Single Stage Virtual Try-on via Deformable Attention Flows

Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang

DAMO Academy, Alibaba Group, China
{baishuai.bs,zhule.zhl,zhikang.lzk,ericzhou.zc,yang.yhx} @alibaba-inc.com

Abstract. Virtual try-on aims to generate a photo-realistic fitting result given an in-shop garment and a reference person image. Existing methods usually build up multi-stage frameworks to deal with clothes warping and body blending respectively, or rely heavily on intermediate parser-based labels which may be noisy or even inaccurate. To solve the above challenges, we propose a single-stage try-on framework by developing a novel Deformable Attention Flow (DAFlow), which applies the deformable attention scheme to multi-flow estimation. With pose keypoints as the guidance only, the self- and cross-deformable attention flows are estimated for the reference person and the garment images, respectively. By sampling multiple flow fields, the feature-level and pixel-level information from different semantic areas is simultaneously extracted and merged through the attention mechanism. It enables clothes warping and body synthesizing at the same time which leads to photo-realistic results in an end-to-end manner. Extensive experiments on two try-on datasets demonstrate that our proposed method achieves state-of-the-art performance both qualitatively and quantitatively. Furthermore, additional experiments on the other two image editing tasks illustrate the versatility of our method for multi-view synthesis and image animation. Code will be made available at https://github.com/OFA-Sys/DAFlow.

Keywords: Virtual try-on, Single stage, Deformable attention flows

1 Introduction

Virtual try-on aims to generate a photo-realistic and reasonable try-on result based on an in-shop garment and a reference person image. In recent years, due to its potential applications in the fashion and e-commerce industries, it has received more and more attention. Recent methods [10,43,4,9,5] have achieved considerable progress in generating realistic results and preserving details. However, this task is still challenging, especially under complex poses and large deformations, where most of existing methods are still suffering from misalignment or obvious artifacts.

Most prior try-on systems adopt a multi-stage approach [15,41,43,9] shown in Fig. 1, including clothes warping, structure estimation, and image synthesis. Clothes warping is to align the garment to the target pose while preserving

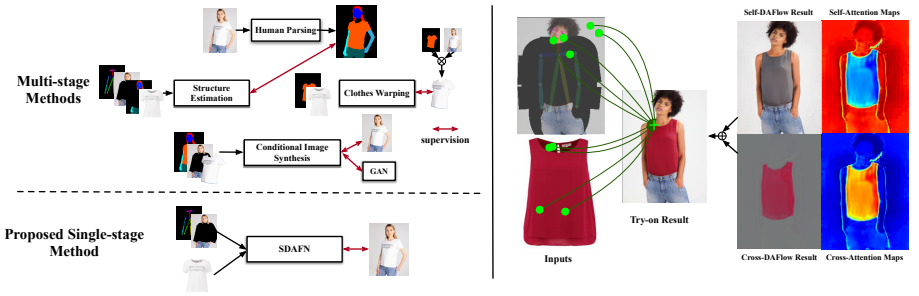

Fig. 1. The comparison between multi-stage methods and our single-stage approach. With multiple flow fields sampled, different regions are associated. Attention maps learn the 3D priors to make the try-on more realistic without 3D information input.

the texture details. Structure estimation predicts the segmentation map of the human body to guide the image synthesis. Given the warped clothing and the intermediate semantic labels, image synthesis is performed as a conditional generation task for pixel-level refinement.

Clothes warping in recent works [14,43,10,5] trains an extra flow network to warp clothes. Such flow operation retains the realistic texture but usually predicts the inaccurate structure. Some approaches [14,10] limit the unsmooth deformation of flow by introducing additional constraints but fail in dealing with complex poses. To further improve the quality of the generated results, some works [43,5] introduce more prior knowledge from some external pre-trained models, such as human parsing [11], densepose [13], and 3D depth. However, the inaccurate intermediate predictions may lead to results with noticeable artifacts. In addition, although the generative adversarial networks (GAN) [12] help preserve the sharp details of the generated images, they modify some attributes of clothes, such as color or style, which are not desirable for virtual try-on.

To address the above problems, we propose a Single-stage Deformable Attention Flow Network (SDAFN) to perform an end-to-end try-on task. we build a Deformable Attention Flow (DAFlow) module by applying a deformable attention scheme to multi-flow estimation. As shown in Fig. 1, it estimates multiple flow fields from different semantic areas and then synchronously merges the feature-level and pixel-level information with the attention mechanism in cascade. This allows the extract deformable attention flows to not only warp clothes but also synthesize photo-realistic human torsos and shadows at the same time. Then, we combine self- and cross-DAFlows to deal with the human body and the garment respectively, generating the fitting results in one single pass. In addition, our method only need pose keypoints as guidance. To our best knowledge, we are the first one-stage pure flow-based virtual try-on method.

The main contributions of this paper can be summarized as follows:

– We propose a single-stage virtual try-on framework, which applies self- and cross-DAFlows to deal with the reference person and garment images in parallel and generate realistic fitting results in an end-to-end manner.

– We propose a novel deformable attention flow module to estimate the reasonable structure while retaining the vivid texture in cascade. It works well even with a large misalignment between the clothes and the person and can be applied to a variety of resolutions.
– Extensive experiments show that our proposed method not only achieves superior performance on virtual try-on, but also can be extended to other image editing tasks, such as multi-view synthesis and image animation.

2 Related Work

2.1 Virtual Try-on.

The study on the virtual try-on task mainly consists of 3D model-based approaches [2,30,1,23] and 2D image-based approaches [15,41,24,45]. Recently, 2D methods have attracted more and more attention because of highly accessible data and photo-realistic results. VITON [15] uses a two-stage coarse-to-fine strategy to generate a clothed person. It first estimates the rough human body shape, then warps clothes and refines the details of the clothed person according to the shape. To improve the accuracy, SwapNet [32] and VTNFP [45] adopt semantic segmentation as guidance. ACGPN [43] introduces an additional stage to predict the semantic layout of the reference image. DCTON [9] and ZFlow [5] add more accurate descriptor, like densepose [13] or UV [8] projection. VITON-HD [4] improves the performance of the conditional GAN structure on high-resolution images. Although the generated images get better quality, the pipeline is becoming more complex and relies on more external information. It results in a reduction in efficiency and obvious artifacts caused by the inaccurate intermediate labels. Recently WUTON [19] and PFAFN [10] adopt the "teacher-tutor-student" scheme to get rid of the extra information at inference time and achieve good performance. This proves that the network has the ability to perform try-on without heavy dependency. Our proposed method is able to obtain realistic fitting performance with only pose information as guidance.

2.2 Spatial Transform Module.

The spatial transform module [20,47,34,38,44] is widely applied in optical flow estimation, image transformation and object recognition tasks. In virtual try-on, the spatial transform module is mainly used to wrap clothes to match the posture of the person. VITON [15] exploits a Thin-Plate Spline (TPS) [7] based warping method to deform the in-shop clothes to the refined result with a composition mask. CP-VTON [41] uses a neural network to learn the transformation parameters of TPS warping rather than using image descriptors. ClothFlow [14] introduces denser flow predictions through a cascade scheme, which makes deformation has a high dimension of freedom. However, dense flows usually present unappealing artifacts. To avoid this problem, ClothFlow uses flow variation regularization to enforce smoothness. PFAPN [10] adds a second-order smooth constraint to encourage the co-linearity of neighboring appearance flows. These constraints are proposed to make the garment smoother after deformation, but they

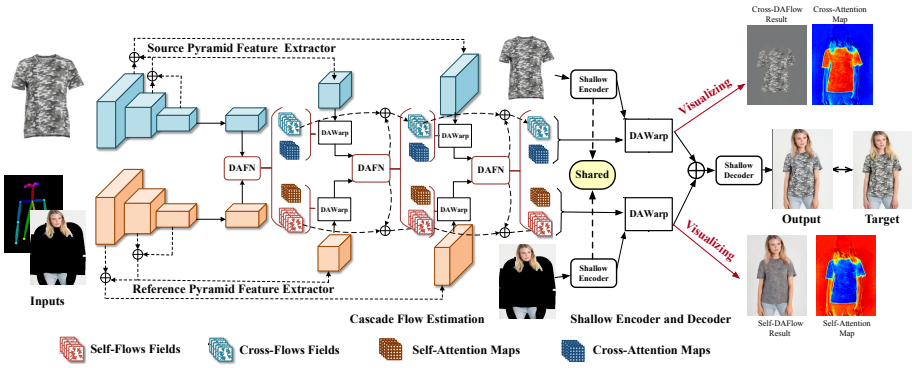

Fig. 2. The overall framework of our SDAFN. The garment, person and pose images are firstly fed into the unshared pyramid feature extractors. Then, both self- and cross-deformable attention flows are estimated in cascade. The final try-on result is obtained by applying shallow encoder and decoder together with the learned flows.

still face the challenge when there exists a huge discrepancy between the target clothes and the original clothes on the model.

2.3 Efficient Attention Mechanism.

The attention mechanism [40] is widely used in transformer models, benefiting from its long-distance association ability, and achieves good performance on image segmentation, object detection, and image generation. However, with the increase in feature resolution, dense attention will make a huge computational cost. This makes the attention mechanism usually used in low-resolution features [33]. Recently sparse attention has been introduced into detection and recognition tasks. Common methods include pre-defined local attention patterns [26,27,31] or learned sparse attention [48,42,22,39]. Combing the effectiveness of flow operation on preserving details and the capability of attention mechanism on estimating accurate structures, our proposed method naturally extend learned sparse attention to pixel-level image transformation.

3 Our Approach

In this section, we introduce the overall framework of our proposed single-stage virtual try-on model called SDAFN. Then, we describe the core DAFlow module as well as the training loss design in detail.

3.1 Architecture

Different from the previous methods who apply the clothes warping and the conditional generation in different stages, we use self- and cross-deformable attention flows to obtain the try-on results directly. In addition, our framework

only relies on the pose keypoints as guidance. Specifically, as shown in Fig. 2, our model is composed of three parts: pyramid feature extraction, cascade flow estimation, and shallow encoder-decoder generation.

Pyramid feature extraction. Our model has two pyramid feature extractors, including a reference branch and a source branch. The reference branch takes the concatenation of the person and pose images as input, where the upper body of the person is masked. The garment image is fed into the source branch. The two feature extractors have the same Feature Pyramid Network (FPN) [25] structure with unshared parameters. The FPN network consists of N encoding layers where each layer has a downsampling convolution with a stride of 2 followed by two residual blocks [16].

Cascade flow estimation. It is difficult to directly and accurately predict large deformations, especially for the virtual try-on application, where the target and source images are not domain-unified. Hence, similar to the recent methods [14,10], we adopt the cascade flow estimation from coarse to fine.

Given the hierarchical reference and source features $\{\boldsymbol{x}_r^n, \boldsymbol{x}_s^n\}_{n=1}^N$ extracted by the pyramid feature extractors, as illustrated in Fig. 2, two types of flows and attention maps are estimated. The first is the self-flow fields and self-attention maps $(\boldsymbol{o}_{daf,r}^n, \boldsymbol{a}_r^n)$ from the reference branch. The second is the cross-flow fields and cross-attention maps $(\boldsymbol{o}_{daf,s}^n, \boldsymbol{a}_s^n)$ from the interaction of the both branches. The features $(\boldsymbol{x}_r^1, \boldsymbol{x}_s^1)$ at the lowest resolution are fed into the proposed Deformable Attention Flow Network (DAFN) to predict the initial flow fields $(\boldsymbol{o}_{daf,r}^1, \boldsymbol{o}_{daf,s}^1)$ and attention maps $(\boldsymbol{a}_r^1, \boldsymbol{a}_s^1)$. Then they will be refined and updated in cascade. Specifically, the reference and source feature maps $(\boldsymbol{x}_r^n, \boldsymbol{x}_s^n), n \in [2, N]$ are first transformed by the self- and cross-DAFlows $(\boldsymbol{o}_{daf,r}^{n-1}, \boldsymbol{o}_{daf,s}^{n-1}, \boldsymbol{a}_r^{n-1}, \boldsymbol{a}_s^{n-1})$ of the previous scale, called Deformable Attention Warping (DAWarp). Then, the transformed features $(\hat{\boldsymbol{x}}_r^n, \hat{\boldsymbol{x}}_s^n)$ are applied to predict the residual flows and the new attention maps in which finer flows are obtained. The above process continues until $n = N$. In our implementation, N is set to 5, and the process shown in Fig. 2 is illustrated with $N = 3$. Both DAFN and DAWarp modules are described in Sec. 3.2 in details.

Shallow encoder-decoder generation. The shallow encoder projects the images from RGB to high dimensional space. With the final estimated flows and attention maps, the reference person and garment images in high dimensional feature space are transformed and merged with DAWarp. Then the merged result is fed into a shallow decoder to obtain the final try-on result. The shallow encoder and decoder both have two convolution layers without downsampling. In particular, both garment and reference person images share the same encoder and decoder. It is found in our experiment that the shallow encoder-decoder structure is a simple and effective way to enhance the representation capability at the pixel level and improve the flow generation quality.

3.2 Deformable Attention Flows

Revisiting the conventional flow field. The flow field estimation [14,5,10] is extensively used in the cloth warping of virtual try-on. It takes the reference (e.g.,

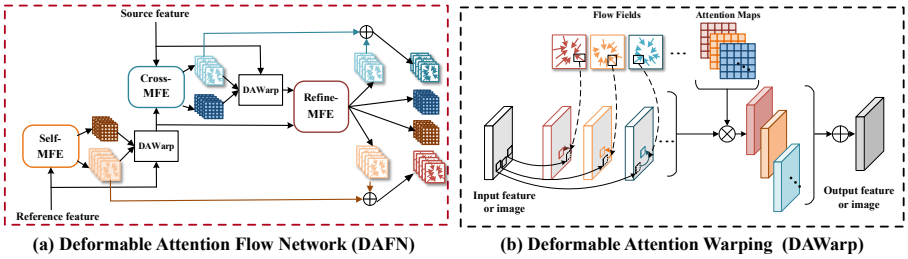

Fig. 3. Illustrations of DAFN network and DAWarp operation.

semantic segmentation, pose keypoints or depth map) feature $\boldsymbol{x}_r \in \mathbb{R}^{C\times H\times W}$ and the source (e.g., clothes) feature $\boldsymbol{x_s} \in \mathbb{R}^{C\times H\times W}$ as inputs, and only estimates one 2D coordinate offset for each sampling position. C is the feature dimension, H and W denote the height and width of feature map. This process of predicting offset map $\boldsymbol{o} \in \mathbb{R}^{2\times H\times W}$ can be written as:

$$\boldsymbol{o} = F([\boldsymbol{x}_r, \boldsymbol{x}_s]), \tag{1}$$

where F indicates flow field estimator network. Accordingly, the warped feature at a 2D reference location p is calculated as :

$$\mathcal{W}(\boldsymbol{x}_s, \boldsymbol{o})_p = \boldsymbol{x}_s(p + \boldsymbol{o}_p), \tag{2}$$

where $\mathcal{W}(\cdot)$ denotes the warp operation, $p+\boldsymbol{o}_p$ is the estimated sampling position. The value $\boldsymbol{x}_s(p+\boldsymbol{o}_p)$ of each sampling point is computed by bilinear interpolation from the nearby grid points on the feature map.

The flow operation directly samples from images, retaining the realistic textures. However, structural information is often associated with multiple locations. For example, the shape of the clothes on the human body is determined by the posture, body shape, and the type of garment. Hence, only one flow field cannot estimate accurate structure, so most existing methods only apply the flow to warp the garment in virtual try-on. To mitigate this weakness, we propose a deformable attention flow (DAFlow) module, which preserves detailed textures and accurately estimates structures. The DAFlow module consists of the Multiple Flow field Estimator (MFE) and the DAWarp operation.

Multiple flow field estimator. Inspired by deformable attention [48], our MFE predicts the fixed number K of sampling points, regardless of the spatial size of the feature maps. In contrast to the conventional single flow field $\boldsymbol{o} \in \mathbb{R}^{2\times H\times W}$, our MFE forecasts multiple flow fields $\boldsymbol{o}_{daf} \in \mathbb{R}^{2K\times H\times W}$ and attention weight maps $\boldsymbol{a} \in \mathbb{R}^{K\times H\times W}$:

$$\boldsymbol{o}_{daf}, \boldsymbol{a} = F([\boldsymbol{x}_r, \boldsymbol{x}_s]), \tag{3}$$

where K is the sampled key number. In the implementation, both $\boldsymbol{o}_{daf}$ and $\boldsymbol{a}$ are obtained via the same ConvNet but different channels.

Deformable attention warping. As shown in the **(b)** of Fig. 3, each location is associated with features from multiple locations. Besides, DAFlows learn a variety of possible flows. The multiple flow fields and the attention operation are simultaneously applied to the features and images. In terms of the feature, the source or reference features $\boldsymbol{x} \in \mathbb{R}^{C\times H\times W}$ are effectively integrated into the desired target position as:

$$\mathcal{W}_{DAF}(\boldsymbol{x}, \boldsymbol{o}_{daf}, \boldsymbol{a})_p = \sum_{k=1}^{K} \frac{exp(\boldsymbol{a}_{pk})}{\sum_{i=1}^{K} exp(\boldsymbol{a}_{pi})} \boldsymbol{x}(p + \boldsymbol{o}_{daf,pk}), \tag{4}$$

where $(p + \boldsymbol{o}_{daf,pk})$ is the k_{th} sampling position for the reference location p, and $\mathcal{W}_{DAF}(\boldsymbol{x}_s, \boldsymbol{o}_{daf}, \boldsymbol{a})$ denotes the warped feature with DAFlows. As for the image level, merging multiple warped images has more generation possibilities and is recombined into new images with reasonable structures and realistic textures. As illustrated in Fig. 7, the neck and arm parts are realistically generated, which are not in the original image. Additionally, the warping operation applies bilinear interpolation, enabling the estimated offsets optimized with back-propagation.

3.3 Deformable Attention Flow Network.

As illustrated in Sec. 3.1, we adopt cascade flow estimation to predict the deformation and attention from coarse to fine. Given the reference and source feature maps $(\boldsymbol{x}_r^n, \boldsymbol{x}_s^n), n \in [2, N]$, the feature maps are transformed to $(\hat{\boldsymbol{x}}_r^n, \hat{\boldsymbol{x}}_s^n)$ in accordance with the flows $(\boldsymbol{o}_{daf,r}^{n-1}, \boldsymbol{o}_{daf,s}^{n-1})$ and attention $(\boldsymbol{a}_r^{n-1}, \boldsymbol{a}_s^{n-1})$ from the previous scale:

$$\hat{\boldsymbol{x}}_{s/r}^n = \mathcal{W}_{DAF}(\boldsymbol{x}_{s/r}^n, \mathcal{U}(\boldsymbol{o}_{daf,s/r}^{n-1}), \mathcal{U}(\boldsymbol{a}_{s/r}^{n-1})), \tag{5}$$

where $\mathcal{U}(\cdot)$ denotes the bilinear samping with the scale factor of 2.

For each scale, DAFN is utilized to estimate the DAFlows. The **(a)** of Fig. 3 illustrates that the DAFN consists of three flow estimators (F_r^n, F_s^n, F_m^n), namely, Self-MFE, Cross-MFE and Refine-MFE. Firstly, the transformed reference feature $\hat{\boldsymbol{x}}_r^n$ is fed into F_r^n to estimate the initial residual self-flows and the self-attention maps $(\Delta\dot{\boldsymbol{o}}_{daf,r}^n, \dot{\boldsymbol{a}}_r^n)$:

$$\Delta\dot{\boldsymbol{o}}_r^n, \dot{\boldsymbol{a}}_r^n = F_r^n(\hat{\boldsymbol{x}}_r^n). \tag{6}$$

Secondly, the reference feature is transformed with $(\Delta\dot{\boldsymbol{o}}_r^n, \dot{\boldsymbol{a}}_r^n)$, and concatenated with source feature $\hat{\boldsymbol{x}}_s^n$ to forecast the residual cross-flow fields and cross-attention maps $(\Delta\dot{\boldsymbol{o}}_{daf,s}^n, \dot{\boldsymbol{a}}_s^n)$:

$$\Delta\dot{\boldsymbol{o}}_{daf,s}^n, \dot{\boldsymbol{a}}_s^n = F_s^n([\hat{\boldsymbol{x}}_s^n, \mathcal{W}_{DAF}(\hat{\boldsymbol{x}}_r^n, \Delta\dot{\boldsymbol{o}}_r^n, \dot{\boldsymbol{a}}_r^n)]). \tag{7}$$

Finally, the source feature is converted with $(\Delta\dot{\boldsymbol{o}}_{daf,s}^n, \dot{\boldsymbol{a}}_s^n)$, and combined with the newly transformed reference features to predict the refinements $(\Delta\ddot{\boldsymbol{o}}_{daf,r}^n, \Delta\ddot{\boldsymbol{o}}_{daf,s}^n)$ and the final attention maps $(\boldsymbol{a}_r^n, \boldsymbol{a}_s^n)$:

$$\Delta\ddot{\boldsymbol{o}}_{daf,s}^n, \Delta\ddot{\boldsymbol{o}}_{daf,r}^n, \boldsymbol{a}_s^n, \boldsymbol{a}_r^n = F_m^n([\mathcal{W}_{DAF}(\hat{\boldsymbol{x}}_s^n, \Delta\dot{\boldsymbol{o}}_s^n, \dot{\boldsymbol{a}}_s^n), \mathcal{W}_{DAF}(\hat{\boldsymbol{x}}_r^n, \Delta\dot{\boldsymbol{o}}_r^n, \dot{\boldsymbol{a}}_r^n)]). \tag{8}$$

The final self- and cross-flow fields at the n_{th} scale are:

$$\boldsymbol{o}_{daf,r}^{n} = \mathcal{U}(\boldsymbol{o}_{daf,r}^{n-1}) + \Delta\dot{\boldsymbol{o}}_{daf,r}^{n} + \Delta\ddot{\boldsymbol{o}}_{daf,r}^{n}, \tag{9}$$

$$\boldsymbol{o}_{daf,s}^{n} = \mathcal{U}(\boldsymbol{o}_{daf,s}^{n-1}) + \Delta\dot{\boldsymbol{o}}_{daf,s}^{n} + \Delta\ddot{\boldsymbol{o}}_{daf,s}^{n}, \tag{10}$$

and the outputs $\boldsymbol{a}_s^n, \boldsymbol{a}_r^n$ of Refine-MFE are applied as the final attention maps.

3.4 Train Losses

Discarding the dependence on intermediate parser-based labels, our model only uses the try-on result as supervision in an end-to-end manner. As for the loss functions, without any constraints or regularization on the flows, the model is directly optimized by comparing the similarity between generated results and ground truths in all scales. The similarity functions consist of the L1 loss, the perceptual loss [21], and the style loss. The L1 loss is formulated as:

$$\mathcal{L}_{L1} = ||\boldsymbol{I}_{out} - \boldsymbol{I}_{target}||_1, \tag{11}$$

where $||\cdot||_1$ indicates the $L1$ distance, and $\boldsymbol{I}_{out}$ and $\boldsymbol{I}_{target}$ represent the predicted image and target image. The perceptual loss proposed in [21] calculates the L1 distance of the feature maps extracted by the VGG-19 [37] network:

$$\mathcal{L}_{prec} = \sum_{i=1}^{5} ||\phi_i(\boldsymbol{I}_{out}) - \phi_i(\boldsymbol{I}_{taregt})||_1, \tag{12}$$

where $\phi_i(\boldsymbol{I}_{out})$ is the features of the i_{th} layer. In addition, the style loss optimizes the statistical error between the feature maps:

$$\mathcal{L}_{style} = \sum_{i} ||G_i^{\phi}(\boldsymbol{I}_{out}) - G_i^{\phi}\boldsymbol{I}_{taregt})||_1, \tag{13}$$

where G_i^{ϕ} denotes the Gram matrix of features. The overall loss is presented as:

$$\mathcal{L} = \sum_{n=1}^{N}(n+1) * (\lambda_{L1}\mathcal{L}_{L1}^{n} + \lambda_{prec}\mathcal{L}_{prec}^{n} + \lambda_{style}\mathcal{L}_{style}^{n}) \tag{14}$$

where $\mathcal{L}^n$ is the loss of the n_{th} scale, and the scale closer to the output is given a larger weight $n-1$.

4 Experiments

4.1 Datasets.

We conduct experiments on VITON [15] and MPV [6] datasets. VITON dataset is commonly used in virtual try-on. it contains a training set of 14,221 image

pairs and a testing set of 2,032 image pairs. Each pair has a front-view photo and an in-shop clothes image with the resolution 256×192. MPV dataset is a recently virtual try-on dataset with multiple views, containing 35,687 / 13,524 person/clothes images at 256×192 resolution where a test set of 4175 image pairs is split out. To fair comparison, following [19,10], the images tagged as back ones are removed since the target garment is only from the front.

4.2 Implementation Details.

Network structure. The two encoders have the FPN [25] structure with five layers, each layer consists of a downsampling convolution with the stride of 2, followed by two residual blocks. The MFE flow estimator in each layer comprises ConvNets with four convolution layers, and the hidden dimensions are $[256, 128, 64, 32]$. To obtain a large receptive field, the kernel size of the last three convolution layers is set to 7. the shallow encoder and decoder are ConvNets with two convolution layers without downsampling, and the hidden dimensions are $[32, 64]$. The $K = 6$ in our experiment. Our approach only has a single stage, the computational efficiency is similar to the clothes warping stage in [14,10].

Training details. We adopt the same training parameters for the two datasets. All our experiments are conducted using Pytorch on Tesla V100 GPUs. The AdamW [28] optimizer is adopted with a batch size of 8. We train the model for 200 epochs where the initial learning rate is 5×10^{-5} and is reduced to 0.1 times the original every 50 epochs. We set the weight of the loss function $\lambda_{L1} = 1, \lambda_{prec} = 1, \lambda_{style} = 100$.

4.3 Qualitative Results.

Results on VITON. To qualitatively evaluate our method, we visually compare our method with three recently proposed virtual try-on works with available code implementations, including CP-VITON+ [29], ACGPN [43], PFAPN[10], as shown in Fig. 4. It shows that all comparing works are able to roughly align clothes with the target person pose. However, noticeable artifacts are observed from their results where there are complex poses or misalignment occurs between the target clothes and the person.

As shown in the first row of Fig. 4, baseline methods fail to preserve the striped pattern after warping, especially around the highly non-rigid body parts like the forearm and waist. The second row shows the results with the side view where baseline methods are not able to deal with large deformation in poses and lead to blur or incorrect fitting results. In comparison, our proposed method is capable of extracting accurate structural and textural information and performing reasonable warping even when there exists huge discrepancy (e.g. large poses or long-sleeve in target clothes while short-sleeve in reference image). The last two columns that parser-based methods like CP-VITON+ [29] and ACGPN [43] are delicate to segmentation errors while learning warping flows. They do not always produce reasonable results for areas like necklines and lower-body parts. Even though PFAPN is a parser-free framework with less

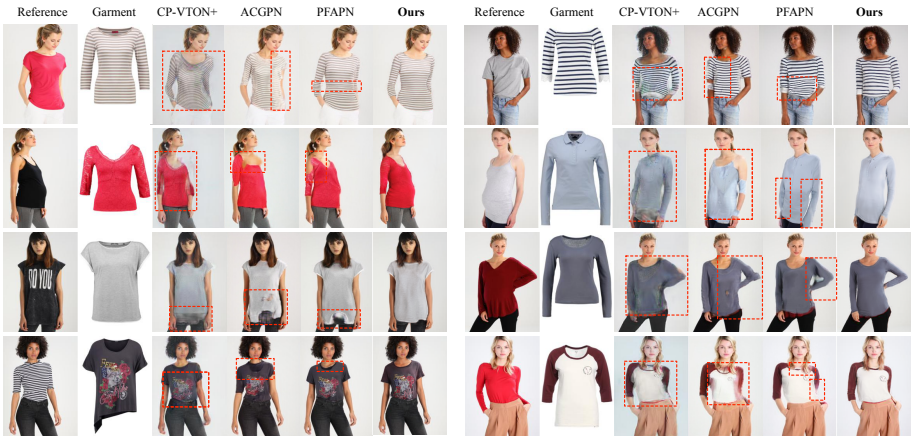

Fig. 4. Visual comparison on VITON dataset. Different regions are marked in red.

Table 1. Comparisons with State-of-the-art methods on VITON under paired setting.

Methods	CP-VTON [41]	ClothFlow [14]	ACGPN [43]	ZFlow [5]	**SDAFN(Ours)**
FID (↓)	30.50	23.68	-	15.17	**10.97**
SSIM (↑)	0.784	0.843	0.845	0.885	**0.888**
IS (↑)	2.757	-	2.829	-	**2.859**
PSNR (↑)	21.01	23.60	-	25.46	**26.48**

distortion in clothes warping, it cannot preserve or generate the body parts well which results in blurry arms and shoulders. Unlike them, our method clearly preserves the characteristics of both the target clothes as well as the body parts, benefiting from the proposed self- and cross-DAFlows.

Results on MPV. To verify the performance of our algorithm on multi-view data, we visualize the results on the MPV [6] dataset. As illustrated in Fig. 5, the reference pose images are shown in the first row while the target garments are shown in the first column. The manipulated results are presented in other columns. It can be seen that our method captures the texture and pose well even with large variations in clothes design and viewpoint change. Besides, it generates realistic body parts even if they are unseen from the reference images which demonstrates the robustness of our method.

Inference at higher resolutions. In contrast to previous methods that can only predict at the inference stage the fixed size of images which is the same as training, our method is able to perform try-on at a higher resolution without re-training. We apply the model trained at 256×192 resolution to test the images with the size of 512×384 from VITON-HD [15]. It is obvious to be seen in Fig. 6 that our pure flow-based approach can retain finer texture information from the original image by linear interpolating the flows, as compared with the image-level interpolation scheme. It helps preserve the clarity of the characters and the patterns of the clothes, which verifies the scalability of our model.

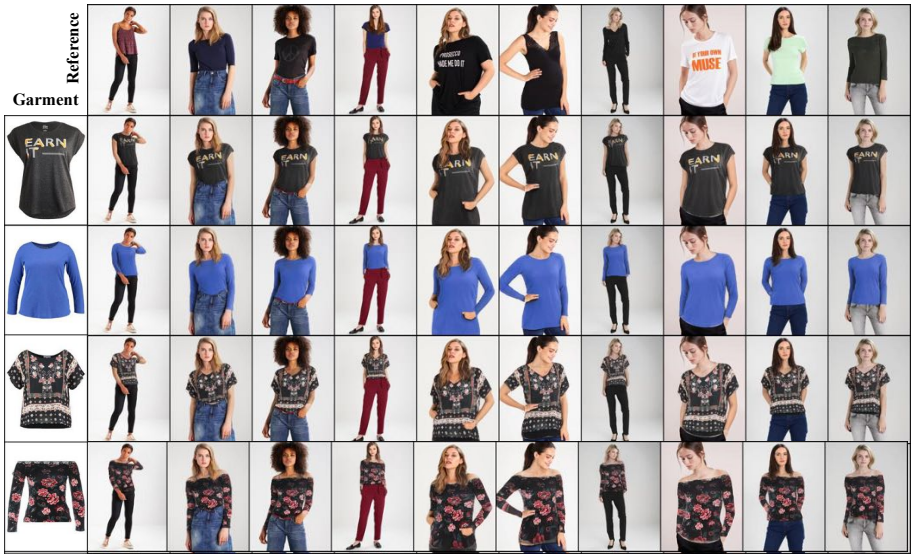

Fig. 5. From diverse perspectives, our approach generates high-quality images and accurately preserves vivid attributes on MPV dataset.

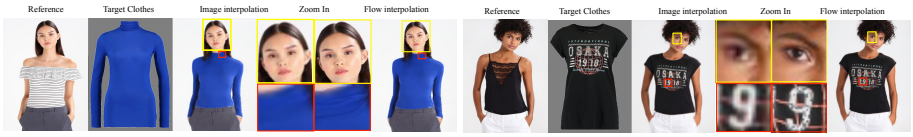

Fig. 6. Comparisons with image interpolation and flow interpolation. Flow interpolation has the advantage of scaling to the higher resolution.

4.4 Quantitative Results.

We conduct both paired and unpaired settings to quantitatively compare our work with baseline methods including CP-VTON [41], Clothflow[14], ACGPN[43], PFAFN [10], WUTON [19] and ZFlow[5]

Paired setting. In the paired setting, we use the Structure Similarity Index Measure (SSIM) [36], the Peak Signal-to-Noise Ratio (PNSR) [18] and the Fréchet Inception Distance(FID) [17] to measure the similarity between the synthesized image and ground truth image. The Inception Score (IS) [35] is applied to evaluate the realism of the generated images. We take the target image (the same person wearing the same clothes) as the ground truth which is used to compare with the synthesized image for computing these metrics. It is noted that PFAFN and WUTON were removed from those measurements as they need to take the target image as input for inference.

As shown in Table 1, our approach consistently outperforms the baselines methods under all the metrics. Achieving the best FID and IS scores shows our

Table 2. Comparisons with state-of-the-art methods on VITON and MPV datasets under unpaired setting. For FID, the lower is the better.

Methods	CP-VTON [41]	ClothFlow [14]	ACGPN [43]	SDAFN (Ours)	WUTON [19]	PF-APN [10]	**SDAFN+ (Ours)**
VITON (FID↓)	24.43	14.43	15.67	12.05	-	10.09	**9.46**
MPV (FID↓)	-	-	-	8.245	7.927	6.429	**5.805**

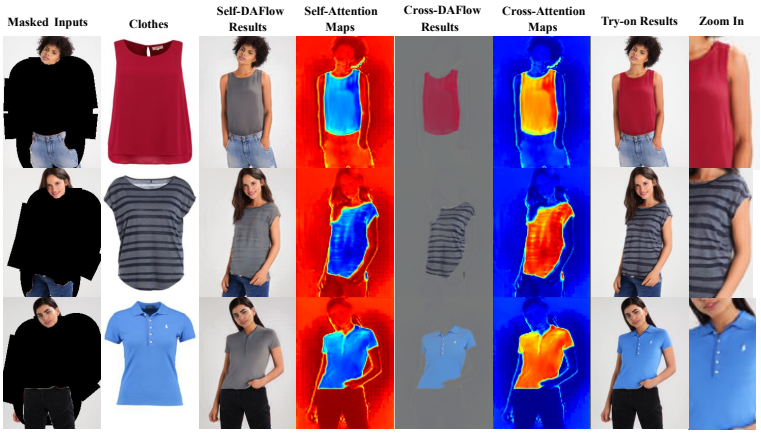

Fig. 7. Visualization of the attention maps and results of deformable attention warping.

ability to obtain better visual quality in results and higher SSIM and PSNR scores demonstrate our advantage of accuracy.

Unpaired setting. Under the unpaired setting, there is no ground truth image for comparison. We directly adopt FID [17] to evaluate the similarity between the generated images and the real images. As reported in Table 2, we compare both parser-free and parser-based methods. In particular, for a fair comparison, we train SDAFN+ model like PFAPN [10] to compare, which allows us to directly input the original image without pose keypoints. In detail, SDAFN+ applies the prediction of SDAFN to replace the masked person and pose keypoints as the input for training. Compared with parser-based methods, we achieve obvious improvements without segmentation guidance in FID. Compared with the parser-free method, the better FID on both VITON and MPV datasets indicates the generality of our method.

4.5 Ablation Study.

In this section, we mainly evaluate the effectiveness of the proposed deformable attention flow module.

Deformable Attention Flow. The choice of the sampling key number K in deformable attention flow is studied. Setting K to 1 can be regarded as the

Table 3. Performances of models with different sampling number.

K	SSIM	PSNR
1	0.833	22.43
2	0.843	23.25
4	0.868	24.71
6	**0.888**	**26.48**
8	0.878	25.35

Table 4. The effect of different modules.

Config	SSIM	PSNR
Baseline	0.818	20.88
+Cascade	0.827	21.83
+Shallow En.	0.833	22.43
+DAFN	**0.888**	**26.48**

traditional flow operation. With the increase of K, the performance gradually improves and drops slightly after it reaches 6, as shown in Table 3. It shows that sampling multiple attention regions can effectively improve the flow operation. We set it to be 6 as a balance between performance and computational cost. In addition, as K increases, more GPU memory is consumed for the sampling and warp operations but increases limited time consumption.

Visualization of Attention. In Fig. 7, we visualize self- and cross-DAFlow along with the intermediate attention maps. After the self-DAFlow process, the clear clothes and torso shape are generated along with shadows which renders the realistic human body. The cross-DAFlow not only predicts accurate warping results but also retrains the fine textural details of the clothes, even with the folds on the t-shirt as shown in the second row. The attention maps show strong evidence that our flow module is able to learn the 3D priors where the clothes fit the skin and make the final try-on more realistic. In general, our method effectively decouples the textural and structural information and learns to add 3D shadow effect without introducing auxiliary models or labels.

Modular Ablation Study. The ablation study on the effectiveness of each module in our framework is reported in Table.4. It is shown that both SSIM and PSNR metrics increase with each of the modules added, including cascaded flow estimation, shallow encoder and decoder, and deformable attention flow network (DAFN). Among them, DAFN contributes the most to the improvement. The best performance is reported when integrating all the modules together.

5 Applications on Other Tasks

In this section, we mainly verify the versatility of the proposed deformable attention flow on two other image editing tasks, namely multi-view synthesis and image animation. To deal with only one image transformation (compared to paired images given in try-on task), we remove the self-MFE in DAFN from our model. More details for implementation can be found in the supplementary.

Multi-view synthesis. View synthesis aims to generate a novel view of an object given one single image as input. We apply our proposed SDAFN to learn the structural correlations of the same object under different viewing. We conduct experiments on the ShapeNet [3] chairs dataset. As demonstrated in

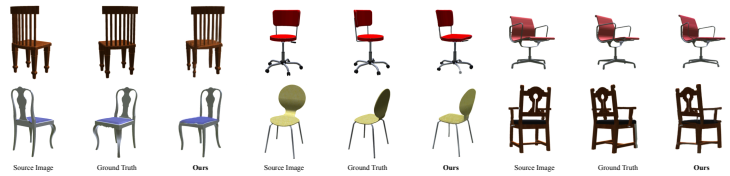

Fig. 8. Qualitative results of the multi-view synthesis on ShapeNet.

Fig. 9. Qualitative results of the image animation task on FashionVideo.

Fig. 8, our method is able to predict the one-shot input chair under different views with accurate textures. It successfully reconstructs the unseen parts which are under occlusion from the input image.

Image animation. Given an input image and a driving video, image animation is to generate a video sequence so that the object in the source image is animated according to the motion of the driving video. We experiment on the Fashion Video dataset [46], using 500 videos for training and 100 videos for testing. Example animations produced by our SDAFN on two actions are shown in Fig.9, where accurate reconstruction of the input pose is generated even in the case of complex motion like the back of the body.

6 Conclusions

In this paper, we present a novel single-stage virtual try-on framework. With only pose map as guidance, our model generates photo-realistic fitting results in an end-to-end manner. The proposed deformable attention flow (DAFlow) module estimates the accurate structure while retaining the vivid texture. It synthesizes photo-realistic human torso and fitting clothes with 3D shadows. Extensive experiments and evaluations show that our proposed method not only achieves superior performance on virtual try-on, but also can be extended to other image editing tasks, such as multi-view synthesis and image animation.

References

1. Bertiche, H., Madadi, M., Escalera, S.: Cloth3d: clothed 3d humans. In: European Conference on Computer Vision. pp. 344–359. Springer (2020)
2. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-garment net: Learning to dress 3d people from images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5420–5430 (2019)
3. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
4. Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14131–14140 (2021)
5. Chopra, A., Jain, R., Hemani, M., Krishnamurthy, B.: Zflow: Gated appearance flow-based virtual try-on with 3d priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5433–5442 (2021)
6. Dong, H., Liang, X., Shen, X., Wang, B., Lai, H., Zhu, J., Hu, Z., Yin, J.: Towards multi-pose guided virtual try-on network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9026–9035 (2019)
7. Duchon, J.: Splines minimizing rotation-invariant semi-norms in sobolev spaces. In: Constructive theory of functions of several variables, pp. 85–100. Springer (1977)
8. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: Proceedings of the European conference on computer vision (ECCV). pp. 534–551 (2018)
9. Ge, C., Song, Y., Ge, Y., Yang, H., Liu, W., Luo, P.: Disentangled cycle consistency for highly-realistic virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16928–16937 (2021)
10. Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., Luo, P.: Parser-free virtual try-on via distilling appearance flows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8485–8493 (2021)
11. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 932–940 (2017)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
13. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7297–7306 (2018)
14. Han, X., Hu, X., Huang, W., Scott, M.R.: Clothflow: A flow-based model for clothed person generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10471–10480 (2019)
15. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7543–7552 (2018)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)

17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
18. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: 2010 20th international conference on pattern recognition. pp. 2366–2369. IEEE (2010)
19. Issenhuth, T., Mary, J., Calauzènes, C.: Do not mask what you do not need to mask: a parser-free virtual try-on. In: European Conference on Computer Vision. pp. 619–635. Springer (2020)
20. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in neural information processing systems **28** (2015)
21. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
22. Kitaev, N., Kaiser, Ł., Levskaya, A.: Reformer: The efficient transformer. arXiv preprint arXiv:2001.04451 (2020)
23. Lahner, Z., Cremers, D., Tung, T.: Deepwrinkles: Accurate and realistic clothing modeling. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 667–684 (2018)
24. Li, K., Chong, M.J., Zhang, J., Liu, J.: Toward accurate and realistic outfits visualization with attention to details. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15546–15555 (2021)
25. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
26. Liu, P.J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., Shazeer, N.: Generating wikipedia by summarizing long sequences. arXiv preprint arXiv:1801.10198 (2018)
27. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
29. Minar, M.R., Tuan, T.T., Ahn, H., Rosin, P., Lai, Y.K.: Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In: CVPR Workshops (2020)
30. Mir, A., Alldieck, T., Pons-Moll, G.: Learning to transfer texture from clothing images to 3d humans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7023–7034 (2020)
31. Qiu, J., Ma, H., Levy, O., Yih, S.W.t., Wang, S., Tang, J.: Blockwise self-attention for long document understanding. arXiv preprint arXiv:1911.02972 (2019)
32. Raj, A., Sangkloy, P., Chang, H., Lu, J., Ceylan, D., Hays, J.: Swapnet: Garment transfer in single view images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 666–682 (2018)
33. Ren, Y., Wu, Y., Li, T.H., Liu, S., Li, G.: Combining attention with flow for person image synthesis. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3737–3745 (2021)
34. Ren, Y., Yu, X., Chen, J., Li, T.H., Li, G.: Deep image spatial transformation for person image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7690–7699 (2020)

35. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016)
36. Seshadrinathan, K., Bovik, A.C.: Unifying analysis of full reference image quality assessment. In: 2008 15th IEEE International Conference on Image Processing. pp. 1200–1203. IEEE (2008)
37. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
38. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8934–8943 (2018)
39. Tay, Y., Bahri, D., Yang, L., Metzler, D., Juan, D.C.: Sparse sinkhorn attention. In: International Conference on Machine Learning. pp. 9438–9447. PMLR (2020)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
41. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 589–604 (2018)
42. Wang, S., Li, B.Z., Khabsa, M., Fang, H., Ma, H.: Linformer: Self-attention with linear complexity. arXiv preprint arXiv:2006.04768 (2020)
43. Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., Luo, P.: Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7850–7859 (2020)
44. Yu, H., Chen, X., Shi, H., Chen, T., Huang, T.S., Sun, S.: Motion pyramid networks for accurate and efficient cardiac motion estimation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 436–446. Springer (2020)
45. Yu, R., Wang, X., Xie, X.: Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10511–10520 (2019)
46. Zablotskaia, P., Siarohin, A., Zhao, B., Sigal, L.: Dwnet: Dense warp-based network for pose-guided human video generation. arXiv preprint arXiv:1910.09139 (2019)
47. Zhou, T., Tulsiani, S., Sun, W., Malik, J., Efros, A.A.: View synthesis by appearance flow. In: European conference on computer vision. pp. 286–301. Springer (2016)
48. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)