换装扩散模型: 两个U型网络的结构

摘要

给定两幅图片描述一个人，和一个由另一个人穿着的服装部件,我们的目标是生成一个可视化服装造型穿在输入点的人上。一个关键的挑战是合成一个真实的服装可视化，同时扭曲服装使之可以适应随主题而变化的一个重要的身体位姿和形状。以前的方法要么专注于服装细节没有有效的姿态和形状变化，要么允许想要的形状和位姿但是缺少服装细节。在这篇文章中，我提出了一个基于Diffusion的架构，统一两个UNet(被称为并行UNet),它允许我们保存衣服的细节，并在一个单一的网络中让我们可以保存服装细节并根据重要的姿态和身体扭曲服装。在并行UNet中的关键想法包括:1）服装通过交叉注意力机制隐含地弯曲。2）服装弯曲和人物混合作为一个统一过程的一部分发生，而不是两个独立的任务序列。

关键词：扩散模型；虚拟换装

1 引言

图 1. 三种方法在512×384分辨率的VITON-HD数据集上的比较结果。可以看出，我们的方法可以产生高质量的结果并确保衣服的恢复

虚拟换装的目标是基于一张人体图片和一张服装图片可视化出一件衣服穿在人身上的效果。其中一个关键的问题是服装穿在人身上会发生非刚性翘曲，但同时不能干扰到服装图案和纹理。当姿势或形体发生显著变化时，服装需要弯曲，根据新的形状或遮挡产生皱纹或变平。

本文提出了TryOnDiffusion，可以处理大型遮挡，姿势变化和体型变化，同时以1024*1024的分辨率保存服装细节.TryOnDiffusion以两张图片作为输入：一张目标图片，和另一张穿在另

一个人身上的服装图片。它会合成目标人物穿着这张图片。这件衣服可能已经被身体或其他衣服部分遮挡，需要显著地形变才能适配到目标人体上。本文方法在4百万图片对上图片，每对图片都有相同的人穿着相同的服装，但是摆出不同的姿势。

TryOnDiffusion基于我们新提出的架构名叫并行-UNet, 由两个子UNets通过交叉注意力机制结合。我们的两个关键设计元素是隐式扭曲翘曲和混合结合在在一个单个过程而不是在一个顺序换装。在目标人和源服装之间的隐式扭曲是通过多个金字塔层次的特征进行交叉注意力来实现，这样就建立了长距离的对应关系。特别是在重遮盖和极端姿势差异下，长距离对应表现良好。而且，使用相同的网络来进行翘曲和混合能使两个过程在特征水平上交换信息而不是在像素水平上，在用感知损失和风格损失时可以证明在特征水平上交换信息是必要的。

为了生成1024*1024分辨率的高质量结果，我们遵循Imagen并且创建了级联扩散模型。值得注意地是，基于扩散模型的并行U网络被用于128*128和256*256像素生成。然后，把256*256像素的结果喂到一个超分辨率扩散网络生成最终的1024*1024图片。

本文的主要贡献是1.合成1024*1024分辨率虚拟换装图片，可以在适配复杂的体态和多样的体型的同时保存服装的细节(包括图片，文本，商标等)。2.提出一个名叫并行U网络的新架构，它用跨模态注意力来隐式地翘曲服装，而且将翘曲和混合合在同一个过程中。我们定性和定量地评测了TryOnDiffusion，对比当前最好的方法，还做了一个扩展的用户调研实验。这项用户研究由15个非专家完成，对超过2K个不同的随机相比进行排名。研究表明，与最近三种最先进的方法相比，我们的结果在92.72%的时间内是最好的

2 相关工作

2.1 虚拟换装

虚拟试穿一直是一个很有吸引力的研究课题，因为它可以显著增强消费者的购物体验。根据 [1, 3]，我们可以将现有的虚拟试穿技术分为2D和3D两类。3D虚拟试穿技术可以带来更好的用户体验，但它依赖于3D参数人体模型，不幸的是，它需要构建大规模的3D数据集, 所以训练时昂贵的。与基于3D的方法相比，基于图像的虚拟试穿，即2D虚拟试穿，虽然不如3D灵活（例如，允许使用任意视图和姿势观看），但更轻，通常更普遍。

以前的许多2D虚拟试穿工作都使用了薄板样条(TPS)方法使衣服灵活变形以覆盖人体。然而，TPS只能提供简单的变形处理，只能将衣服粗略地迁移到目标区域，无法处理一些较大的几何变形。此外，已经提出了许多基于流的方法 [3]，它们对衣服和人体相应区域之间的外观流场进行建模，以更好地将衣服与人相匹配。以前的工作大多是在低分辨率条件下完成虚拟试穿任务，并取得了理想的效果。还有一些方法 [1, 5]可以在高分辨率条件下处理虚拟试穿任务，这无疑对服装的打磨和图像的合成有更高的质量要求。这些工作大多可分为两个阶段。第一阶段是早期的翘曲阶段，第二阶段是合成阶段，主要基于GANs。随着分辨率的增加，GANs生成的这些图像很难保留衣服的特征，甚至保真度也会随着更多的模糊和伪影而显著降低。

GANs的生成能力限制了先前方法的结果。即使衣服有更好的翘曲效果，当衣服与人结合在一起时，它仍然会失去很多真实感。已经证明，扩散模型能够产生高分辨率的高质量图像，并且具有更强的生成能力。在这项创新的帮助下，我们打算进一步提高虚拟试穿性能。

2.2 扩散模型

已经提出了的去噪扩散模型DDPM [4]，通过逐渐的反向去噪过程，从正态分布生成真实的图像。DDPM可以生成真实的、多样化的图像，但其采样速度慢阻碍了其广泛应用。最近，宋飏博士提出了DDIM，将采样过程转换为非马尔可夫过程，从而实现更快和确定性的采样。为了进一步降低扩散模型的计算复杂度和计算资源需求，潜在扩散模型（LDM） [8]采用了一组冻结的编码器-解码器对隐空间进行扩散和去噪处理。随着扩散模型的发展和成熟，它已经成为GANs在生成领域的强大竞争对手。

与此同时，研究人员也在探索如何更有效地控制扩散模型的生成。文生图技术可以极大地帮助用户进行富有想象力的创作。许多工作 [8]将文本信息作为去噪过程中的一个条件，以指导模型生成与文本相关的图像。ILVR和SDEdit可以通过干预去噪过程在空间层面上指导扩散模型。最近， [7, 10]提出用于更容易地将扩散模型转移到不同的任务。然而，仍然没有适用于解决虚拟换装问题的扩散模型。为了描绘衣服的各种外观，通过文本到图像的方式来完成试穿任务显然是不现实的。参考 [9]，我们可以使用修复的思想来完成任务，但这种方法不能很好地控制修复的细节。为了解决这个问题，我们将粗略的结果输入到扩散模型中进行微调，有效地指导生成的结果。此外，我们在去噪过程中引入了局部条件，这些条件与全局条件一起约束模型生成。

3 本文方法

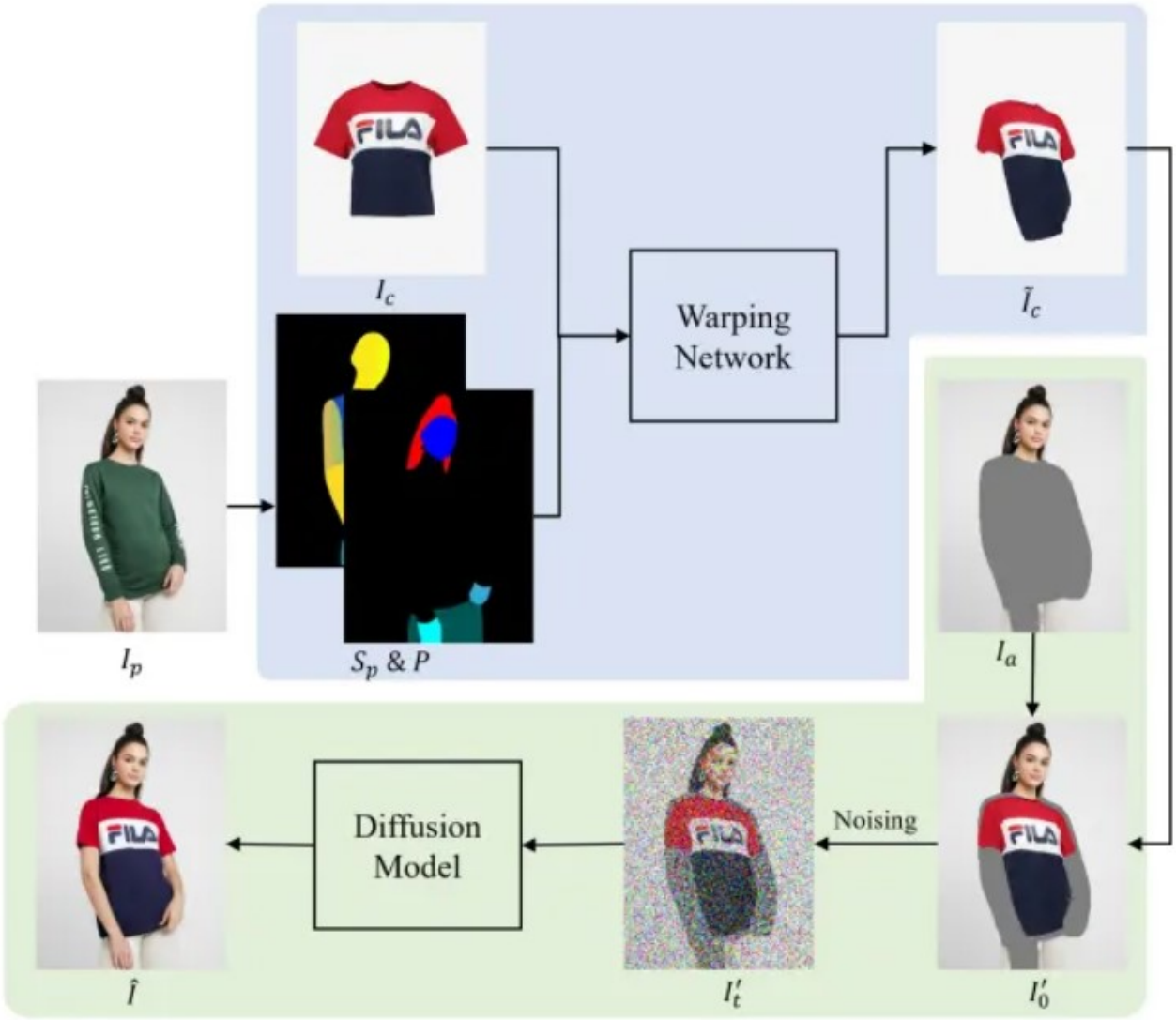

图 2. 我们的方法概述。首先通过预处理得到目标人物图像的分割结果、密度和人体无关的服装表示。衣服图像通过扭曲网络与人大致对齐。然后我们将翘曲的服装和人体无关表示结合在一起，并加以噪声作为扩散模型的输入，通过去噪的方式产生最终的输出

给定一张人体图像I_p和一张穿着服装g的另一个人图像I_g,我们的方法生成穿着服装g的人p的结果I_{tr}. 我们的方法被训练在配对的数据上训练，每对数据I_p和I_g是相同的人穿相同的服装但是具有不同的姿势。在推理期间，I_p和I_g是不同的人穿不同的衣服且具有不同的姿势的图片集。我们以描述我们的预处理步骤和一个简要的扩散模型架构图为开始。接着，我们会在子小节中描述我们的贡献和设计选择。

输入预处理。 我们首先使用现有的模型 [2, 6]预测人体和服装的解析分割图(S_p, S_g)和2D关键点(J_p, J_g)。对于服装图片，我们进一步地通过解析分割图分割出了服装图片I_c.对于人体图片，我们生成了服装无关的RGB图片I_a，I_a去除了原始的服装部分但是保存了人体特征。注意在VITON-HD描述中的服装无关RGB缺少泄露了原始服装的信息，模型对于用于具有挑战性的人体姿势和宽松服装的泛化性不足。我们因此采用了一个更具激进的方式来去除服装信息。具体来说，我们首先将前景的整个边界框区域遮盖住，然后将头部、手部和下半身复

制粘贴上去。我们使用S_p和J_p来提取非服装的人体部分。在将姿态关键点输入到网络之前我们把它们归一化到[0,1]。我们虚拟换装的条件输入可以被表示为$c_{tryon} = (I_a, J_p, I_c, J_g)$。

扩散模型的简要概括。 扩散模型是一类生成模型，它通过一系列去噪过程学习目标分布。它们由一个马尔科夫前向过程和一个可学习的逆向过程组成，前向过程将数据样本x干扰为高斯噪声z_T,逆向过程将z_T迭代地转换为x。扩散模型可以以类标签、文本或者图片作为条件信号。一个条件扩散模型$\hat{x}_\theta$可以使用加权去噪分数匹配目标进行训练

$$\mathbb{E}_{x,c,\epsilon,t}[w_t||\hat{x}_\theta(\alpha_t x + \sigma_t \epsilon, c)||_2^2] \tag{1}$$

x是目标数据样本，c是条件输入，$\epsilon \sim \mathcal{N}(0, I)$是噪声项。$\alpha_t, \sigma_t, w_t$是时间戳$t$的函数，它可以影响采样指令。在实际过程中$\hat{x}_\theta$重参数化为$\hat{\epsilon}_\theta$来预测噪声，噪声将x扰动成z，$z_t := \alpha_t x + \sigma_t \epsilon$.在推理阶段，数据样本可以从高斯噪声$z_t \sim \mathcal{N}(0, I)$使用类似DDPM或者DDIM的采样器采样生成。

3.1 用于虚拟换装的级联扩散模型

我们的级联扩散模型由一个基础扩散模型和两个超分辨率(SR)扩散模型组成。基础扩散模型是一个128×128的并行UNet。它接受条件输入的c_{tryon}预测128×128的换装结果I_{tr}^{128}.因为I_a和I_c可以是有噪声的，因为不准确的人体分割和姿态估计，我们将噪声条件增加到其中。具体地，在其他处理之前，随机高斯噪声被加到I_a和I_c, 按照[18]这种噪声增强的大小也被作为条件输入。 128×128到256×256的SR扩散模型也是参数化的256×256并行UNet。它同时以128×128的试穿结果I_{tr}^{128}和256×256分辨率的试穿输入c_{tryon}作为条件生成256×256试穿结果I_{tr}^{256}。在训练过程中，I_{tr}^{128}是标签直接下采样得到的。在测试阶段，他被设置为基础扩散模型的预测。在这个阶段，将噪声条件增强应用于所有条件输入图像，包括I_{tr}^{128},I_a和I_c. 256×256到1024×1024的超分辨率扩散模型是一个从Imagen引入的高效UNet.这个阶段是纯超分辨率网络，没有试穿条件。在训练时，从1024×1024的图片随机裁剪出256×256作为标签，输入被设置为裁剪区域下采样到64×64的图片。在推理阶段，模型以之前的并行UNet模型中的256×256试穿结果作为输入并最终合成1024×1024的结果I_{tr},为了方便这种设置，我们通过移除所有注意力层来使网络完全卷积。与前面的两个模型一样，将噪声条件增强应用于条件输入图像。

3.2 并行UNet

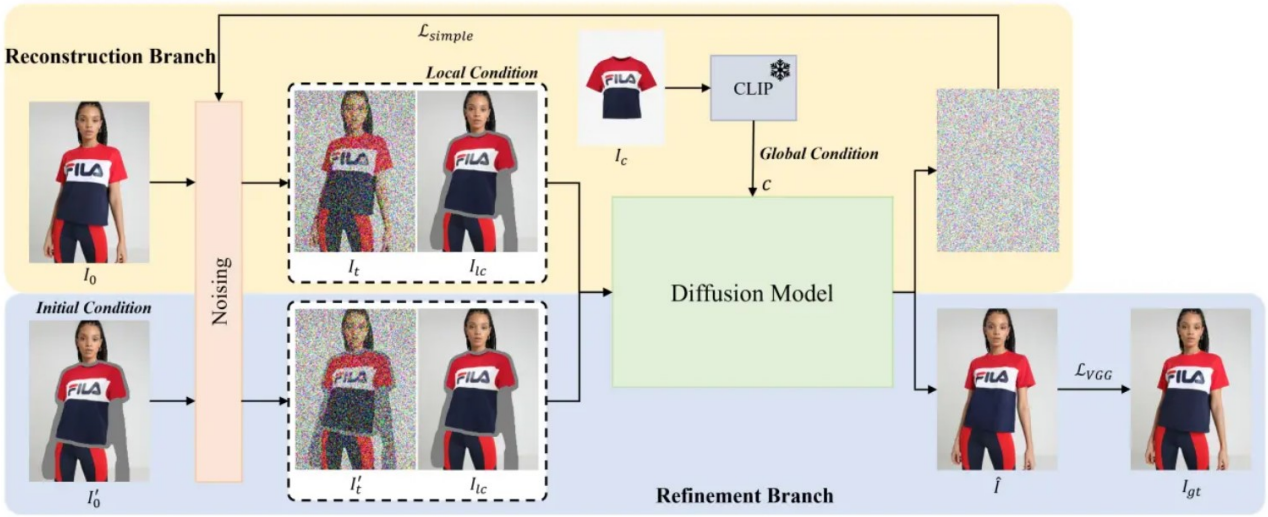

图 3. 我们的方法中的扩散模型的训练管道。我们的培训管道有两个分支：上面的构建分支和下面的细化分支。它们之间的主要区别在于输入对象和优化目标的不一致性。为了更好地可视化，我们显示了与隐空间中的变量相对应的图像。

128*128并行UNet可以表示为

$$\epsilon_t = \epsilon_\theta(z_t, t, c_{tryon}, t_{na}) \tag{2}$$

这里t是扩散时间步，z_t是标签在t时间步扰动下的噪声图片，c_{tryon}是试穿条件输入，t_{na}是加载不同条件图片噪声增强水平的集合，ϵ_t是预测的噪声可以用于从z_t中恢复出标签。256*256并行UNet以试穿结果I_{tr}^{128}作为输入，此外还将256*256分辨率的c_{tryon}作为试穿条件。接下来，我们描述并行UNet的两个关键设计元素。

隐式翘曲。第一个问题是：我们如何在神经网络中实现隐式扭曲？一个自然而然的解决方案是使用传统的UNet，然后将分割好的服装I_c和噪声图片z_t沿着通道维度进行拼接。然而，通道级别的拼接不能处理复杂的变换例如服装扭曲。这是因为传统UNet的计算原语是空间卷积和空间自注意力，并且这些原语具有很强的像素结构偏见。为了解决这个挑战，我们提出在信息流(I_c, z_t)中使用交叉注意力机制来实现隐式翘曲。这个交叉注意力基于缩放点积注意力:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{3}$$

这里$Q \in \mathbb{R}^{M\times d}$, $K \in \mathbb{R}^{N\times d}, V \in \mathbb{R}^{N\times d}$是查询、键和值的堆叠的向量，$M$是查询向量的数量，$N$是键和值向量的数量，$d$是向量的维度。在我们的例子中，查询和键值对来自不同的输入。具体来说，Q是z_t展平的特征,并且K, V是I_c展平的特征。注意力图$\frac{QK^T}{\sqrt{d}}$通过点乘计算告诉我们目标人体和源服装的相似度，对于虚拟换装提供了一种可学习的表达相关性的方式。我们也使用多头交叉注意力，使得模型可以从不同的表征子空间中学习。在一个通路中结合翘曲和混合。与之前的工作先翘曲服装再把翘曲的服装融合到目标人体不同的是，我们将两种操作结合在一个通路中。正如图3展示的那样，我们通过两个UNets分别处理服装和人体。人体UNets把服装无关的RGB图I_a和有噪声的z_t作为输入。因为I_a和z_t是像素对齐的，我们直接将它们沿着通道维度进行拼接送入UNet.

服装UNet把分割的服装图片I_c作为输入。服装特征通过上述的交叉注意力机制融合到目标图片。为了保存模型参数，我们在32*32上采样之后提前停了服装UNet，在那里人体UNet最后的交叉注意力模块已经完成了。人体姿态和服装廓形对于指导翘曲和融合的过程也是必要的。他们首先被喂入线程层分别计算位姿嵌入。位姿嵌入通过注意力机制融合到人体UNet，使用将位姿嵌入与每一个自注意力层的键值对拼接起来。而且，位姿嵌入延关键点维度使用CLIP风格的1D注意力池化，然后与扩散时间步t的位置编码以及噪声增强水平t_{na}相加。一维嵌入的结果使用FiLM在两个UNet所有尺度上进行调制特征。

4 复现细节

4.1 与已有开源代码对比

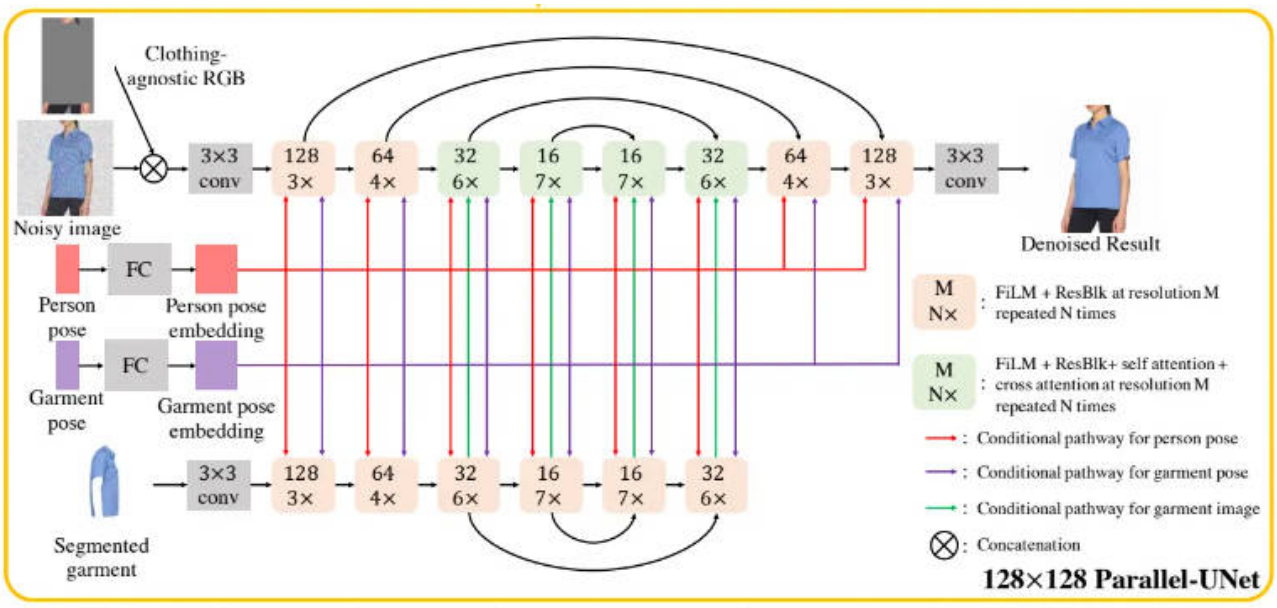

图 4. 并行UNet。

TryOnDiffusion论文没有开源代码。故借鉴DCI-VTON项目中的代码,基于此代码作为代码库借鉴TryOnDiffusion的结构进行修改。其中主要贡献量是如图4并行UNet架构的实现。其中获得Person pose embedding和Garment pose embedding是通过一个预训练好的自动编码器的编码器部分，自动编码器通过进行自监督来训练。关键是实现将分割图片UNet支路的特征通过交叉注意力机制合并到主支路UNet。

4.2 定性分析

图 5. 单个服装图片和人体图片不配对数据的测试结果，可以看出服装可以根据人的姿态、形体而发生变化。

图 6. 人体服装图片和人体图片不匹配的测试结果，可以看出服装可以在进行换装时，发生了错误的廓形改变。

定性评估通过各种方法在512×384的VITONHD数据集上产生的合成图像如图5和图6所示。尽

管之前的一些虚拟试穿方法可以正确地合成人体和衣服，但处理两者之间的互动是困难的。可以看出，我们的方法可以产生更逼真、更合理的结果，并且可以充分恢复衣服的纹理特征。图5是配对数据的测试结果，基本可以保存服装的图案和廓形，图5是非配对数据测试的结果，即虚拟换装真实的应用场景，模型可以生成逼真的服装图片，但是在试穿过程中，可能由于模型在填充掩码过程中自由发挥导致服装廓形改变，但服装的纹理细节保存良好。

4.3 定量分析

表 1. VITON-HD数据上定量实验

模型	FID↓	KID↓
Method	9.53	0.0045
HR-VITON	9.90	0.188
PF-AFN	11.30	0.283
TryOnDiffusion	13.45	6.964

将我们的方法与之前的虚拟换装方法进行对比：HR-VITON，PF-AFN，TryOnDiffusion。表1表明与这些方法的定量比较。可以看出HR-VITON具有之前方法最好的效果。结合扩散模型强大的生成能力，我们的模型在指标上可以超过HR-VITON，并且我们的模型可以生成真实自然的图像，同时最大限度地保留原始衣服。

5 总结与展望

该文件主要讨论了一种使用级联扩散模型架构进行虚拟试穿的新方法，该架构专门为高分辨率服装试穿而设计，同时保留复杂的细节并适应不同的体型和姿势。关键特征包括平行的UNet结构和交叉注意力机制，用于在统一的过程中进行隐含的服装翘曲和混合。该方法优于现有的虚拟试穿技术，尤其是在处理复杂的身体姿势和服装纹理方面。

对于未来的研究，潜在的方向可能包括提高模型在处理不同服装风格和复杂体型时的准确性和稳健性，提高试穿图像的分辨率和真实性，并探索将这种方法与实时应用程序相结合。此外，研究降低计算成本和提高模型效率的方法可能很有价值，尤其是对于面向消费者的应用程序的实际部署。

参考文献

[1] Ruili Feng, Cheng Ma, Chengji Shen, Xin Gao, Zhenjiang Liu, Xiaobo Li, Kairi Ou, Deli Zhao, and Zheng-Jun Zha. Weakly supervised high-fidelity clothing model generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3440–3449, 2022.

[2] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018.

[3] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3470–3479, 2022.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[5] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pages 204–219. Springer, 2022.

[6] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):871–885, 2018.

[7] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[9] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023.

[10] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.