文章编号：2095-2163(2023)12-0114-06 中图分类号：TP 391.43 文献标志码：A

基于光流改进的视频虚拟试衣

胡安康
(东华大学 计算机科学与技术学院，上海 201620)

摘 要：随着计算机视觉的发展，基于图像的虚拟试衣方法已经取得了很大的进展，但是，直接将此方法应用到视频虚拟试穿的任务时，由于缺乏时空一致性，会导致视频帧之间的不一致和不连贯，会对视频的视觉效果产生很大的影响，为了解决上述问题，本文提出了一种基于光流改进的视频虚拟试穿模型。首先使用正则化校正损失的薄板样条插值变换方法对服装进行扭曲，然后使用一个 U-Net 网络进行服装试穿，同时使用原视频服装区域的光流监督，合成视频服装区域的光流。实验结果表明，本文基于光流改进的视频虚拟试衣方法能很好解决虚拟试衣视频时空一致性问题。

关键词：视频虚拟试衣；光流；时空一致性；薄板样条插值变换

Video virtual try-on based on optical flow improvement

HU Ankang
(School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

Abstract: With the development of computer vision, image-based virtual try-on methods have made great progress, however, when they are directly applied to the task of video virtual try-on, the lack of spatio-temporal consistency will lead to inconsistency and incoherence between frames, and this will have a great impact on the visual effect of the video, to solve the above problems, this paper proposes a video virtual try-on model based on optical flow improvement To solve the above problems, this paper proposes a video virtual try-on model based on optical flow improvement, which first distorts the clothing using a regularization-corrected loss thin-slab sample interpolation transformation method, and then uses a U-Net network for clothing try-on while using the optical flow of the original video clothing region to supervise the optical flow of the synthetic video clothing region. The experimental results show that this paper's improved video virtual try-on method based on optical flow can well solve the virtual try-on video spatio-temporal consistency problem.

Key words: video virtual try-on; optical flow; spatio-temporal consistency; thin slab-like interpolation transform

0 引 言

随着互联网的发展，在线购物变得越来越普及，人们可以很方便的在购物网站上买到自己喜欢的服装。但是，由于图片的展示效果有限，网购服装因不能试穿而导致的退换货现象降低了用户购物体验，也在一定程度上降低了消费者的消费意愿[1]。幸运的是，这个问题可以通过视频虚拟试衣技术得到缓解。视频虚拟试穿旨在通过视频和服装合成高质量、连贯的试衣视频，用户可以通过换好服装之后的视频，全方位的观察服装的合身性和美观性。视频试衣的难点在于在保持用户身材比例和特征的同时，自然贴合地将服装扭曲到目标用户身上，同时保留服装纹理和褶皱，在每一帧图像得到较好的试穿效果的情况下，生成连贯、时空一致的虚拟试穿视频。

1 网络设计

基于上述问题，本文设计了一种端到端的视频虚拟试衣模型，可以进行视频虚拟试衣。模型由服装扭曲模块和试衣模块组成。首先是服装扭曲模块，通过改进的薄板样条插值神经网络，在目标服装和人物身上的服装之间建立对应关系，根据这种对应关系对目标服装进行扭曲；试衣模块把服装扭曲到目标人物身上的同时，利用原视频的前后两帧的光流信息；由于在换完服装之后和换完服装之前服装区域和身体区域的光流场不会产生大的变化，让原视频的每两帧之间的光流场去监督换服装之后的光流场，使生成的视频帧之间产生时空一致，解决视频中服装边缘剧烈抖动、视频不连贯的问题。

作者简介：胡安康(1996-)，男，硕士研究生，主要研究方向：计算机视觉。

收稿日期：2022-12-18

1.1　服装扭曲模块

服装扭曲就是将待试穿服装根据人的体型和姿态，扭曲成适合人物穿着的形状。薄板样条插值经常用于图像处理中的形变分析[2]，是通过估计一个带有形状上下文匹配的参数来扭曲目标服装。根据目标服装的掩码和人的服装掩码计算带有形状上下文的薄板样条插值参数，然后将目标服装根据薄板样条插值参数进行扭曲。其中，薄板样条插值的物理意义是假定给定两张图片中一些互相对应的控制点，薄板样条插值使其中一张图片进行特定的形变，使得两张图片中控制点重合，如图 1 所示。但是，此方法由于受有限的控制点限制，无法实现精细图案的变形，在处理复杂图案的服装的变形时效果不佳。

基于上述问题，本文采用改进的预测薄板样条转换的方式来扭曲服装。首先，利用目标服装信息和目标人物的外观信息之间潜在的对应关系训练网络；然后引入正则化校正损失来对薄板样条插值变换，缓解有限的控制点的影响；最后利用训练好的薄板样条插值参数对目标服装进行细致扭曲变形，达到服装扭曲自然的同时保留纹理细节。

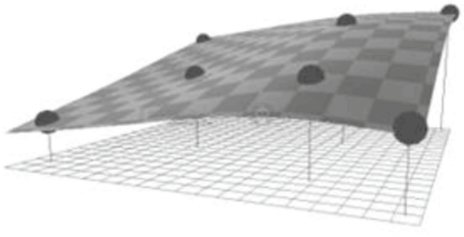

图 1　薄板样条插值变换示意图

Fig. 1　Schematic diagram of the interpolation transformation of a thin slab-like strip

服装变形模块使服装与目标人物的体型、姿势对齐，然后对服装进行自然、准确地扭曲，网络结构如图 2 所示。首先，使用两个特征提取网络，分别提取人体特征图和目标服装的特征，且并行卷积；然后，使用一个相关层，将特征提取网络提取的两个特征合并为一个张量，作为回归网络的输入；最后，通过预测空间变化参数的网络，预测出薄板样条插值参数，通过参数对服装进行扭曲。

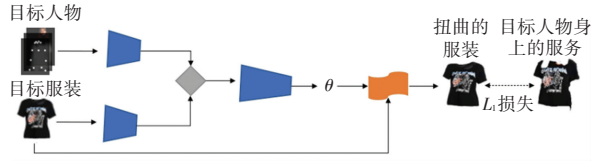

图 2　服装扭曲模块网络结构

Fig. 2　Network structure of clothing distortion module

人物特征提取网络与服装特征提取网络结构类似，唯一的区别是两个特征提取网络的输入通道数量不同，网络结构包括卷积层、下采样层和 ReLU 激活函数 3 部分。其中下采样层是 4 个步长为 2 的跨步卷积层，提取得到的特征为 f_1、f_2。提取到的两个特征图 f_1、f_2 的张量形状均用 (C,H,W) 表示，其中 C 表示图像的通道数，H 表示图像的高度，W 代表图像的宽度。在匹配层将调整 f_1 的形状为 $(C,H\times W)$，调整 f_2 的形状为 $(H\times W,C)$，然后通过矩阵相乘 $f_{12}=f_1\times f_2$，得到 f_{12} 的形状为 $(H\times W,H\times W)$，最后将张量形状调整为 $(H,W,H\times W)$。这样在得到的张量 f_{12} 中，消除了服装图片与人物图片的通道数，同时包含了服装和人的每个像素对之间的密切对应关系。在回归层中，从匹配层里得到的服装和人物体型的对应关系中，预测出 100 个薄板样条插值参数对目标服装进行扭曲。回归层的结构由卷积层、下采样层、ReLU 激活函数以及全连接层组成。其中，下采样层是步长为 2 的跨步卷积层。

1.2　基于正则化校正损失的薄板样条插值变换

由于薄板样条插值变换方法受有限的控制点限制，当服装图案复杂、色彩丰富或控制点过少时，虽然服装能按照人的体型和姿态进行扭曲，但服装的图案会扭曲的不自然。为了解决这个问题，本文引入了一个正则化校正损失，来对薄板样条插值变换。

在薄板样条插值变换控制点的上下左右各取一个点，分别为 p_0、p_1、p_2、p_3，然后利用式(1)对这 4 个点进行约束，如图 3 所示，防止内部花纹经过薄板样条插值变化后产生较大形变，从而丢失了内部信息。也就是说，使薄板点变化比较均匀，防止对服装过度扭曲。

$$L_1=\sum_{p\in P}\lambda_r\left|\ \|pp_0\|_2-\|pp_1\|_2\right|+\left|\ \|pp_2\|_2-\|pp_3\|_2\right|+\lambda_s\left(\left|S(p,p_0)-S(p,p_1)\right|+\left|S(p,p_2)-S(p,p_3)\right|\right)\tag{1}$$

式中：L_1 表示约束的正则化校正损失函数，λ_r、λ_s 表示权重超参数，$p(x,y)$ 表示一个特定的采样点，$p_0(x_0,y_0)$、$p_1(x_1,y_1)$、$p_2(x_2,y_2)$、$p_3(x_3,y_3)$ 分别表示 $p(x,y)$ 上下左右周围 4 个点，$\|pp_0\|_2$ 表示两点之间的 2-范数，用以限制扭曲距离。$S(p,p_i)$ 表示两点之间的斜率，用以限制扭曲的幅度，表达式如式(2)所示：

$$S(p,p_i)=\frac{y_i-y}{x_i-x},\ i=0,1,2,3\tag{2}$$

形变损失函数表示形变后的服装与其原服装之间的 1-范数，其公式如式(3)所示：

$$L_2 = \| C_A - C_B \|_1 \quad (3)$$

其中，L_2 表示服装形变的损失函数；C_A 是变形后的服装；C_B 是未变形的服装。

因此，最后的损失函数如式(4)所示：

$$L_W = L_1 + L_2 \quad (4)$$

其中，L_W 表示总的变形损失函数。

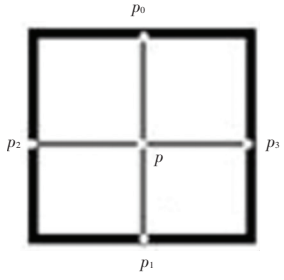

图 3 薄板样条插值示意图

Fig. 3 Schematic diagram of the interpolation of the thin slab sample strip

通过正则化校正损失，保持了薄板样条转换的共线性和不变性，从扭曲距离和扭曲幅度两个方面减少了过度扭曲的发生。使用约束函数与未使用约束函数的图像效果对比如图 4 所示，可以看到当有较多纹理的服装扭曲时，使用正则化校正的服装扭曲效果明显好于未使用正则化的效果。

(a) 目标人物 (b) 目标服装 (c) 未用正则化校正 (d) 用正则化校正

图 4 使用正则化校正与未使用效果对比图

Fig. 4 Comparison of the effect of using regularization correction and not using

1.3 基于光流的视频试穿

给定一段目标人物运动的视频序列 $I_i(i = 0,1,2,3\cdots)$，以及目标服装 c，其中 I_i 表示视频中某一帧的图像。其目标是让视频的每一帧图像穿上目标服装 c，让视频图像变为 $I_{ic}(i = 0,1,2,3\cdots)$，即目标人物穿着目标服装的视频。在视频试穿模块，不能只是简单考虑每一帧的图像，要使帧之间保持时空的一致性。通过实验发现，如果直接使用常用的穿衣模块对视频中的每一帧图像进行服装试穿，然后将所有的帧连接起来合成视频，都会出现一个问题，即当人物在运动时服装边缘会在人物身上剧烈抖动，即使人物保持静止时，这种情况也很明显，服装在人身上的图像视觉效果比较割裂。如图 5 所示，两条线的距离代表服装抖动的幅度，这种视频效果显然是无法被用户接受的，不能让用户沉浸式地体验到视频虚拟试衣。

为了解决上述问题，本文引入了光流，光流是指视频中一个物体或者单个像素在相邻帧中的位移[3]。本文在试穿模块中加入了一个光流估计模块，由于在服装换完前后，服装区域和身体区域的光流场不会产生大的变化，本文将服装区域的光流场与人物和背景区域的光流场分开，让原视频的每两帧之间服装区域的光流场去监督换服装之后的服装区域的光流场，使生成的视频帧之间产生时空一致，解决视频中服装边缘抖动，视频不连贯的问题。要做到这一点，网络不仅要学会穿衣，还要学会解释两个输入图像的外观。

图 5 视频中服装边缘抖动效果图

Fig. 5 The effect of garment edge shaking in the video

U-Net[4] 是一个全卷积神经网络，由一个编码器和一个解码器组成，编码器和解码器之间具有相同空间分辨率的跳跃式连接，其网络结构可以渲染平滑的合成图像，因此 U-Net 架构非常适合视频虚拟试戴。

视频试穿网络架构如图 6 所示。U-Net 输出的是粗略渲染的人物图像 p，然后使用前一帧的图像去更加细化渲染 p，如式(5)所示：

$$\hat{p} = \alpha * \hat{c} * m + (1 - \alpha) * p * (1 - m) \quad (5)$$

其中，$\hat{p}$ 表示人物穿着服装的图像；p 为 U-Net 输出的粗略图像；α 是权重参数，表示两个图像的贡献；$\hat{c}$ 表示按照人的形态和姿势扭曲的服装；m 表示服装位置的掩膜。

粗略渲染试衣图像 p 由当前帧的图像进行试穿得到，同时加入了光流损失，让前后两帧产生时空一致性，将服装的光流场区域与其他地方的光流场区域分开，重点关注服装区域的光流场，使换完服装之后的视频两帧之间的光流场与原视频两帧之间的光流场相似，即让生成视频的服装运动模仿原视频的服装运动。但是，对于原视频和换完服装之后的相同时刻的两张图片来说，可能出现相同位置的某一像素点，在换服装之前此像素点代表人物，换完服装之后此像素点代表服装的情况。此时，引入一个概念光流区分性 F_V，$F_{V\to 1}$ 表示此像素点的光流更接近

于服装，$F_{V\to 0}$ 表示此像素点的光流更接近于人物，$F_{V\to 0.5}$ 表示此像素点的光流取人物的光流和此区域服装光流的平均值。生成粗略试衣图像 p 过程如式(6)所示：

$$p = g(I_0, F_V) \tag{6}$$

其中，I_0 表示前一帧的图像，$g(,)$ 表示根据光流和前一帧生成图像的函数。

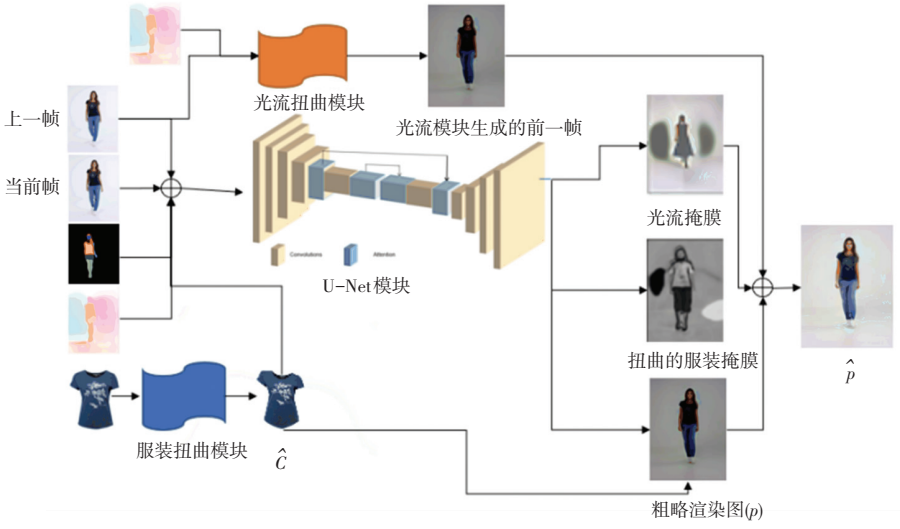

图 6　视频试穿模块网络架构图

Fig. 6　Network architecture diagram of video try-on module

在试穿任务中，本文使用较为简单的基于解析器的方法进行试穿任务。即在训练中，分割出用户穿着目标服装的区域，然后遮住此区域，用模型来训练服装和遮住服装区域的图片，合成试穿服装图像，这样就可以计算出真实图像和合成图像之间的损失。

最后的损失函数是试穿图像的损失加上光流场的损失，如式(7)所示：

$$L = L_{L1} + L_{mask} + L_{VGG} + L_{flow} \tag{7}$$

本文使用了几种常见的损失函数 L_{L1}、L_{mask}、L_{VGG}。如 Pix2Pix[5] 所说，L_{L1} 损失是实现图像翻译必不可少的，也是许多虚拟试穿常用的损失函数。视频虚拟试穿也需要保持服装图案、设计、纹理等高级结构；L_{mask} 可以保留布料的特征细节，并与目标人物的身材保持一致，L_{VGG} 也是解决这个问题的关键组件。L 代表总的损失函数，L_{flow} 代表光流场的损失函数，其表达式如式 8 所示：

$$L_{flow} = \frac{1}{N}\sum_{i=1}^{N} \| F_{o_i} - F_{c_i} \|_1 \tag{8}$$

其中，F_{o_i} 表示原视频的第 $i-1$ 帧和第 i 帧之间的光流场，F_{c_i} 表示合成视频的第 $i-1$ 帧和第 i 帧之间的光流场。

2　实验评估

2.1　实验数据及设置

本文使用 Dong[6] 等人提供的虚拟试穿视频数据集 VVT 进行虚拟试穿任务，数据集包含 791 个时装模特走秀视频。其中，训练集包含 661 个视频，测试集包含 130 个视频。

本文首先训练服装扭曲模块，使用 $\beta_1 = 0.8$、$\beta_2 = 0.899$ 的 Adam 优化器，每批次训练样本为 16，学习率为0.000 02，共进行 100 轮次迭代。当迭代至 50 轮时，使用学习率逐步衰减的方式进行训练。在训练视频试穿模块时，需要将第一阶段生成的服装一起输入到网络中，使用 $\beta_1 = 0.5$，$\beta_2 = 0.999$ 的 Adam 优化器，每批次训练样本为 32，学习率为 0.000 01，共进行 400 轮次迭代。当迭代至 200 轮时，使用学习率逐步衰减的方式进行训练。

2.2　实验结果与分析

实验从图像和视频两方面来评估本文方法在虚拟视频试衣上的表现。初始分数(Inception Score，IS)[7] 从图片的清晰度和多样性来衡量生成图片的质量，Fréchet 起始距离(Fréchet Inception Distance，FID)[8] 用来衡量生成图片和真实图片的特征相似；结构相似性指数测量(Structure Similarity Index Measure，SSIM)[8] 为结构相似性评价，从图片的亮度、对比度和结构 3 个关键特征来考量，更接近人眼的直观感受；学习感知图像块相似度(Learned Perceptual Image Patch Similarity，LPIPS)通过计算真实图像和生成图像的感知图像相似度，来衡量生成图像的质量。视频多方法评估融合(Video Multimethod

Assessment Fusion, VMAF)[9]是一种全参考的评价方式,将视频的空间域特征和时间域特征使用机器学习的支持向量机方法融合为一个指标,从视频的时间一致性和空间一致性来衡量视频质量。

本文选择特征保留试穿模型(CP-VTON[10])、光流导航翘曲模型(FW-GAN[6])、实用虚拟试衣模型(ShineOn[11])作为基准方法进行定量对比。其中,CP-VTON 学习薄板插值参数,然后对服装进行变形,最后通过一个新的几何匹配模块,将目标服装变换为适合目标人的体形。FW-GAN 用于学习基于人物图像、所需服装图像和一系列目标姿势合成虚拟试穿视频,分别对帧和服装图像进行扭曲,在全局视图和局部视图下保留细节。ShineOn 建立了自下而上的方式来处理试穿任务,同时建立了一系列的科学实验来对比虚拟视频试穿网络中的不同模型组合。

本文方法与其它模型的定量对比结果见表 1。从表中数据可以看出,本文方法在虚拟试穿任务中取得了优异的成绩。与 CP-VTON 相比,本文方法在引入了正则化校正损失,除 IS 指标外,都取得了更好的结果;与 FW-GAN 相比,本文方法在 FID 指标和 SSIM 指标上都有了大幅度的提升,表明本文生成的虚拟试穿图片更接近人眼的直观感受;与 ShineOn 相比,所有的评价指标都有大幅度的提升,由于本文引入了人体关键点解析,能更好地保留人物关键点特征,无论是在人眼视觉上,还是在图片清晰度上,本文的图像质量更好。且在 VMAF 上也取得了较好的成绩,证明了本文方法能够通过光流很好地解决视频时空一致性问题,生成连贯、清晰的视频。

表 1 虚拟试衣图像定量结果指标对比表

Table 1 Comparison table of quantitative results indicators of virtual fitting images

	IS ↑	FID ↓	SSIM ↑	LPIPS ↓	VMAF ↑
特征保留试穿模型	**4.68**	37.8	0.78	0.53	0.56
光流导航翘曲模型	4.07	30.6	0.69	**0.31**	**0.85**
实用虚拟试衣模型	3.17	41.7	0.63	0.45	0.65
本文算法	4.23	**26.2**	**0.81**	0.36	0.72

总体来看,本文方法在 FID 和 SSIM 指标上都取得了最好的分数,在 IS、LPIPS 和 VMAF 指标上均取得了较好的分数。结果表明,本文方法能够产生更为清晰的图片,在保留整体特征上效果更好,生成的视频时空一致性更好。因此,从定量方面证明了本文方法的有效性。

此外,也将本文所提模型与 CP-VTON、FW-GAN、ShineOn 等进行了定性比较,其中部分帧的图像效果对比如图 7 所示。

图 7 视频中部分帧效果对比

Fig. 7 Comparison of the effect of some frames in the video

本文方法在处理较为复杂的纹理的服装时,由于加入了人体特征解析信息和人体关键点信息,服装能够很好地按照人物的身材和姿势进行对齐。在试穿服装时,也能够完全保留人物的各部分特征信息,较好地解决了 FW-GAN 的面部区域图像模糊和 ShineOn 的脖子“消失”问题。此外,在扭曲试衣时,

使用正则化校正损失对薄板样条插值，也能较好地保留服装花纹等细节信息。定性结果验证了本文方法在视频虚拟试衣上的有效性。

下面将从视频的连贯性方面来对虚拟试衣视频进行定性分析。通过实验发现，若直接将基于图像的虚拟试衣技术运用到视频上时，服装的边缘会出现很剧烈的抖动，即使在人物静止时，也会出现抖动的情况。从视觉效果上看，人物和服装具有很强的割裂感，是不可接受的，连贯性主要从服装边缘在人物身上的抖动进行分析。

为了更加直观地进行定性比较，本文使用基于蒸馏外观流的无解析器虚拟试衣模型（PF-AFN[12]）和基于视频的虚拟试衣方法 ShineOn 两种模型作为对比方法。如图 8 所示，实验中每隔 10 帧取一张图片展示，图中两条线之间的距离代表衣服下摆边缘抖动的幅度。从中可以看出，在基于图像的虚拟试衣方法 PF-AFN 中，当人物在视频中运动时，由于帧之间没有时空一致性，服装的下摆会产生剧烈的抖动。

(a) PF-AFN模型

(b) ShineOn模型

(c) 本文算法

图 8　虚拟试衣视频对比

Fig. 8　Virtual try-on video comparison

ShineOn 使用光流来扭曲前一帧的图像，然后利用前一帧的图像来改善当前帧的图像，在一定程度上增加了视频的时间一致性，改善了视频的连贯性，但是出现了部分帧人物脖子“消失”的情况。本文的方法与 ShineOn 相比，可以看出本文方法更好地保留人物的特征信息，没有出现人物的脖子等位置消失情况。同时，本文利用光流更好地使视频具有时空一致性，生成的视频更加连贯，服装在人物身上更加自然，很好地解决了服装下摆及边缘在视频中剧烈抖动的问题。

定量和定性结果说明了本文方法在视频虚拟试衣上的有效性，可以生成连贯、时空一致的虚拟试衣视频。

3　结束语

本文提出的视频虚拟试衣模型，解决了视频虚拟试衣中帧之间的时空一致性问题。实验结果表明，本文方法生成的虚拟试衣视频在图像质量和视频质量上都取得了较好的结果，证明了本文方法能生成清晰、时空一致的虚拟试衣视频。

参考文献

[1] 王浩然，赵永琨，陈琦. 浅析虚拟试衣的发展现状、技术原理及前景展望[J]. 中国信息化，2022(7)：88-89.

[2] 吴敏，郭田德，韩丛英. 带方向信息的薄板样条插值函数及其应用[J]. 应用数学学报，2021，44(5)：659-677.

[3] HORN B K P, SCHUNCK B G. Determining optical flow[J]. Artificial Intelligence, 1981, 17(1-3): 185-203.

[4] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2015: 234-241.

[5] HAN X, WU Z, WU Z, et al. Viton: An image-based virtual try-on network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7543-7552.

[6] DONG H, LIANG X, SHEN X, et al. Fw-gan: Flow-navigated warping gan for video virtual try-on[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1161-1170.

[7] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv: 1412.6980, 2014.

[8] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[C]// Advances in Neural Information Processing Systems, 2017: 30.

[9] 卓力，杨硕，张菁，等. 基于多模双线性池化和时间池化聚合的无参考 VMAF 视频质量评价模型[J]. 北京工业大学学报，2022，48(7)：721-728.

[10] WANG B, ZHENG H, LIANG X, et al. Toward characteristic-preserving image-based virtual try-on network[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 589-604.

[11] KUPPA G, JONG A, LIU X, et al. ShineOn: Illuminating design choices for practical video-based virtual clothing try-on[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 191-200.

[12] GE Y, SONG Y, ZHANG R, et al. Parser-free virtual try-on via distilling appearance flows[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8485-8493.