Bidirectionally Deformable Motion Modulation For Video-based Human Pose Transfer

Anonymous CVPR submission

Paper ID 6538

Abstract

Video-based human pose transfer is a video-to-video generation task that animates a plain source human image based on a series of target human poses. Considering the difficulties in transferring highly structural patterns on the garments and discontinuous poses, existing methods often generate unsatisfactory results such as distorted textures and flickering artifacts. To address these issues, we propose a novel Deformable Motion Modulation (DMM) that utilizes geometric kernel offset with adaptive weight modulation to simultaneously perform feature alignment and style transfer. Different from normal style modulation used in style transfer, the proposed modulation mechanism adaptively reconstructs smoothed frames from style codes according to the object shape through an irregular receptive field of view. To enhance the spatio-temporal consistency, we leverage bidirectional propagation to extract the hidden motion information from a warped image sequence generated by noisy poses. The proposed feature propagation significantly enhances the motion prediction ability by forward and backward propagation. Both quantitative and qualitative experimental results demonstrate superiority over the state-of-the-arts in terms of image fidelity and visual continuity. The source code will be publicly available.

1. Introduction

The video-based human pose transfer is a task to animate the plain source image according to a series of desired postures. It is challenging due to problems of spatio-temporally discontinuous poses and highly structural texture misalignment as depicted in Figure 1. In this paper, we aim to tackle these problems with an end-to-end generative model to maximize the value of applications in various domains including person re-identification [49], fashion recommendation [13, 20], and virtual try-on [8, 33, 42].

Existing works focus on three categories to solve the spatial misalignment problem, including prior generation [7, 27, 28, 45, 46], attention module [34, 39, 52], and flow warping [35, 48]. There are many side effects in these methods such as spatially misaligned content, blurry visual quality and unreliable flow prediction. Some methods [22, 24, 25] proposed to obtain the spatial transformation flow by computing the vertex matching in 3D neural rendering process. The main advantage is to preserve more texture details of the source image. However, the generative networks struggle to render new content for occluded regions since flows in such regions are not accurate.

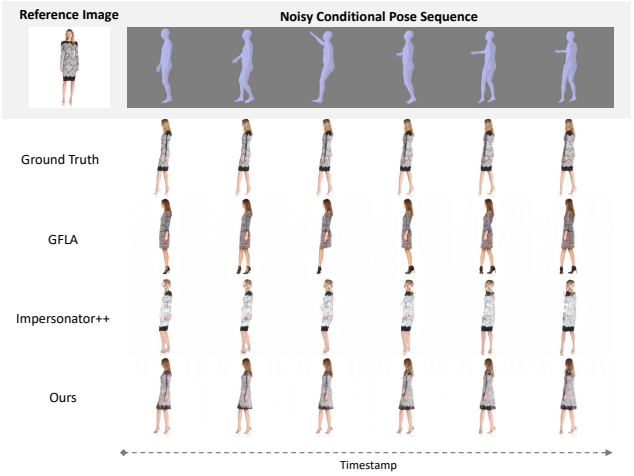

Figure 1. Examples of a synthesized video clip based on some noisy poses. Existing methods such as GFLA [35] and Impersonator++ [25] fail to generate realistic videos due to problems of spatio-temporally discontinuous poses and highly structural texture misalignment while our method can generate highly plausible texture with seamless transition between consecutive frames. Please zoom in for more details.

To obtain animated sequences with smooth human gesture movements, the temporal coherence is the main determinant. Different from most of the generative tasks such as image inpainting or image super-resolution, the conditional inputs of the sequence in this task are noisy. It is because the existing third-party human pose extractors [2, 10, 21] fail to extract accurate pose labels in the video frames. It increases the difficulty to predict the temporal correspon-

dence for generating a smooth sequence of frames, especially the highly structural patterns on garments and occluded regions. In general, previous works [24, 25, 35, 35] mainly use recurrent neural networks to solve this problem by taking the previously generated result as the input of current time step. However, the perceptual quality is still unsatisfactory due to limited receptive field of view along time space. We observe that solely relying on unidirectionally hidden states in recurrent units to interpolate the missing content is insufficient. It motivates us to utilize all the frames within the mini batch to stabilize the temporal statistics in the generated sequence.

To alleviate the aforementioned problems, we propose a novel modulation mechanism – Deformable Motion Modulation (DMM) incorporated with bidirectional recurrent feature propagation to perform spatio-temporal affine transformation and style transfer simultaneously. It is designed with three major components, including motion offset, motion mask and the modulated style weight. To strengthen the temporal consistency, the offset and mask are responsible for estimating the local geometric transformation based on the features of two spatially misaligned adjacent frames. One of the feature branches comes from forward propagation branch or backward propagation branch. The bidirectional feature propagation encapsulates the temporal information of the entire sequence so that a long-range temporal correspondence of a sequence from the forward flow to the backward flow can be captured at current time. By maintaining more semantic details from the source image to process the coarsely aligned features, the style weights are modulated by the style codes extracted from the source image. The corresponding affine transformation is enhanced with the augmented spatial-temporal sampling offset. It can produce a dynamic receptive field of view to track semantics so that it can synthesize a sequence of plausible and smooth video frames. The main contributions of this work can be summarized as follows:

- We propose a novel Deformable Motion Modulation that utilizes geometric kernel offset with adaptive weight modulation to perform spatio-temporal affine transformation and style transfer simultaneously;
- We design a bidirectionally recurrent feature propagation on coarsely warped images to generate target images on top of noisy poses so that a long-range temporal correspondence of the sequence can be captured at current time;
- We demonstrate the superiority of our method in both quantitative and qualitative experimental results with a significant enhancement in perceptual quality in terms of visual fidelity and temporal consistency.

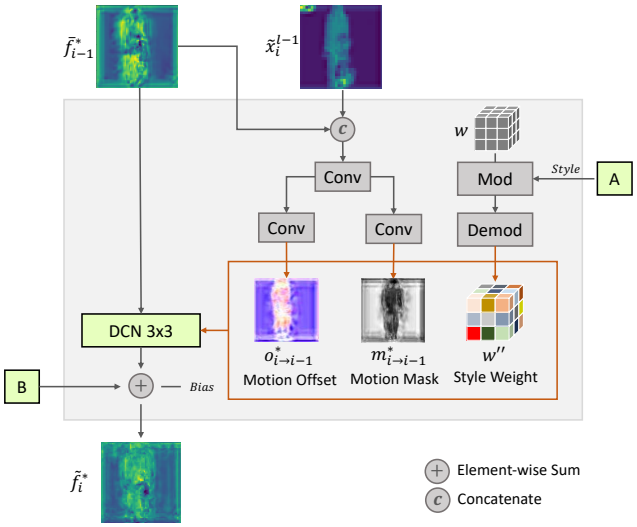

Figure 2. Illustration of the proposed Deformable Motion Modulation (DMM) module. The motion offset and motion mask are parametrized by the output of coarsely warped features f^*_{i-1} in forward branch or b^*_{i+1} (skipped for simplicity) in backward branch, the output results generated from previous layer $\tilde{x}^{l-1}_i$ at time i, and the affine transformation based on I_s.

2. Related Work

Human Pose Transfer. Recent research in image-based human pose transfer can be categorized as prior-based, attention-based, and flow-based. Initial methods [28, 45] proposed a prior-based generative model to combine the generated results with residual priors. In addition to residual maps, some solutions [7, 27] proposed to pre-generate the target parsing maps in order to enhance the semantic correspondence. Yu *et al.* [46] also introduced an edge prior to reconstruct the fragile high frequency on the characteristics of garments. Although these priors are tailor-made to reconstruct details of the source image, inaccurately generated priors limit the ability to synthesize new content, especially when encountering large occlusion variations. Some attention-based methods proposed to compute dense correspondences in feature space via activated pose attention [52] and spatial attention [34, 39]. Despite the fact that these kinds of attentional operation can achieve better scores in some quantitative evaluation metrics such as FID, the qualitative visualizations show a blurry effect on the generated images due to insufficient texture and shape guidance. In view of this problem, flow-based methods [22, 35, 48] warped the features of the source image by estimating the pose correspondence. Notwithstanding that they can preserve the characteristics of the source image, unreliable optical flow prediction is a bottleneck for these methods to transfer complex texture patterns.

Apart from spatial transformation, the video-based human pose transfer has an additional challenge on maintaining temporal consistency. Current approaches [24, 25, 35]

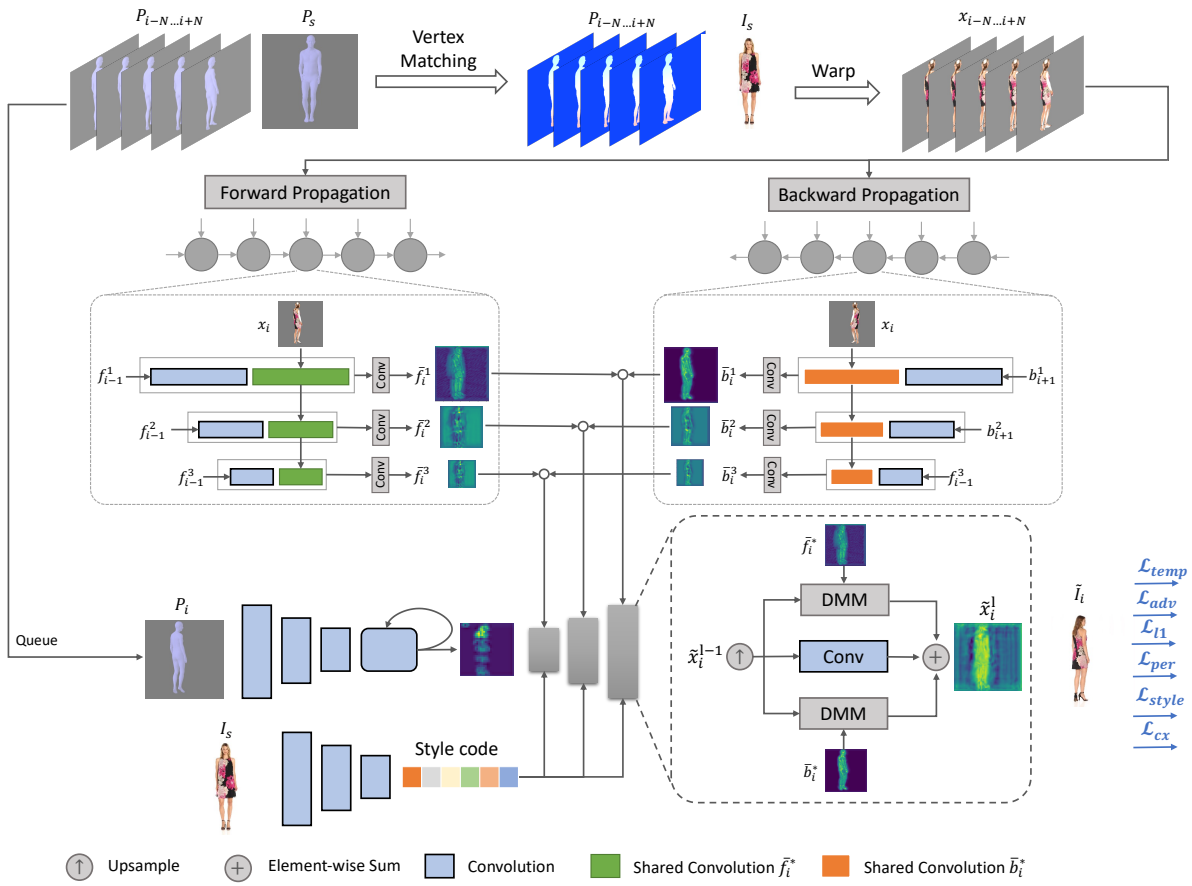

Figure 3. Overview of the proposed model. We use a bidirectional propagation mechanism to manipulate coarsely spatial-aligned sequence rendered by vertex matching. The pose is encoded to capture structural guidance by a self-recurrent convolution unit by a Structural Encoder. The generator decoder progressively synthesizes target image by fusing features from forward and backward propagation branches via the proposed Deformable Motion Modualtion (DMM) block and the source style code extracted by a Style Encoder.

employed unidirectional forward propagation in recurrent networks to extract the hidden temporal information. However, it is insufficient to produce a spatio-temporally smooth sequence due to the problem of noisy pose that cannot be detected at certain time steps. To address this issue, Ren *et al.* [35] used a convolution network to preprocess the 2D skeletons by transferring knowledge of 3D pose estimation in advance. Due to the domain gap between different datasets, reducing the number of key points in the heatmap limits the ability of flow prediction. Without training an extra network to perform noisy pose recovery, our method is still able to generate temporally coherent videos transferred from source images.

Video-to-video Generation. With the success of conditional Generative Adversarial Networks [9, 31](cGANs), video-to-video models convert semantic input videos to photorealistic videos. Wang *et al.* [43] introduced a sequential generative model to extract feature correlations from adjacent frames. Due to weak spatial transformation ability, it failed to produce plausible images. Siarohin *et al.* [36, 37] suggested to simulate the motion directly from the driving images by using zeroth-order and first-order Taylor series expansions to estimate the transformation flow. However, it sacrificed the controllability of generating images on arbitrary poses because of domain gaps.

Deformable Convolutional Networks. Due to the shortcoming of geometric transformations in Convolutional Neural Networks (CNNs) [19], Deformable Convolutional Networks (DCNs) [5, 51] suggested to learn the kernel offsets by augmenting the spatial sampling locations. The deformable alignment regressed by flow-guided features demonstrated effective spatial transformation capabilities in several generative tasks, including image inpainting [23] and image super-resolution [4,40]. Inspired by these works, we have the motivation to enhance the style transfer ability and the temporal coherence by modulating affine transformations from the source image.

3. Methodology

To begin with, we define some notations used in this paper. Given a source person image I_s, the corresponding source pose P_s, and a sequence of spatially arbitrary

target pose $P(1:M)$, where M is the total numbers of frames in a sequence. The goal of video-based human pose transfer is to animate the I_s according to $P(1:M)$ with desired movements including free-form view angles, postures, or body shapes, etc. The proposed end-to-end and recurrent generative model $\mathcal{G}$ can be formulated as $\hat{I}_{1:M} = \mathcal{G}(I_s, P_s, P_{1:M})$.

3.1. Deformable Motion Modulation (DMM)

The major challenge of video-based pose transfer is to maintain the spatio-temporally misaligned characteristics of I_s while synthesizing unseen content according to the target poses. In this subsection, we introduce a new modulation mechanism – Deformable Motion Modulation (DMM) to synthesize continuous frame sequences by modulating the affine transform of I_s with an augmentation of spatio-temporal sampling locations. It aims to estimate local geometric transformations on an initially aligned feature space so that it can enhance the smoothness of the propagated features in forward and backward branches. We design the proposed DMM with three components, namely motion offset, motion mask and style weight, inspired by the success of Deformable Convolution Network (DCN) [5,51] and StyleGANv2 [14, 15]. As depicted in Figure 2, we parametrize them as the output of coarsely warped features f^*_{i-1} in the forward branch or b^*_{i+1} in the backward branch, the output results generated from previous layer $\widetilde{x}_i^{l-1}$ at time i, and the source style code from I_s. We firstly initialize the standard convolution as

$$\widetilde{f}_i\ (p) = \sum_{k=1}^{K} w_k \cdot f_i\ (p) + bias\ , \tag{1}$$

where K is a set of the sampling location of a kernel, $y\,(p)$ is the convoluted result of input x at position p with the sampled weight w_k. To equip convolution with modulation and irregular receptive field of view, we formulate our proposed DMM as

$$\widetilde{f}_i\ (p) = \sum_{k=1}^{K} w''_k \cdot\ m_{i\to i-1}(p) \cdot f_i\ (p + p_k + o_{i\to i-1}(p)), \tag{2}$$

where p_k is the pre-defined kernel offset depending on K, $o_{i\to i-1} \in \mathbb{R}^{2K}$ and $m_{i\to i-1} \in \mathbb{R}^{K}$ are both learnable shift offsets and a non-negative modulation scalar for a kernel at p location regressed by the geometric relationship between the propagated features f_i or b_i and the previous generation layer $\widetilde{x}_i^{l-1}$, w''_k is the stylized weights modulated by the incoming statistics of style code extracted from I_s. More specifically, w''_k is responsible for manipulating the style transfer accompanied with the motion mask $m_{i\to i-1}$ so that a long-range spatio-temporal correspondence of the sequence can be captured at the current time. This goal can be achieved by computing the weights with demodulation [15], which is expressed as

$$w'_{jhk} = A_j \cdot\ w_{jhk}, \tag{3}$$

$$w''_{jhk} = {w'_{jhk}}\Big/\sqrt{\sum_{jk} {w'_{jhk}}^2 + \epsilon}, \tag{4}$$

where w_{jhk} represents the weights of j-th input feature and h-th output feature map on k-th sampling kernel location, i.e., $w_k \subset w_{jhk}$, A_j is the j-th scalar from the source style vector, w'_{jhk} is computed for estimating the affine transformation based on the statistics of incoming style code, ϵ is a small number to prevent computation from numerical error. The demodulation can well preserve the semantic details of the source image while it is able to interpolate unseen content by considering the forward and backward propagation features. The augmented spatio-temporal sampling offsets can also produce dynamic receptive fields of view to track the semantics of interest so that it can synthesize a sequence of good-looking and smooth video frames.

3.2. Bidirectional Propagation

It has been a challenge to produce stable and smooth videos simply by relying on current pose to generate the target person image due to discontinuous noisy poses extracted by some third-party human skeleton extractors [2, 10, 21]. We introduce a simple bidirectional propagation mechanism to interpolate the probability of missing structural guidance from both forward and backward propagation.

Mesh Flow. Following previous work [25], we apply SPIN [18] as the 3D human pose and shape estimator to predict parametric representations by inferencing RGB images into the implicit differentiable model SMPL [26]. The SMPL representation consists of three major elements, including a weak perspective camera vector $C \in \mathbb{R}^3$, a pose vector $\theta \in \mathbb{R}^{72}$ and a shape vector $\beta \in \mathbb{R}^{10}$. It can parametrize a triangulated mesh to produce the explicit pose representation by computing the corresponding function $SMPL\,(\theta, \beta) \in \mathbb{R}^{6890\times 3}$. To evaluate the transformation flow $F_{i\to s} \in \mathbb{R}^{2\times H\times W}$ between P_s and P_i, where H and W are height and width of the generated image resolution, P_s and P_i are the source image and target pose at time i. The existing mesh renderer such as Neural Mesh Renderer (NMR) [16] comes in handy to formulate the displacement of each vertex between P_s and P_i.

Recurrent propagation. Once we obtain the transformation flow $F_{i\to s}$, we perform feature propagation to extract the latent temporal information in a recurrent manner. We leverage a bidirectional propagation mechanism to manipulate the coarsely spatial-aligned sequence before feeding it into the generator. As shown in Figure 3, the pre-warped frames are formulated as

$$x_{i-N:i+N} = warp\,(F_{i\to s}\,(P_s, P_{1:M})\,, I_s)\,, \tag{5}$$

where $N = M/2$. We use a shared 2D CNN encoder to independently extract features of $x_{i-N:i+N}$ in both forward branch $\mathcal{F}$ and backward branch $\mathcal{B}$, respectively. With the recurrent propagation, the extracted features at time i are encapsulated with the spatio-temporal information across the entire input sequence in the feature space. The temporal forward features and backward features computed at time i are represented as

$$f_i^* = conv\left(\mathcal{F}\left(x_i\right)|_{l=*} \circledcirc f_{i-1}^*\right) \tag{6}$$

$$b_i^* = conv\left(\mathcal{B}\left(x_i\right)|_{l=*} \circledcirc b_{i+1}^*\right) \tag{7}$$

where $\mathcal{F}\left(x_i\right)|_{l=*}$ indicates the feature maps of the forward encoder $\mathcal{F}$ at layer $l = *$, $\mathcal{B}$ is also used for backward encoder, and $\circledcirc$ denotes concatenation operator. With the recurrent features from forward and backward propagations, the model can expand the field of view across the whole input sequence so that a more robust spatio-temporal consistency is captured during the generation process. Moreover, the outliers of input noisy pose at time i can also be interpolated by the warped features from $x_{i-N:i-1}$ to $x_{i+1:i+N}$. With the assistance of Equation 2, the probability of estimating generative result can be formulated as

$$q\left(x_i|I_s\right) = \prod_{i-N}^{i} q\left(f_i|f_{i-1}\right) + \prod_{i}^{i+N} q\left(b_i|b_{i+1}\right) \tag{8}$$

The combinations of $q\left(x_i|I_s\right)$ can dramatically provide positive gain to the network in synthesizing new content by feature interpolation.

3.3. Objective Loss Function

Following similar training strategies in current pose transfer frameworks [25, 35], the final objective loss function in our model is composed of six terms including a spatial adversarial loss $\mathcal{L}_{adv}$, a spatio-temporal adversarial loss $\mathcal{L}_{temp}$, an appearance loss $\mathcal{L}_{l1}$, a perceptual loss $\mathcal{L}_{per}$ a style loss $\mathcal{L}_{gram}$, and a contextual loss $\mathcal{L}_{cx}$ as follows:

$$\begin{aligned}\mathcal{L}_{full} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{temp}\mathcal{L}_{temp} + \lambda_{l1}\mathcal{L}_{l1} \\ +\lambda_{per}\mathcal{L}_{per} + \lambda_{gram}\mathcal{L}_{gram} + \lambda_{cx}\mathcal{L}_{cx}\ ,\end{aligned} \tag{9}$$

where λ_{adv}, λ_{temp}, λ_1, λ_{per}, λ_{gram}, and λ_{cx} are the hyper-parameters to optimize the convergence of the network.

Spatial adversarial loss. We utilize the traditional generative adversarial loss [9, 29] $\mathcal{L}_{adv}$ to mimic the distribution of the training set with a convolutional discriminator D_s. It is formulated as:

$$\mathcal{L}_{adv} = \mathbb{E}\left[\log\left(D_s\left(I_s,\ I_i\right)\right) + \log\left(1 - D_s\left(I_s,\ \hat{I}_i\right)\right)\right] \tag{10}$$

where $(I_s, I_i) \in \mathbb{I}_{real}$, $\hat{I}_i \in \mathbb{I}_{fake}$, and $i \in 1 \ldots M$ indicate samples from the distribution of real person image, generated person image, the numbers of an input patch.

Temporal adversarial loss. Similar with $\mathcal{L}_{adv}$, the temporal adversarial loss $\mathcal{L}_{temp}$ optimizes the temporal consistency in time and feature channels of a mini patch with a 3D CNN discriminator D_t.

Appearance loss. To enforce discriminatively pixel-level supervision, we employ a pixel-wise L1 loss to provide guidance on synthesizing photo-realistic appearance compared to the ground-truth image.

Perceptual loss. To minimize the distance in feature-level space, we apply a standard perceptual loss [12]. It computes the L1 difference of a selected layer $\ell = Conv1\_2$ from a VGG-19 [38] model $\theta_\ell\left(\cdot\right)$ pre-trained in ImageNet [6]. It is defined as

$$\mathcal{L}_{per} = \sum_{C_\ell H_\ell W_\ell} \|\theta_\ell(\hat{I}_i) - \theta_\ell(I_i)\|\|_1. \tag{11}$$

where C_ℓ is the number of channels, H_ℓ and W_ℓ are the height and width of the feature maps in a particular layer ℓ respectively.

Style loss. Similar to the perceptual loss to minimize the L1 distance in feature-level space, we further calculate the Gram matrix of some activated feature maps at the selected layers to maximize the similarity.

$$\mathcal{L}_{gram} = \sum_{C_\ell H_\ell W_\ell} \|Gram(\theta_\ell(\hat{I}_i)) - Gram(\theta_\ell(I_i)|)\|_1. \tag{12}$$

where the used layers are the same as in perceptual loss.

Contextual loss. To maximize the similarity between two non-aligned images in context space, we utilize the contextual loss [30] to allow spatial alignment according to contextual correspondence during the deformation process.

$$\mathcal{L}_{cx} = -\sum_{C_\ell H_\ell W_\ell} log\left[CX\left(\delta_\ell\left(\hat{I}_i\right), \delta_\ell\left(I_t\right)\right)\right], \tag{13}$$

where $\ell = relu\left\{3\_2, 4\_2\right\}$ layers from a pre-trained VGG-19 model $\theta\left(\cdot\right)$, the $CX\left(\cdot\right)$ function is the similarity measurement defined in [30].

4. Experiment and Result

4.1. Implementations

Dataset. We conducted experiments on two publicly available high-resolution video datasets for video-based human pose transfer, including FashionVideo [47] and iPER [24]. Both are collected from a human-centric manner with diverse garments, poses, viewpoints, and occlusion scenarios. The FashionVideo consists of 600 videos with around 350 frames per video. It is partitioned into 500 videos for training and 100 videos for testing. It is collected from a static camera and a clean white background. The iPER dataset contains 206 videos with roughly 1100 frames each. There are 164 videos for training and 42 videos for testing

Models	FashionVideo							iPER						
	SSIM↑	PSNR↑	l1↓	FID↓	LPIPS↓	FVD-Train128f↓	FVD-Test128f↓	SSIM↑	PSNR↑	L1↓	FID↓	LPIPS↓	FVD-Train128f↓	FVD-Test128f↓
Impersonator [24]	0.870	21.094	0.0498	18.220	0.0799	333.951±2.054	374.034±6.749	0.714	17.001	0.131	36.182	0.230	1034.278±17.523	1178.021±48.840
GFLA [35]	0.892	21.309	0.0459	16.308	0.0922	195.205±3.036	256.430±7.459	0.797	20.898	0.085	25.075	0.149	684.101±11.215	796.112±37.071
Impersonator++ [25]	0.873	21.434	0.0502	22.363	0.0761	197.668±2.309	175.663±5.857	0.755	18.689	0.103	33.629	0.173	714.519±13.813	742.394±30.208
DPTN [48]	0.907	23.996	0.0335	15.342	0.0603	215.078±2.252	206.345±6.522	0.742	17.997	0.110	34.204	0.209	1003.598±14.715	1143.603±33.631
NTED [34]	0.890	22.025	0.0425	14.263	0.0728	278.854±3.505	324.128±7.753	0.771	19.320	0.091	**20.164**	0.162	784.509±12.908	916.489±46.471
Ours	**0.918**	**24.071**	**0.0302**	**14.083**	**0.0478**	**168.275±2.564**	**148.253±6.781**	**0.803**	**21.797**	**0.0724**	22.291	**0.120**	**500.226±11.670**	**536.084±29.200**

Table 1. Quantitative comparisons with some state-of-the-art methods on the FashionVideo and iPER benchmarks. The best scores are highlighted in bold format.

Figure 4. Qualitative comparisons of pose transfer with some state-of-the-art methods on DanceFashion and iPER benchmarks. Please zoom in for more details.

purposes. Different focal lengths and genders are included to capture various poses and views in some indoor or natural backgrounds.

Evaluation metrics. To evaluate structural similarity, the SSIM [44] index is used to achieve this goal by applying covariance and mean. The PSNR computes the power of maximum value and its mean squared error. The L1 distance represents the pixel-wise fidelity. We also employ two supervised perceptual metrics including Fréchet Inception Distance (FID) [11] and Learned Perceptual Image Patch Similarity (LPIPS) [50]. The FID is used to measure the distribution disparity between the generated images and the training images by computing the perceptual distances. The LPIPS is targeted on evaluating the Wasserstein-2 distance between the distributions of the generated samples and real samples. To measure the temporal coherence, we utilize Fréchet Video Distance (FVD) [41] to extract features on time and feature space by a pre-trained I3D [3] network. It considers a distribution over the entire video, thereby avoiding the drawbacks of frame-level metrics. The term "FVD-Train128f" denotes the protocol of computing the FVD on randomly selected consecutive 128 frames for a sequence on training set and generated images with 50 iterations, likewise for "FVD-Test128f" on testing set.

Training strategy. We implement the proposed method with the public framework PyTorch. We adopt the Adam [17] optimizer with momentum $\beta_1 = 0.5$ and $\beta_2 = 0.999$ to train our model for $50,000$ iterations in total. The learning rate is set to 10^{-4}. To keep the original aspect ratio of the images, we resize the video frames to 256×256 by thumbnail approach. The negative slope of LeakyReLU [32] is set to 0.2. The weighting hyperparameters λ_{adv}, $\mathcal{L}_{temp}$, λ_1, λ_{per}, $\mathcal{L}_{gram}$, and λ_{cx} are set to 5, 5, 2, 500, 0.5, and 0.1. All models are trained and tested on an NVIDIA GeForce RTX 2080 Ti GPU with 11GB memory.

4.2. Comparison with SOTAs

To demonstrate the superiority of the proposed model, we compare our model with several state-of-the-art approaches including Impersonator [24], GFLA [35], Impersonator++ [25], DPTN [48], and NTED [34]. Due to different system environment settings, we reimplement these methods on the same machine based on their open-source repositories.

Quantitative Comparison. As shown in Table 1, our model achieves the best results on most evaluation metrics. The large margin of enhancement on FVD score indicates the best performance of our method in terms of spatio-temporal consistency. It represents the merits of the proposed bidirectionally deformable motion modulation in modulating long-range motion sequences with minimum discontinuity. Our model can achieve the best result on those image-based perceptual measuring metrics in the challenging FashionVideo dataset. For some images with natural backgrounds like those in iPER dataset, our model is also able to get highly competitive performance. It quantifies that our model has a better style transfer and video synthesis ability against current methods.

Qualitative Comparison. Apart from quantitative comparison, we also conduct a comprehensive qualitative measurement to compare the perceptually visual quality with the state-of-the-arts. We illustrate some generated results with various poses in Figure 4. We demonstrate a wide variety of viewpoints including front view, left side of body, right side of body, and back view on the Fashion dataset (row 1 – row 4). These results can highlight the superiority

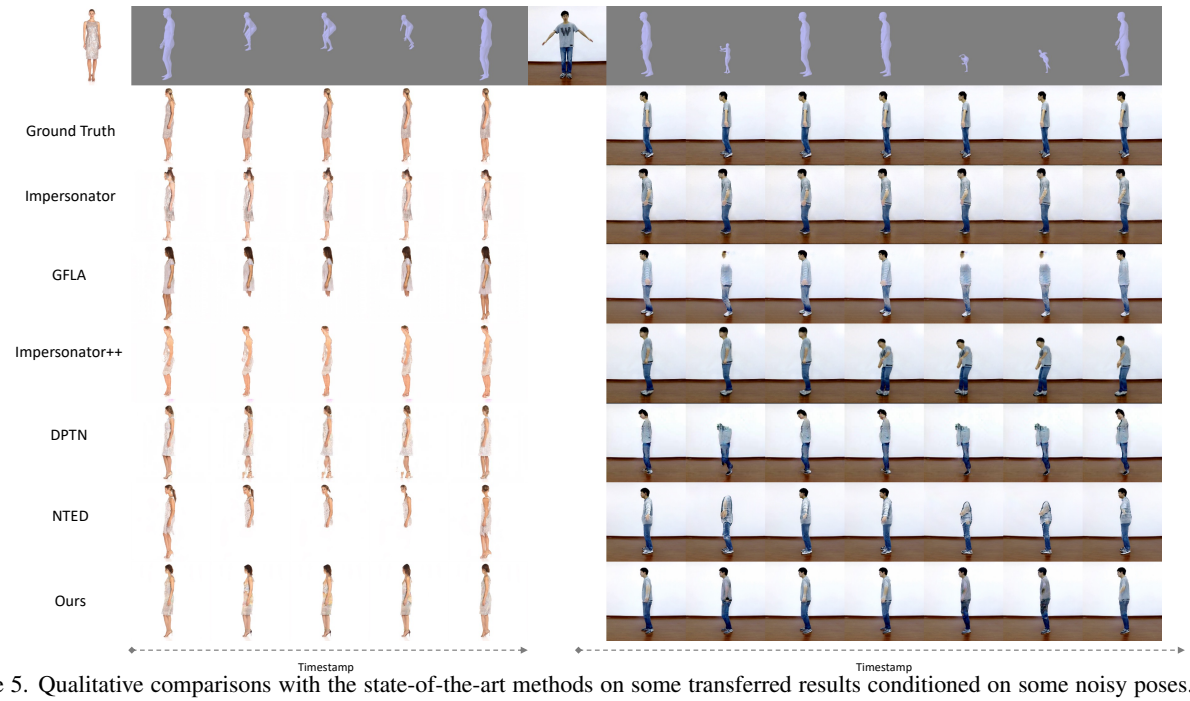

Figure 5. Qualitative comparisons with the state-of-the-art methods on some transferred results conditioned on some noisy poses. Noted that the input poses are evenly sampled from a random video clip. Please zoom in for more details.

Models	SSIM↑	PSNR↑	L1↓	FID↓	LPIPS↓	FVD-Train128f↓	FVD-Test128f↓
Deformable Motion Modulation							
w/o DMM	0.892	21.802	0.0435	16.636	0.0935	200.363±2.992	267.476±7.093
w/o DCN	0.916	23.849	0.0317	14.358	0.0484	187.390±2.617	167.529±7.566
w/o Style weight	0.914	23.483	0.0330	14.497	0.0513	176.107±2.745	161.549±7.245
w/o Feature concat	0.911	23.529	0.0338	14.933	0.0523	199.483±2.224	172.651±8.439
w/o Motion mask	0.912	23.492	0.0343	14.554	0.0519	191.984±2.336	176.882±7.004
Bidirectional Propagation							
w/o Forward propagation	0.914	23.715	0.0327	15.794	0.0510	208.354±2.833	188.869±9.678
w/o Backward propagation	0.908	23.179	0.0349	14.345	0.0538	171.649±2.289	156.469±6.671
w/o Recurrent structural flow	0.910	23.440	0.0337	15.951	0.0527	202.555±2.712	199.854±7.192
Ours	**0.918**	**24.071**	**0.0302**	**14.083**	**0.0478**	**168.275±2.564**	**148.253±6.781**

Table 2. Quantitative comparisons with some state-of-the-art methods on the FashionVideo and iPER benchmarks. The best scores are highlighted in bold format.

of our method from transferring person facial characteristics and complex texture on the garments in different points of view. To evaluate the synthesis quality in natural background, we present some generated results on uncommon gestures in iPER dataset (row 5 – row 8). It shows that our method can confidently handle arbitrary poses, shapes and backgrounds with minimum generated artifacts compared with others. It is benefited from the irregular field of view constructed by the deformable motion offset so that the multi-scale features can be effectively activated.

As a video-based solution, our method can generate temporally coherent sequences conditioned on some noisy poses without pre-processing, as shown in Figure 5. In general, the majority of structural guidance is hampered due to statistical outliers, especially in some occluded scenarios. It leads to an uncompleted shape and artifacts on the generated images, even though recurrent neural networks are applied in [24, 25, 35]. With the proposed bidirectional modulation

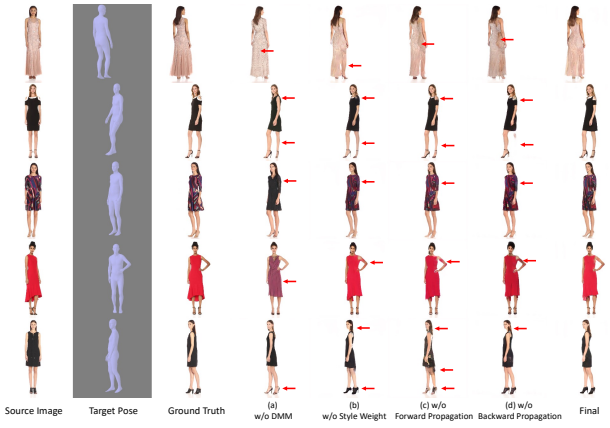

Figure 6. Quantitative results of ablation study. The best scores are highlighted with bold format.

mechanism, our method can synthesize smooth sequences with high-fidelity transferring effects.

4.3. Ablation Study

Deformable Motion Modulation. The proposed DMM is used to synthesize continuous frame sequences by modulating the 1D style code of source image with an augmentation of spatio-temporal sampling locations. We use a sum operation to fuse the features extracted from bidirectional propagation. Compared to the model *w/o DMM*, our model achieves superior results for all evaluation metrics in Table 2. The lack of style modulation mechanism leads to failure in style transferring results, in spite of the simple appearance style from the source image, as shown in Fig-

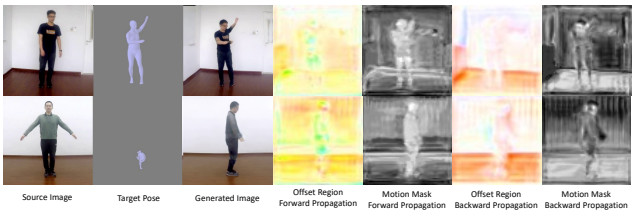

Figure 7. Demonstration on the region of interest for DMM. We highlight the activated regions of the estimated motion offsets and motion mask for both forward and backward propagation. Please zoom in for more details.

ure 6 (a). Moreover, based on the result of the model *w/o DCN*, we observe that there is a positive gain in synthesizing new content if the receptive field of view during convolution is expanded. It can achieve higher FID and LPIPS scores for image-based perception. The enhancement on FVD demonstrates the importance of capturing temporal information from adjacent frames. Furthermore, we compare the result with the model *w/o Style Weight*. The modulated style weight is an important component to perform affine transformation decomposed from style codes to structural poses. As depicted in Figure 6 (b)., the generated images are not with style consistence due to lack of a generalization on style transfer. It verifies that our proposed DMM can provide benefits to the fusion of style statistics so that it can minimize the distribution between real-world images and the synthesized.

Forward / Backward Propagation. The proposed bidirectional propagation mechanism is used to interpolate the probability of missing structural guidance from both forward and backward propagation flows in order to enhance the temporal consistency. The result of evaluation metrics in Table 2 reports that they both have an effective contribution in generating realistic images and maintaining temporal coherence between adjacent frames. In addition, the qualitative result in Figure 6 (c-d) demonstrates that the forward and backward propagation can preserve more details on structural shape and appearance details. Both comparisons on different measurements verify the efficacy of the proposed bidirectional propagation flow.

4.4. Visualization of DMM

The proposed DMM uses geometric kernel offset to transform regular receptive field of view to some irregular shapes [5, 51]. To investigate the effectiveness of the proposed deformable motion modulation, we illustrate some visualizations on the DMM module in feature space.

Region of Interest. The region of interest for DMM is to highlight the global area with effective motion offsets and motion mask. As demonstrated in Figure 7, we plot the total kernel offsets as a kind of optical flow by following [1] so that we can observe the activated regions of interest in each propagation branch. The visualizations for the motion mask also highlight the activated magnitude along with the motion offsets. It is reasonable that the motion offsets and masks are not aligned for both forward and backward branches because they are designed to capture the temporal information in two different sequences. Based on the global shape on the offset regions and masks in both forward and backward propagation, we can clearly point out the human body shape with a predictable movement. The regions with more semantic information are with higher density. The activated regions provide geometric guidance for the network to modulate the style code extracted from the source image.

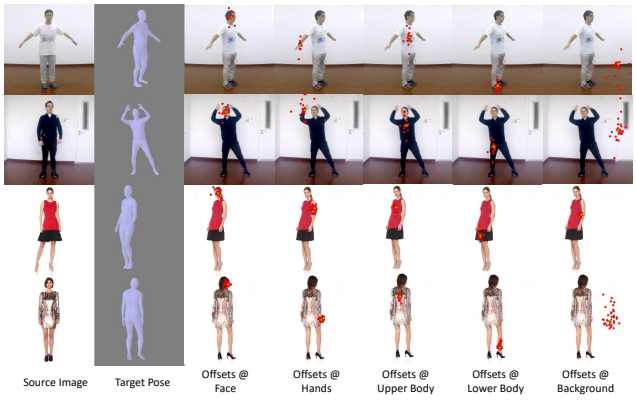

Figure 8. Demonstration on the motion offset applied on the activated units for DMM. The green points and red points represents the activation units for the corresponding augmented sampling locations. Please zoom in for more details.

Activated Unit. The success of the proposed DMM relies on the augmentation of spatio-temporal sampling locations. We visualize the behavior of the deformable filters in Figure 8. The activation units are highlighted with green points and red points for the corresponding augmented sampling locations. It is obvious that the proposed semantics on the sampling locations are dependent on the activated units. It is certified that the proposed DMM module can produce a dynamic receptive field of view to keep track of interested semantics so that it can synthesize a sequence of high-quality and smooth video frames.

4.5. Conclusion

In this paper, we present a novel end-to-end framework for video-based human pose transfer. The proposed Deformable Motion Modulation (DMM) employs geometric kernel offsets with adaptive weight modulation to perform spatio-temporal alignment and style transfer concurrently. The bidirectional propagation is employed to strengthen the temporal coherence. Comprehensive experimental results show that our method can effectively deal with the problems of spatial misalignment for complex structural patterns and noisy poses. Our framework has an excellent synthesis ability in human pose video generation and has great research potential for industrial development.

References

[1] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International journal of computer vision*, 92(1):1–31, 2011. 8

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1, 4

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[4] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5972–5981, 2022. 3

[5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 3, 4, 8

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[7] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. *Advances in neural information processing systems*, 31, 2018. 1, 2

[8] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9026–9035, 2019. 1

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3, 5

[10] 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. 1, 4

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[13] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian McAuley. Visually-aware fashion recommendation and design with generative image models. In *2017 IEEE international conference on data mining (ICDM)*, pages 207–216. IEEE, 2017. 1

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 4

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 4

[16] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3907–3916, 2018. 4

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[18] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 4

[19] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995. 3

[20] Chenyi Lei, Dong Liu, Weiping Li, Zheng-Jun Zha, and Houqiang Li. Comparative deep learning of hybrid representations for image recommendations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2545–2553, 2016. 1

[21] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019. 1, 4

[22] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019. 1, 2

[23] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17562–17571, 2022. 3

[24] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5904–5913, 2019. 1, 2, 5, 6, 7

[25] Wen Liu, Zhixin Piao, Zhi Tu, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan with attention: A unified framework for human image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2, 4, 5, 6, 7

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 4

[27] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10806–10815, 2021. 1, 2

[28] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in neural information processing systems*, 30, 2017. 1, 2

[29] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 5

[30] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 768–783, 2018. 5

[31] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[32] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010. 6

[33] Assaf Neuberger, Eran Borenstein, Bar Hilleli, Eduard Oks, and Sharon Alpert. Image based virtual try-on network from unpaired data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5184–5193, 2020. 1

[34] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural texture extraction and distribution for controllable person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13535–13544, 2022. 1, 2, 6

[35] Yurui Ren, Ge Li, Shan Liu, and Thomas H Li. Deep spatial transformation for pose-guided person image generation and animation. *IEEE Transactions on Image Processing*, 29:8622–8635, 2020. 1, 2, 3, 5, 6, 7

[36] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019. 3

[37] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[38] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[39] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In *European Conference on Computer Vision*, pages 717–734. Springer, 2020. 1, 2

[40] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3360–3369, 2020. 3

[41] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

[42] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 589–604, 2018. 1

[43] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. *arXiv preprint arXiv:1910.12713*, 2019. 3

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[45] Lingbo Yang, Pan Wang, Xinfeng Zhang, Shanshe Wang, Zhanning Gao, Peiran Ren, Xuansong Xie, Siwei Ma, and Wen Gao. Region-adaptive texture enhancement for detailed person image synthesis. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 1, 2

[46] Wing-Yin Yu, Lai-Man Po, Yuzhi Zhao, Jingjing Xiong, and Kin-Wai Lau. Spatial content alignment for pose transfer. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 1, 2

[47] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*, 2019. 5

[48] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7713–7722, 2022. 1, 2, 6

[49] Quan Zhang, Jianhuang Lai, Zhanxiang Feng, and Xiaohua Xie. Seeing like a human: Asynchronous learning with dynamic progressive refinement for person re-identification. *IEEE Transactions on Image Processing*, 31:352–365, 2021. 1

[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[51] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019. 3, 4, 8

[52] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019. 1, 2

Supplementary Materials

Anonymous CVPR submission

Paper ID 6538

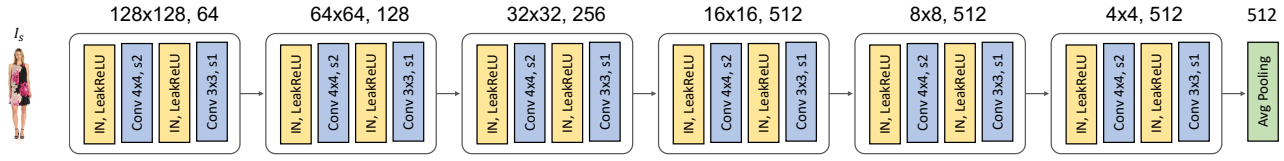

Figure 1. Network architecture of Style Encoder.

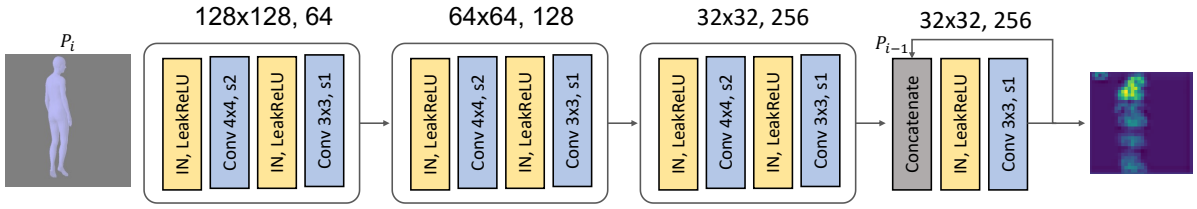

Figure 2. Network architecture of Structural Encoder.

1. Network Architecture

We present the design details of the proposed network architecture for different components. To be consistent, the resolution of all the input is 256×256. The input channels are set to 3 for a RGB image and 18 for a structural pose map. As depicted in Figure 1– 5, to simplify the notations, we use "IN" to represent Instance Normalization [5], "Conv $k \times k$, s#" to represent a convolutional layer with kernel size $k \times k$ and stride #. For example, "Conv 4×4, $s2$" indicates kernel size 4×4 and stride 2. With appropriate padding, we set the convolution layer with stride 2 to down-scale the features to half of the input resolution.

1.1. Encoder

Style Encoder. The Style Encoder is designed to extract style code of the source image which is a vector that consists of dense semantic features from the source image. As shown in Figure 1, it includes 6 encoder blocks that progressively downsample the input features from 256×256 to 4×4. At the bottleneck of the encoder, we use an adaptive average pooling layer with kernel size 4×4 to compute the style vector.

Structural Encoder. The Structural Encoder is used to encode the spatial details of the target pose and shape so that it can produce a spatially aligned content in the final output image. Apart from cascaded convolutional blocks, we also leverage a recurrent flow at the bottleneck to maintain spatio-temporal information, as indicated in Figure 2. In this recurrent operation, we concatenate the input features and the output with same resolution. We visualize an example of output features in the Figure 2. The regions with key structural guidance such as eyes, hands or legs are well highlighted. The features with fading effect represent the hidden motion information. It indicates the effectiveness on extracting temporal information of the proposed recurrent Structural Encoder.

1.2. Discriminator

The discriminator is an essential element in our network to formulate the adversarial loss. As shown in Figure 3, there are two discriminators including the Spatial Discriminator and Temporal Discriminator in our framework. During training implementation, we randomly select a frame (4D tensor) from a mini-batch to calculate the spatial adversarial loss while using the whole mini-batch (5D tensor) to compute the temporal adversarial loss.

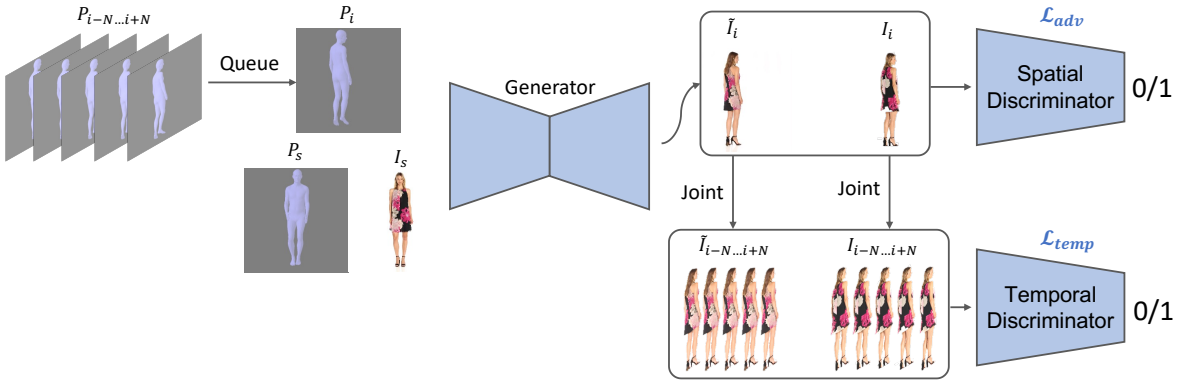

Figure 3. Overview of network architecture, including Spatial Discriminator and Temporal Discriminator.

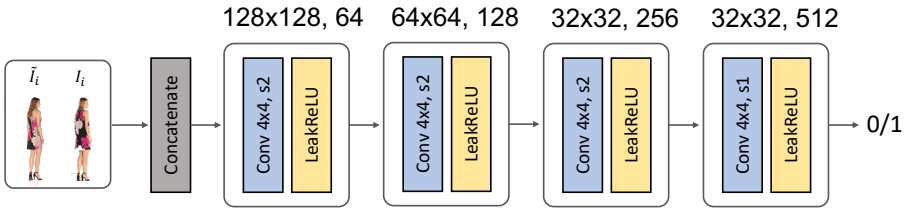

Figure 4. Network architecture of Spatial Discriminator.

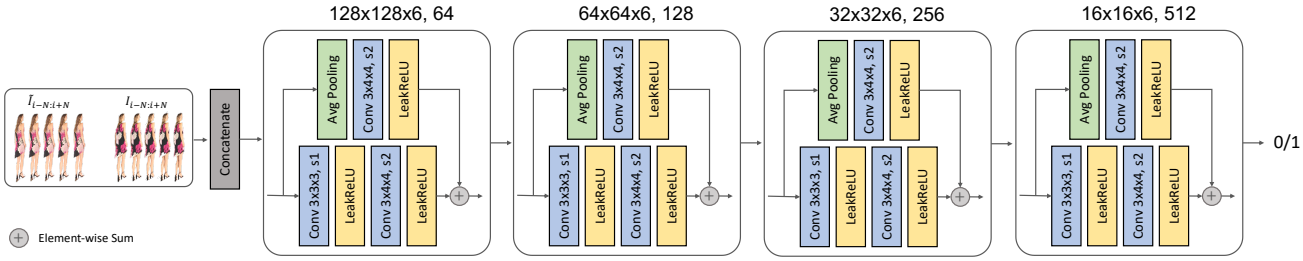

Figure 5. Network architecture of Temporal Discriminator.

Spatial Discriminator. The Spatial Discriminator is used to mimic the distribution of the training set by discriminating whether the input pair is real or fake. We demonstrate the network architecture in Figure 4. Different from traditional GANs that using a single image as the input, we concatenate a generated image and a ground truth by channel dimension as a paired input, like PatchGAN [1]. There are 3 encoder blocks that progressively to reduce the resolution of input features from 256×256 to 32×32. We use a convolution layer with kernel size 4×4 and the LeaklyReLU [4] activation function to extract the patched features. Finally, we apply the least square error [3] to compute the statistical distance.

Temporal Discriminator. The temporal Discriminator is used to optimize the temporal consistency in time and feature channels of a mini-batch by using a 3D CNN model. During training, we collect the output image $\widetilde{I}_i$ one by one from time step $i - N$ to $i + N$. Similar with the Spatial Discriminator that concatenating the paired input images, we concatenate the generated sequence $\widetilde{I}_{i-N:i+N}$ and the target sequence $I_{i-N:i+N}$ by channel dimension as input. As indicated in Figure 5, there are 4 encoder blocks that progressively to downsample the input features from 256×256 to 16×16. The N is equal to 3 if the total iteration sequence length is 6. The major design of the encoder block is similar with one of Spatial Discriminator. We use 3D convolutional layer with kernel size $3 \times 3 \times 3$ or $3 \times 3 \times 4$ to downscale the input features by half. Moreover, we apply the average pooling layer to produce a downsampled residual map to preserve the feature signals. We use the weighted sum to fuse the output features and the residual branch.

Models	SSIM↑	PSNR↑	L1↓	FID↓	LPIPS↓	FVD-Train128f↓	FVD-Test128f↓
w/o L1 loss	N/A	15.147	0.117	379.232	0.388	2575.616±14.795	2626.936±29.864
w/o perceptual loss	0.914	23.767	0.0323	15.336	0.0518	174.383±2.413	159.217±6.818
w/o style loss	0.905	22.933	0.0373	14.956	0.0569	180.374±2.169	172.143±15.127
w/o CX loss	0.908	23.179	0.0349	14.345	0.0538	171.824±2.143	155.876±7.937
w/o spatial adv loss	0.913	23.576	0.0329	14.394	0.0506	201.065±3.101	187.118±8.642
w/o temporal adv loss	0.916	23.892	0.0312	14.466	0.0487	178.727±2.317	165.524±7.654
Final	**0.918**	**24.071**	**0.0302**	**14.083**	**0.0478**	**168.275±2.564**	**148.253±6.781**

Table 1. Quantitative ablation study on the objective loss functions evaluated on the FashionVideo benchmark. The best scores are highlighted in **bold** format.

2. Ablation Study on Loss Function

We conduct a comprehensive quantitative experiment on the analysis of objective loss functions in Table 1. We formulate the loss functions in three domains - pixel domain, semantic domain, and spatio-temporal domain to not only synthesize high-fidelity person image but also maintain details of person identity with characteristics of garments in the source image. The evaluation protocol is designed to observe the importance of each loss function by excluding the target function.

2.1. Pixel Domain

We mainly use L1 loss to minimize the absolute value of pixel distance between the generated image and the target image. The worst results on the model *w/o L1 loss* indicates the crucial role on generating acceptable images in our model. It is because pixel-wise comparison can preserve more global statistics such as basic appearance and shape of a person.

2.2. Semantic Domain

The losses on semantic domain are used to enhance the vividness of the generated images by comparing the features with different correspondence operations. It includes model *w/o perceptual loss*, model *w/o style loss*, and model *w/o CX loss*. The outcome shows that they all provide positive gains to the evaluation metrics in different aspects. The model *w/o perceptual loss* shows an 8% increment on the FID score. It represents the effectiveness on minimize the distribution distance between the generated results and the training set. The model *w/o style loss* and model *w/o CX loss* have major contributions on SSIM, PSNR, L1, and LPIPS scores. It indicates that these two losses can maintain more structural details on the generated images.

2.3. Spatio-temporal Domain

We mainly use adversarial losses to strengthen the spatio-temporal consistency of the generated sequence. The results of model *w/o spatial adv loss* and model *w/o temporal adv loss* demonstrates a large margin on the FVD-train128f and FVD-test128f scores compared to the final model. It can certify that these two losses can maintain spatio-temporally coherent information in our framework.

3. Video Comparison for SOTAs

We randomly select some video demonstrations to compare the visual quality with some state-of-the-art methods. The videos are from FashionVideo [6] and iPER [2] benchmarks. Please find the attached cvpr_video_supplementary.zip file to enjoy the video clips. The default frame rate is 30fps.

References

[1] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[2] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5904–5913, 2019. 3

[3] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 2

[4] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010. 2

[5] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 1

[6] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*, 2019. 3