersonHD ((a) clean background; (b) complex background)

准的探索还需要大量的工作。为了更好地在低分辨率方法和高分辨率需求之间架起桥梁,本文方法提出了一个两阶段框架。未来,本研究将继续探索更高效的单阶段方法。

2)高分辨率带来的挑战。高分辨率人物图像的生成带来了更多的挑战,包括:如何保证模型的稳定性,如何保留人体和服装的细节等。PersonHD的高质量数据为未来的高分辨率人体姿态迁移研究提供了坚实的基础。PersonHD提供了语义分割图和骨架关键点,对当前主流的人体姿态迁移方法都很友好。在未来,PersonHD可以扩展到如高分辨率姿态引导的人物视频生成等相关任务中。

4 结 论

本文提出一个面向姿态迁移高清人体图像生成任务的数据集 PersonHD,该数据集的主要特点包括:数据规模大、数据多样性强、图像数据分辨率高且具备高质量的多模态的辅助数据标注(人体关键点和语义分割图)。该数据集有助于对姿态引导的人物图像生成算法进行全面评估。

在提出的高清人体图像生成数据集上,本文建立了更为全面的评估基准。本文首先建立了低分辨率基准,用于对当下主流的先进姿态迁移方法进行

评估。为了进一步探索姿态迁移高分辨率人物图像的生成,本文提出了第1个高分辨率基准,并在该基准上进行了广泛的实验评估,这将为分析最新最先进的人物图像生成方法提供一个重要的平台。

为了提供一种简单而有效的高清人物图像生成方法,本文提出了一个二阶段框架来生成更高分辨率和更高质量的人体图像。首先设计了一种语义增强的部分增强方法来解决人体图像生成中具有挑战性的过拟合问题。其次,该框架引入了条件上采样模块来生成和进一步细化高分辨率图像。本文实验系统地评估了当前具有代表性的姿态引导人物图像生成方法在两个不同分辨率评估基准上的性能,并系统地验证了本文提出的框架中每个模块的有效性。实验结果表明,该框架以较低的计算成本取得了良好的效果。

结合实验与分析结果,本文方法仍存在以下不足:1)本文提出的高清人物图像生成框架是一个二阶段框架,无法端到端实现高清人物图像的姿态迁移与生成,如何设计更高质量与低复杂度的端到端单阶段高清人物图像生成方法,是值得继续探索的方向;2)针对高分辨率基准下的图像生成实验结果,本文框架在FID指标上仍有较大的提升空间。如何提升高清人物图像的生成质量,是本研究未来需要关注的方向。

参考文献(References)

- Albahar B, Lu J W, Yang J M, Shu Z X, Shechtman E and Huang J B. 2021. Pose with style: detail-preserving pose-guided image synthesis with conditional StyleGAN. *ACM Transactions on Graphics*, 40(6): #218 [DOI: 10.1145/3478513.3480559]
- Andriluka M, Pishchulin L, Gehler P and Schiele B. 2014. 2D human pose estimation: new benchmark and state of the art analysis//*Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, USA: IEEE: 3686-3693 [DOI: 10.1109/CVPR.2014.471]
- Cao Z, Simon T, Wei S E and Sheikh Y. 2017. Realtime multi-person 2D pose estimation using part affinity fields//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, USA: IEEE: 1302-1310 [DOI: 10.1109/CVPR.2017.143]
- Dong H Y, Liang X D, Shen X H, Wang B C, Lai H J, Zhu J, Hu Z T and Yin J. 2019. Towards multi-pose guided virtual try-on network//*Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*. Seoul, Korea (South): IEEE: 9025-9034 [DOI: 10.1109/ICCV.2019.00912]
- Fu J L, Li S K, Jiang Y M, Lin K Y, Qian C, C. Loy C C, Wu W and Liu Z W. 2022. StyleGAN-human: a data-centric odyssey of human generation//*Proceedings of the 17th European Conference on Computer Vision*. Tel Aviv, Israel: Springer: 1-19 [DOI: 10.1007/978-3-031-19787-1_1]
- Heusel M, Ramsauer H, Unterthiner T, Nessler B and Hochreiter S. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium//*Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, USA: Curran Associates Inc.: 6629-6640 [DOI: 10.5555/3295222.3295408]
- Isola P, Zhu J Y, Zhou T H and Efros A A. 2017. Image-to-image translation with conditional adversarial networks//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, USA: IEEE: 5967-5976 [DOI: 10.1109/CVPR.2017.632]
- Jiang Y M, Chan K C K, Wang X T, Loy C C and Liu Z W. 2021. Robust reference-based super-resolution via C^2 -matching//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville, USA: IEEE: 2103-2112 [DOI: 10.1109/CVPR46437.2021.00214]
- Karmakar A and Mishra D. 2019. A robust pose transformational GAN for pose guided person image synthesis//*Proceedings of the 7th National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics*. Hubballi, India: Springer: 89-99 [DOI: 10.1007/978-981-15-8697-2_8]
- Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J and Aila T. 2020. Analyzing and improving the image quality of styleGAN//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, USA: IEEE: 8107-8116 [DOI: 10.1109/CVPR42600.2020.00813]
- Khatun A, Denman S, Sridharan S and Fookes C. 2023. Pose-driven attention-guided image generation for person re-identification. *Pattern Recognition*, 137: #109246 [DOI: 10.1016/j.patcog.2022.109246]
- Li J, Yang J, Wang L Y and Wang Y G. 2024. Incorporating variational auto-encoder networks for text-driven generation of 3D motion human body. *Journal of Image and Graphics*, 29(5): 1434-1446 (李健, 杨钧, 王丽燕, 王永归. 2024. 融入变分自编码网络的文本生成三维运动人体. *中国图象图形学报*, 29(5): 1434-1446) [DOI: 10.11834/jig.230291]
- Li K, Zhang J S, Liu Y B, Lai Y K and Dai Q H. 2020. PoNA: pose-guided non-local attention for human pose transfer. *IEEE Transactions on Image Processing*, 29: 9584-9599 [DOI: 10.1109/TIP.2020.3029455]
- Li T J, Zhang W, Song R, Li Z H, Liu J, Li X L and S J Lu. 2021. PoT-GAN: pose transform GAN for person image synthesis. *IEEE Trans-*

- actions on Image Processing, 30: 7677-7688 [DOI: 10.1109/TIP.2021.3104183]
- Li Y N, Huang C and Loy C C. 2019. Dense intrinsic appearance flow for human pose transfer//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3688-3697 [DOI: 10.1109/CVPR.2019.00381]
- Liang X D, Gong K, Shen X H and Lin L. 2019. Look into person: joint body parsing and pose estimation network and a new benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(4): 871-885 [DOI: 10.1109/TPAMI.2018.2820063]
- Liu M C, Yan X, Wang C H and Wang K J. 2021. Segmentation mask-guided person image generation. Applied Intelligence, 51(2): 1161-1176 [DOI: 10.1007/s10489-020-01907-w]
- Liu W, Piao Z X, Min J, Luo W H, Ma L and Gao S H. 2019. Liquid warping GAN: a unified framework for human motion imitation, appearance transfer and novel view synthesis//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5903-5912 [DOI: 10.1109/ICCV.2019.00600]
- Liu W, Piao Z X, Tu Z, Luo W H, Ma L and Gao S H. 2022. Liquid warping GAN with attention: a unified framework for human image synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(9): 5115-5133 [DOI: 10.1109/TPAMI.2021.3078270]
- Liu Z W, Luo P, Qiu S, Wang X G and Tang X O. 2016. DeepFashion: powering robust clothes recognition and retrieval with rich annotations//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 1096-1104 [DOI: 10.1109/CVPR.2016.124]
- Ly Z Y, Li X M, Li X, Li F, Lin T W, He D L and Zuo W M. 2021. Learning semantic person image generation by region-adaptive normalization//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10801-10810 [DOI: 10.1109/CVPR46437.2021.01066]
- Ma L Q, Jia X, Sun Q R, Schiele B, Tuytelaars T and Van Gool L. 2017. Pose guided person image generation//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 405-415
- Ma L Y, Huang K J, Wei D X, Ming Z Y and Shen H B. 2023. FDA-GAN: flow-based dual attention GAN for human pose transfer. IEEE Transactions on Multimedia, 25: 930-941 [DOI: 10.1109/TMM.2021.3134157]
- Ma T X, Peng B, Wang W and Dong J. 2021. Must-GAN: multi-level statistics transfer for self-driven person image generation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 13617-13626 [DOI: 10.1109/CVPR46437.2021.01341]
- Men Y F, Mao Y M, Jiang Y N, Ma W Y and Lian Z H. 2020. Controllable person image synthesis with attribute-decomposed GAN//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5083-5092 [DOI: 10.1109/CVPR42600.2020.00513]
- Mirza M and Osindero S. 2014. Conditional generative adversarial nets [EB/OL]. [2024-03-01]. <https://arxiv.org/pdf/1411.1784.pdf>
- Pumarola A, Agudo A, Sanfeliu A and Moreno-Noguer F. 2018. Unsupervised person image synthesis in arbitrary poses//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8620-8628 [DOI: 10.1109/CVPR.2018.00899]
- Qian X L, Fu Y W, Xiang T, Wang W X, Qiu J, Wu Y, Jiang Y G and Xue X Y. 2018. Pose-normalized image generation for person re-identification//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 661-678 [DOI: 10.1007/978-3-030-01240-3_40]
- Ren J, Chai M L, Woodford O J, Olszewski K and Tulyakov S. 2021. Flow guided transformable bottleneck networks for motion retargeting//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10790-10800 [DOI: 10.1109/CVPR46437.2021.01065]
- Ren Y R, Yu X M, Chen J M, Li T H and Li G. 2020. Deep image spatial transformation for person image generation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7687-7696 [DOI: 10.1109/CVPR42600.2020.00771]
- Roy P, Bhattacharya S, Ghosh S and Pal U. 2023. Multi-scale attention guided pose transfer. Pattern Recognition, 137: #109315 [DOI: 10.1016/j.patcog.2023.109315]
- Siarohin A, Sangineto E, Lathuilière S and Sebe N. 2018. Deformable GANs for pose-based human image generation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3408-3416 [DOI: 10.1109/CVPR.2018.00359]
- Song S J, Zhang W, Liu J Y and Liu T. 2019. Unsupervised person image generation with semantic parsing transformation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2352-2361 [DOI: 10.1109/CVPR.2019.00246]
- Sun Y, Ding J W, Zhang Q and Deng Q Y. 2024. Images super-resolution reconstruction of transposed self-attention with local feature enhancement. Journal of Image and Graphics, 29(4): 908-921 (孙阳, 丁建伟, 张琪, 邓琪瑶. 2024. 局部特征增强的转置自注意力图像超分辨率重建. 中国图象图形学报, 29(4): 908-921) [DOI: 10.11834/jig.230320]
- Tang H, Bai S, Zhang L, Torr P H S and Sebe N. 2020. XingGAN for person image generation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 717-734 [DOI: 10.1007/978-3-030-58595-2_43]
- Tang H, Xu D, Liu G W, Wang W, Sebe N and Yan Y. 2019. Cycle in

- cycle generative adversarial networks for keypoint-guided image generation//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM: 2052-2060 [DOI: 10.1145/3343031.3350980]
- Wang T C, Liu M Y, Zhu J Y, Tao A, Kautz J and Catanzaro B. 2018. High-resolution image synthesis and semantic manipulation with conditional GANs//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8798-8807 [DOI: 10.1109/CVPR.2018.00917]
- Wang Z, Bovik A C, Sheikh H R and Simoncelli E P. 2004. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4): 600-612 [DOI: 10.1109/TIP.2003.819861]
- Xiong W, Xiong C Y, Gao Z R, Chen W Q, Zheng R H and Tian J W. 2023. Image super-resolution with channel-attention-embedded Transformer. Journal of Image and Graphics, 28(12): 3744-3757 (熊巍, 熊承义, 高志荣, 陈文旗, 郑瑞华, 田金文. 2023. 通道注意力嵌入的Transformer图像超分辨率重构. 中国图象图形学报, 28(12): 3744-3757) [DOI: 10.11834/jig.221033]
- Yang F Z, Yang H, Fu J L, Lu H T and Guo B N. 2020a. Learning texture Transformer network for image super-resolution//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5790-5799 [DOI: 10.1109/CVPR42600.2020.00583]
- Yang H, Zhang R M, Guo X B, Liu W, Zuo W M and Luo P. 2020b. Towards photo-realistic virtual try-on by adaptively generating $\leftrightarrow$ preserving image content //Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7847-7856 [DOI: 10.1109/CVPR42600.2020.00787]
- Yoon J S, Liu L J, Golyanik V, Sarkar K, Park H S and Theobalt C. 2021. Pose-guided human animation from a single image in the wild//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 15034-15043 [DOI: 10.1109/CVPR46437.2021.01479]
- Zhang J S, Li K, Lai Y K and Yang J Y. 2021. PISE: person image synthesis and editing with decoupled GAN//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 7978-7986 [DOI: 10.1109/CVPR46437.2021.00789]
- Zhang P Z, Yang L X, Lai J H and Xie X H. 2022. Exploring dual-task correlation for pose guided person image generation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 7703-7712 [DOI: 10.1109/CVPR52688.2022.00756]
- Zhang R, Isola P, Efros A A, Shechtman E and Wang O. 2018. The unreasonable effectiveness of deep features as a perceptual metric//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 586-595 [DOI: 10.1109/CVPR.2018.00068]
- Zhang Z F, Wang Z W, Lin Z and Qi H R. 2019. Image super-resolution by neural texture transfer//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7974-7983 [DOI: 10.1109/CVPR.2019.00817]
- Zheng L, Shen L Y, Tian L, Wang S J, Wang J D and Tian Q. 2015. Scalable person re-identification: a benchmark//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1116-1124 [DOI: 10.1109/ICCV.2015.133]
- Zhou X R, Zhang B, Zhang T, Zhang P, Bao J M, Chen D, Zhang Z F and Wen F. 2021. CoCosNet v2: full-resolution correspondence learning for image translation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 11460-11470 [DOI: 10.1109/CVPR46437.2021.01130]
- Zhu J Y, Park T, Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2242-2251 [DOI: 10.1109/ICCV.2017.244]
- Zhu Z, Huang T T, Shi B G, Yu M, Wang B F and Bai X. 2019. Progressive pose attention transfer for person image generation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2342-2351 [DOI: 10.1109/CVPR.2019.00245]

作者简介

徐正国,男,工程师,主要研究方向为智能信息处理。

E-mail: 657884041@qq.com

宋纯锋,通信作者,男,副研究员,主要研究方向为模式识别与计算机视觉。E-mail: chunfeng.song@nlpr.ia.ac.cn

普碧才,男,工程师,主要研究方向为自动控制技术。

E-mail: pbc_121403@163.com

秦建明,男,工程师,主要研究方向为智能信息处理。

E-mail: 379252293@qq.com

项炎平,男,高级工程师,主要研究方向为计算机视觉。

E-mail: xiangyanping@aipark.com

彭振江,男,助理工程师,主要研究方向为计算机视觉。

E-mail: zhenjiang.peng@ia.ac.cn