C2F-FWN: Coarse-to-Fine Flow Warping Network for Spatial-Temporal Consistent Motion Transfer

Dongxu Wei[1], **Xiaowei Xu**[2], **Haibin Shen**[1], **Kejie Huang**[1*]

[1] Department of Information Science and Electronic Engineering, Zhejiang University
[2] Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences

Abstract

Human video motion transfer (HVMT) aims to synthesize videos that one person imitates other persons' actions. Although existing GAN-based HVMT methods have achieved great success, they either fail to preserve appearance details due to the loss of spatial consistency between synthesized and exemplary images, or generate incoherent video results due to the lack of temporal consistency among video frames. In this paper, we propose Coarse-to-Fine Flow Warping Network (C2F-FWN) for spatial-temporal consistent HVMT. Particularly, C2F-FWN utilizes coarse-to-fine flow warping and Layout-Constrained Deformable Convolution (LC-DConv) to improve spatial consistency, and employs Flow Temporal Consistency (FTC) Loss to enhance temporal consistency. In addition, provided with multi-source appearance inputs, C2F-FWN can support appearance attribute editing with great flexibility and efficiency. Besides public datasets, we also collected a large-scale HVMT dataset named SoloDance for evaluation. Extensive experiments conducted on our SoloDance dataset and the iPER dataset show that our approach outperforms state-of-art HVMT methods in terms of both spatial and temporal consistency. Source code and the SoloDance dataset are available at https://github.com/wswdx/C2F-FWN.

Introduction

Human Video Motion Transfer (HVMT) refers to the task of synthesizing videos that one person imitates motions of other persons, which has attractive potential applications in movies, interactive games, virtual shopping, etc. With the development of Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) and GAN-based image-to-image translation techniques (Wang et al. 2018b,a; Park et al. 2019), HVMT works have achieved great success.

In general, existing HVMT methods have two main streams: personalized HVMT and general-purpose HVMT. Personalized methods (Chan et al. 2019; Liu et al. 2019a) focus on learning the mapping from motion inputs (e.g., body poses or semantic layouts that describe the desired motions) to video frames for a specific person, with a large number of frames from this person collected as the training data to fit the model for his/her appearance. To generate videos for another person, they have to perform a new round of data collection and model training, which requires massive human resources and computation costs. The recent emergence of general-purpose methods (Wang et al. 2019; Liu et al. 2019b; Wei et al. 2020) manages to solve this by providing additional appearance inputs (e.g., exemplary images that describe the desired appearances) for GANs. Thus they can generate videos for new persons by altering the input exemplary images. However, most of these methods directly utilize GANs to generate values for all the pixels from scratch without preserving their spatial consistency with pixels in the exemplary images, which results in the loss of appearance details such as decorative patterns and colors of clothes. Besides, they either don't consider temporal consistency or only focus on implicit temporal consistency among frame images when synthesizing videos, which causes low temporal coherence in their video results. Moreover, most of them don't support HVMT with fully editable appearances, lacking flexibility and efficiency for real applications.

In this paper, to address these limitations, we propose Coarse-to-Fine Flow Warping Network (C2F-FWN) to ensure both spatial and temporal consistency for HVMT. For spatial consistency, our C2F-FWN synthesizes motion transfer videos through warping based on coarse-to-fine transformation flows rather than direct generation based on GANs. Thus we can precisely model geometric deformations caused by motions to preserve spatial correlations between synthesized and exemplary image pixels. Moreover, Layout-Constrained Deformable Convolution (LC-DConv) is utilized to extract deformable features for C2F-FWN, further improving the spatial consistency. For temporal consistency, we propose Flow Temporal Consistency (FTC) Loss with optical flows as the constraints to enforce explicit temporal consistency among transformation flows instead of frame images, radically ensuring the video coherence.

In our experiments, we evaluate our method on both iPER dataset (Liu et al. 2019b) and a large-scale SoloDance dataset collected by ourselves. Both quantitative and qualitative results demonstrate that videos generated by our method have significantly better spatial and temporal consistency

[*] Corresponding author: huangkejie@zju.edu.cn

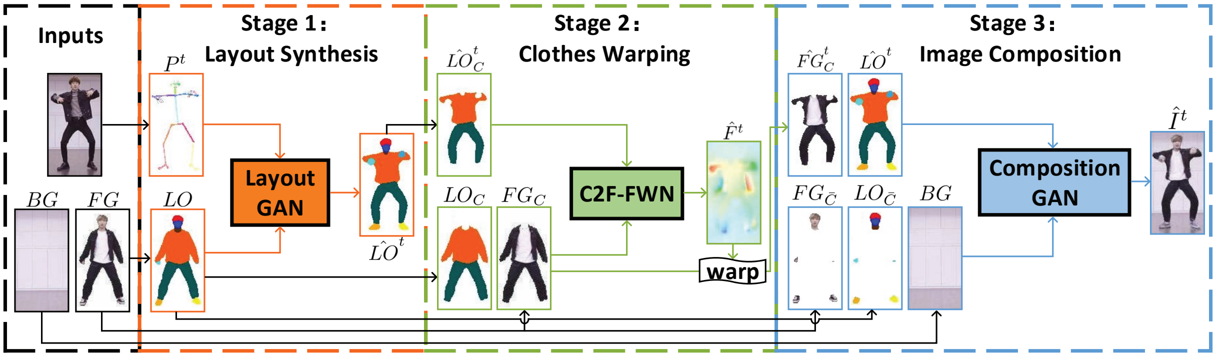

Figure 1: Overview of our method. Orange, green and blue rectangles specify processes in Stages 1, 2 and 3, respectively. Black arrows denote ordinary data flows like pose detection, layout detection and body division. Orange, green and blue arrows denote data flows of the layout GAN in Stage 1, our C2F-FWN in Stage 2, and the composition GAN in Stage 3, respectively.

than existing personalized and general-purpose methods. We also show that our approach can utilize multi-source appearance inputs to enable full appearance attribute editing (e.g., change identities, tops, bottoms, backgrounds) for HVMT, which has promising application prospects.

Related Work

Personalized HVMT

Personalized HVMT (Chan et al. 2019; Liu et al. 2019a; Aberman et al. 2019; Yang et al. 2020) only learns the mappings from motion inputs to video frames, with appearances learned individually in different models. Once trained, one model can only generate videos with specific appearances. To generate videos with new appearances, they need to train new models. Although such approaches can generate high-fidelity videos, they lack the efficiency for applications.

General-Purpose HVMT

General-purpose HVMT can be divided into direct generation methods (Wang et al. 2019; Wei et al. 2020) and warping-based methods (Liu et al. 2019b; Dong et al. 2018; Han et al. 2019). Both utilize additional appearance inputs to control the synthesized appearances in addition to motions.

Direct Generation Methods leverage GAN-based image-to-image translation techniques (Wang et al. 2018b,a; Park et al. 2019) to generate video frames from appearance and motion inputs directly. Wang et al. utilize several SPADE blocks (Park et al. 2019) to adaptively propagate the appearance information throughout the network, which can achieve appearance control by altering the appearance inputs. Wei et al. propose an appearance-consistency discriminator to force the generator to generate appearances consistent with the alterable appearance inputs, which also achieves the appearance control. However, these methods don't consider the spatial consistency between pixels in outputs and appearance inputs. Thus they can't preserve appearance details such as textures and colors well. Moreover, they only consider temporal consistency among frame images, which is implicit and hard to learn. On account of mode collapse and over-fitting problems of GANs (Webster et al. 2019), these methods often obtain low-fidelity results.

Warping-Based Methods focus on generating images through warping to preserve spatial consistency. Dong et al. utilize Thin-Plate-Spline (TPS) transformation for warping to align features of appearance inputs with those of motion inputs before GAN-based generation. Similar feature warping can also yield fancy facial animation results for face video synthesis (Chen et al. 2020). Unfortunately, the TPS transformation is decided by a few control points, which restricts its warping capability due to the low degree of freedom. Thus it can't precisely model the geometric deformations. Moreover, Liu et al. propose liquid warping based on 3D SMPL models to achieve similar feature alignment. However, the SMPL models (Loper et al. 2015) only describe naked human bodies. Thus they can't model surface deformations for clothes and hair. Instead of warping features, Han et al. propose to warp images using flows, which is similar to our approach in spirit. However, they directly estimate the dense flow field from misaligned features of appearance and motion inputs, failing to model large deformations when the two inputs greatly differ from each other in motion. Besides, none of these warping-based methods considers the temporal consistency between warping operations of neighbored frames, making them not capable of synthesizing coherent videos. Moreover, these methods all adopt standard convolutions in their networks, where the fixed receptive fields can't accommodate shape variances. Hence they can't extract appropriate features for human subjects with various shapes to estimate warping functions.

Method

Overview

The overview of our method is shown in Figure 1, which contains three stages: layout synthesis (Stage 1), clothes warping (Stage 2) and image composition (Stage 3). For ease of discussion, the used symbols are presented as follows. Given an exemplary foreground image FG describing the desired human appearance and an exemplary background image BG describing the desired background appearance, we aim at synthesizing videos that the exemplary human foreground FG performs motions described by the pose sequence $P^{1\sim T}$ in the exemplary background BG. For the exemplary foreground FG, we detect its semantic layout LO and divide it to further obtain layout LO_C and foreground FG_C for the clothing parts (i.e., tops and bottoms), as well as layout $LO_{\bar{C}}$ and foreground $FG_{\bar{C}}$ for the non-clothing parts (i.e., hair, face, torso and shoes). Provided with these processed inputs, we synthesize the corresponding output video sequence $\hat{I}^{1\sim T}$ by three stages described in Figure 1. Since $\hat{I}^{1\sim T}$ is generated frame by frame, we take the synthesis of the t-th frame $\hat{I}^t$ as an example for brevity.

In Stage 1, we utilize a layout GAN to generate the semantic layout $\hat{LO}^t$, which has the same motion as P^t and the same appearance as LO. We further divide $\hat{LO}^t$ to obtain the clothing layout $\hat{LO}_C^t$. **In Stage 2**, taking $\chi_1 = \{LO_C, FG_C\}$ as the appearance input and taking $\chi_2 = \hat{LO}_C^t$ as the motion input, our C2F-FWN computes the transformation flow $\hat{F}^t$ to warp the exemplary clothing foreground FG_C into the foreground $\hat{FG}_C^t$, which precisely aligns with the generated clothing layout $\hat{LO}_C^t$. **In Stage 3**, we utilize a composition GAN to generate the remaining parts including the non-clothing foreground and the background, and compose them with the clothing foreground $\hat{FG}_C^t$ from Stage 2 to generate the full frame image $\hat{I}^t$. Note that we don't generate the non-clothing parts through warping for two reasons. First, the appearance of the non-clothing parts varies sharply in different views, making it extremely hard to model their appearance changes through warping. Second, texture and color patterns of the non-clothing parts are simple and easy to generate using GANs. Therefore, we utilize the composition GAN to synthesize the non-clothing parts.

Particularly, the layout GAN and the composition GAN follow the Vid2Vid design presented in (Wang et al. 2018a). Vid2Vid is a general image-to-image translation backbone consisting of two encoders and two decoders (E_1,E_2,D_1,D_2 for brevity). E_1 and E_2 aim to encode features for two inputs $\mathcal{I}_1$ (i.e., current conditional inputs) and $\mathcal{I}_2$ (i.e., previous generated results), respectively. D_1 and D_2 aim to decode the added features of $\mathcal{I}_1$ and $\mathcal{I}_2$ to output $\mathcal{O}_1$ (i.e., a raw result) and $\mathcal{O}_2$ (i.e., an optical flow). Then we can obtain the current frame result by using $\mathcal{O}_2$ to warp the last frame result and add it to $\mathcal{O}_1$. For the **layout GAN**, $\mathcal{I}_1$ denotes the concatenated $\{P^t, LO\}$. $\mathcal{I}_2$ denotes the concatenated $\{\hat{LO}^{t-1}, \hat{LO}^{t-2}\}$. Thus we can utilize the Vid2Vid backbone to generate $\hat{LO}^t$. Besides, to better synthesize the one-hot semantic layouts rather than RGB images, we replaced image reconstruction losses of Vid2Vid with a structure-sensitive pixel-wise softmax loss introduced in human parsing works (Liang et al. 2018). Similarly, for the **composition GAN**, $\mathcal{I}_1$ denotes the concatenated $\{\hat{LO}^t, \hat{FG}_C^t, LO_{\bar{C}}, FG_{\bar{C}}, BG\}$. $\mathcal{I}_2$ denotes the concatenated $\{\hat{I}^{t-1}, \hat{I}^{t-2}\}$. The Vid2Vid backbone learns to automatically attach non-clothing parts to clothes synthesized in Stage 2, and thus obtain the full image $\hat{I}^t$.

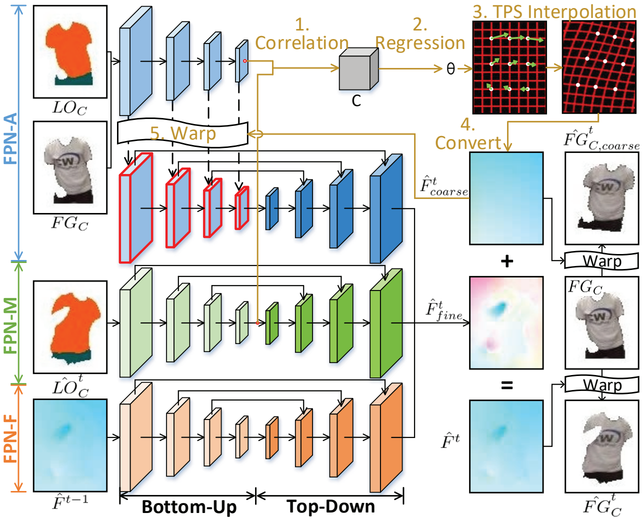

Figure 2: Illustration of C2F-FWN. We use feature maps in different colors to differentiate the three FPNs, where blue, green, and orange feature maps specify FPN-A, FPN-M, and FPN-F, respectively. Each FPN has two pathways connected by lateral connections (horizontal arrows), where we use light and dark colors to differentiate the features of bottom-up and top-down pathways, respectively. Steps 1∼5 drawn in gold describe the procedure of our coarse flow warping.

In the following, the details of our **C2F-FWN** including coarse-to-fine flow warping, Layout-Constrained Deformable Convolution (LC-DConv) and Flow Temporal Consistency (FTC) Loss are presented. At last, the unique characteristic of our C2F-FWN, multi-source appearance attribute editing is discussed.

Coarse-to-Fine Flow Warping

Before diving into details of the coarse-to-fine flow warping, we first explain its motivation and mechanism. In cases that LO_C greatly differs from $\hat{LO}_C^t$ in motion, there would be far distances between pixels in the appearance input χ_1 and their correlated pixels in the motion input χ_2. If we estimate the whole transformation flow directly based on the concatenation of misaligned appearance and motion features, we would fail because standard convolutions with limited kernel sizes can't build correlations between pixels far away from each other in position. Differently, our C2F-FWN first estimates a coarse Thin-Plate-Spline (TPS) flow $\hat{F}^t_{coarse}$ based on the smallest bottom-up features to coarsely warp the appearance features into the desired motion, where the effects of far distances w.r.t. the size of inputs can be ignored due to the large receptive fields of small-size features. Thus the appearance and the motion features are aligned. Then we can concatenate the largest top-down appearance and motion features to further compute the refinement flow $\hat{F}^t_{fine}$ for fine warping, where the effects of far distances have been eliminated after preliminary feature alignment.

As shown in Figure 2, C2F-FWN contains three feature pyramid networks (FPN) (Lin et al. 2017): FPN-A, FPN-M and FPN-F, responsible for extracting pyramidal features of appearance input $\chi_1 = \{LO_C, FG_C\}$, motion input $\chi_2 = \hat{LO}^t_C$ and previously estimated transformation flow $\chi_3 = \hat{F}^{t-1}$. Each FPN has two symmetrical pathways (bottom-up and top-down). Specifically, the top-down pathway is built upon the bottom-up pathway via lateral connections, with sizes of the bottom-up features growing smaller and sizes of the top-down features growing larger. Benefiting from such symmetric design, both coarse and fine warpings can be realized in the unified C2F-FWN.

Coarse Flow Warping The procedure of our coarse flow warping is described in steps 1∼5 in Figure 2. First, we compute the correlation map C with each position containing the pairwise similarities between the smallest bottom-up features of FPN-A and FPN-M. The correlation map is then fed into a regression layer to compute $K \times 2$ parameters (θ), which represent positions of K control points (K=3×3 in this paper). Based on TPS interpolation (Rocco, Arandjelovic, and Sivic 2017), we can generalize the mapping between the estimated K control points and their corresponding predefined grid points to all the pixels of FG_C, and hence move each pixel to its new position to obtain the coarsely warped clothes $\hat{FG}^t_{C,coarse}$. To enable the supervision of the coarse warping, we utilize a VGG loss (Johnson, Alahi, and Fei-Fei 2016) L^{coarse}_{VGG} to minimize the difference between $\hat{FG}^t_{C,coarse}$ and the ground truth.

Then, to make the TPS transformation compatible with our transformation flow, we convert it to a coarse flow $\hat{F}^t_{coarse}$ by computing the position difference before and after transformation for each pixel. Let $P = (x, y)$ denote the position of a pixel in the warped clothes $\hat{FG}^t_{C,coarse}$, and let $P^{'} = (x^{'}, y^{'})$ denote the position of the same pixel in the exemplary clothes FG_C. The coarse flow at position (x, y) can be given by: $\hat{F}^t_{coarse}(x, y) = \overrightarrow{PP^{'}} = (x^{'} - x, y^{'} - y)$, which is the same for all the other positions.

Then we downsample $\hat{F}^t_{coarse}$ to different sizes to warp all the bottom-up features of FPN-A, roughly aligning them with the generated layout $\hat{LO}^t_C$, which represents the desired motion. Thus we can compute the corresponding roughly-aligned top-down appearance features via lateral connections, with pixels located at positions close to the correlated pixels in the top-down motion features, facilitating the subsequent estimation of the refinement flow.

Fine Flow Warping As shown in Figure 2, we predict the refinement flow $\hat{F}^t_{fine}$ based on the concatenation of the largest top-down features of the three FPNs, where we include features of FPN-F to allow for learning the temporal consistency with previous transformation flows. Specifically, the refinement flow $\hat{F}^t_{fine}$ has the same size as the coarse flow $\hat{F}^t_{coarse}$, adding pixel-wise offsets to $\hat{F}^t_{coarse}$ to precisely align with the generated layout $\hat{LO}^t_C$. Thus our final transformation flow is given by: $\hat{F}^t = \hat{F}^t_{coarse} + \hat{F}^t_{fine}$. Using $\hat{F}^t$ to warp the exemplary clothes FG_C, we can obtain the final warped clothes $\hat{FG}^t_C$.

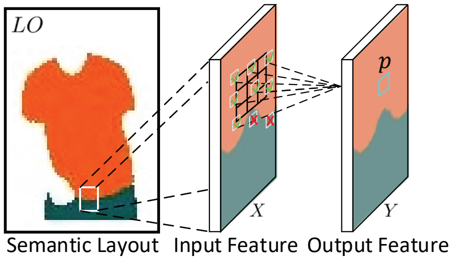

Figure 3: Illustration of LC-DConv. Here we take the LC-DConv in the first layer of FPN-M as an example, which is the same for FPN-A expect FPN-A takes foregrounds in addition to layouts as its inputs. We let orange and dark green represent semantic classes of tops and bottoms, respectively. In the small patch, ✓ in green and × in red denote valid and invalid sampling positions, respectively.

During training, we also utilize a VGG loss (2016) L_{VGG} to minimize the difference between $\hat{FG}^t_C$ and the ground truth, which enables the supervision of the fine warping.

Layout-Constrained Deformable Convolution

Since both the coarse and the refinement flows are predicted based on FPN-A and FPN-M features, feature extraction in these two FPNs directly affects the quality of our warping results. In motion transfer tasks, clothes items may change to various shapes along with body poses. Thus the extracted features should be able to generalize to various shapes correspondingly. Unfortunately, standard CNN features are transformation-invariant, which means receptive fields remain fixed no matter how the shape changes and hence can't accommodate the geometric deformations for different shapes. Besides, such fixed receptive fields are not large enough to accommodate the misalignment between appearance and motion features.

Therefore, we replace all the standard convolutions in bottom-up pathway layers of FPN-A and FPN-M with Deformable Convolutions (DConv) (Dai et al. 2017), which can model geometric deformations adaptively with deformable receptive fields. As shown in Figure 3, DConv learns additional 2D offsets to shift regular sampling locations of the standard convolution, which enables deformable and larger receptive fields. However, the unconstrained offsets may result in invalid sampling from positions not semantically related to the output position, causing the loss of semantic information in the output feature. Therefore, our Layout-Constrained Deformable Convolution (LC-DConv) utilizes input semantic layouts as priors to set amplitudes of features sampled from invalid positions to zero, precisely preserving the layout boundaries and thus enhancing the semantic information in the output feature. Taking the convolution with a 3x3 kernel of dilation 1 as an example, we explain how our LC-DConv works. Let $X(p)$ and $Y(p)$ be the input and the output features at position p respectively, and let w_k, Δp_k and $p_k \in \{(-1, -1), (-1, 0), ..., (1, 1)\}$, $k = 1$∼9 represent

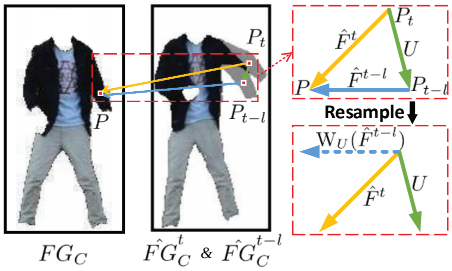

Figure 4: Illustration of FTC loss. The first image is the exemplary clothes FG_C. The second image is a combination of two warped clothes $\hat{FG}_C^{t-l}$ and $\hat{FG}_C^{t}$. We take the position P in the right arm region as an example, which moves from P_{t-l} to P_t during time $t-l{\sim}t$. Yellow, blue and green solid arrows represent $\hat{F}^t$, $\hat{F}^{t-1}$ and U, respectively. Blue dotted arrow denotes the resampled $\hat{F}^{t-1}$.

the k-th kernel weight, the k-th sampling offset and the k-th regular sampling position respectively, we can derive the LC-DConv as follows:

$$
\begin{aligned}
&Y(p) = \sum_{k=1}^{K} w_k \cdot X(p + p_k + \Delta p_k) \cdot \Delta m_k, \\
&\Delta m_k = \begin{cases} 0, & LO(p) \neq LO(p + p_k + \Delta p_k), \\ 1, & \text{otherwise}, \end{cases}
\end{aligned}
\tag{1}
$$

where $K = 9$, Δm_k is the modulation scalar determined by the layout prior LO, deciding the validity of the k-th offset sampling position. For FPN-A, LO refers to the exemplary clothing layout LO_C. For FPN-M, LO refers to the generated clothing layout $\hat{LO}_C^t$. As depicted in Figure 3, we set feature amplitudes to zero if they belong to semantic classes different from the class at the output position, which can effectively avoid any invalid sampling.

Flow Temporal Consistency Loss

Compared to other methods (Chan et al. 2019; Wang et al. 2019) that only learn implicit temporal consistency among frame images, our FTC loss uses optical flows to enforce explicit temporal consistency among transformation flows, also enabling direct supervision on transformation flows instead of warped clothes. Specifically, benefiting from the flow format of our transformation, we can build the correlation between two transformation flows $\hat{F}^t$ and $\hat{F}^{t-l}$ using the optical flow U between the corresponding two frames. Let $P = (x, y)$ denote the position of a pixel in the exemplary clothes FG_C, and let $P_t = (x_t, y_t)$ and $P_{t-l} = (x_{t-l}, y_{t-l})$ denote positions of the same pixel in the warped clothes $\hat{FG}_C^t$ and $\hat{FG}_C^{t-l}$. Thus, for this pixel, the transformation flow vectors at time t and $t-l$ are given by:

$$
\begin{aligned}
\hat{F}^t(x_t, y_t) &= (x - x_t, y - y_t), \\
\hat{F}^{t-l}(x_{t-l}, y_{t-l}) &= (x - x_{t-l}, y - y_{t-l}).
\end{aligned}
\tag{2}
$$

Note that all the flows used in this paper are backward flows. Therefore, the flow vectors at time steps t and $t-l$ are actually located at the transformed positions P_t and P_{t-l} w.r.t. the warped clothes, rather than the original position P w.r.t. the exemplary clothes. Such backward format can ensure each pixel in the warped clothes has a flow vector to indicate its original position to be sampled from the exemplary clothes, further ensuring the warping operation is valid.

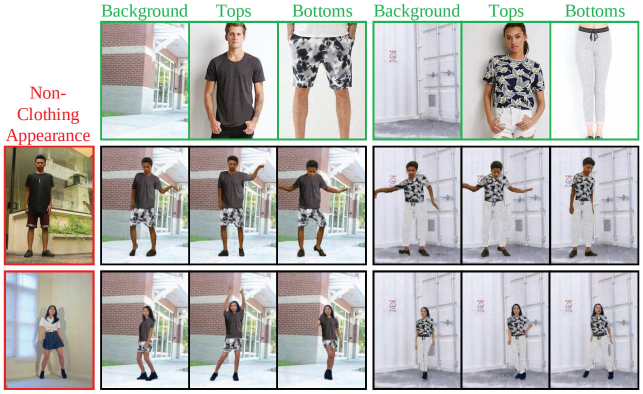

Figure 5: Examples of our multi-source appearance synthesis. Red-edged images describe the non-clothing foreground appearances (hair, face, torso, shoes). Green-edged images from left to right describe the appearances of background, tops, bottoms. Black-edged images are our synthesized motion transfer results. *Please zoom in for a better view.*

In principle, if the frames at t and $t-l$ are temporally consistent, $\hat{F}^t(x_t, y_t) - \hat{F}^{t-l}(x_{t-l}, y_{t-l})$ should be equal to the ground-truth optical flow vector $U(x_t, y_t)$, which is from t to $t-l$ and equal to $(x_{t-l} - x_t, y_{t-l} - y_t)$. As shown in Figure 4, to generalize this equation to the whole image rather than a single pixel, we should resample $\hat{F}^{t-l}(x_{t-l}, y_{t-l})$ at P_{t-l} to the same position P_t as $\hat{F}^t(x_t, y_t)$, and do the same to all the remaining pixels of $\hat{F}^{t-l}$ to make them share positions with those of $\hat{F}^t$. We can realize this by using U to warp $\hat{F}^{t-l}$. Thus our FTC loss is given by:

$$
L_{FTC,l} = \|\hat{F}^t - \mathrm{W}_U(\hat{F}^{t-l}) - U\|_1, \tag{3}
$$

where W denotes the warping operation based on U. To guarantee both short-term and long-term temporal consistency, we set $l = 1, 3, 9$ to compute FTC losses at three time scales and sum them together as our full FTC loss L_{FTC}.

We further utilize a TVL1 loss (Fan et al. 2018) L_{TVL1} to minimize the difference between flow vectors at neighbored positions of $\hat{F}^t$, which smooths the warping. Summarily, the full objective is a weighted sum of several losses, given by:

$$
L_{full} = L_{VGG} + L_{VGG}^{coarse} + \lambda_1 L_{FTC} + \lambda_2 L_{TVL1}, \tag{4}
$$

where λ_1 and λ_2 denote the weights of FTC and TVL1 losses, respectively.

Multi-Source Appearance Attribute Editing

Compared with existing HVMT methods, C2F-FWN can support multi-source appearance attribute editing when transferring motions. As described above, we divide the exemplary appearance into the background, clothing and non-

Table 1: Quantitative results tested on our SoloDance dataset and iPER (2019b) dataset. SSIM, PSNR, TCM are similarity metrics, the higher the better (SSIM and TCM range from 0 to 1). LPIPS and FID are distance metrics, the lower the better. Note that TCM measures temporal consistency while other metrics measure spatial consistency.

Datasets	Metrics	EDN(2019)	FSV2V(2019)	LWGAN(2019b)	SGWGAN(2018)	ClothFlow(2019)	w/o FTC loss	w/o LC-DConv	Ours
SoloDance	SSIM	0.811	0.721	0.786	0.763	0.843	0.849	0.850	**0.879**
	PSNR	23.22	20.84	20.87	20.54	22.06	23.05	23.19	**26.65**
	LPIPS	0.051	0.132	0.106	0.124	0.072	0.065	0.063	**0.049**
	FID	53.17	112.99	86.53	99.24	76.61	64.92	61.03	**46.49**
	TCM	0.347	0.106	0.176	0.166	0.322	0.319	0.401	**0.641**
iPER(2019b)	SSIM	0.840	0.780	0.825	0.818	0.814	0.824	0.822	**0.849**
	PSNR	23.39	20.44	21.43	22.41	21.87	22.76	22.52	**24.27**
	LPIPS	0.076	0.110	0.091	0.086	0.088	0.082	0.084	**0.072**
	FID	56.29	110.99	77.99	101.99	71.21	64.40	63.72	**55.07**
	TCM	0.361	0.184	0.197	0.260	0.422	0.411	0.499	**0.687**

clothing foregrounds. The clothing foreground can be further divided into tops and bottoms, which decide how the exemplary human subject is dressed in the synthesized videos. The non-clothing foreground can be further divided into hair, face, torso and shoes, with the first three parts deciding the human identity. With the help of semantic layouts, the background and each part of the foregrounds can be extracted from different sources to achieve the multi-source exemplary appearance. For example, the background can be replaced by arbitrary fixed images. Tops and bottoms in the clothing foreground can be extracted from arbitrary fashion or portrait images, which is the same for parts of the non-clothing foreground. With such multi-source appearance inputs, our proposed method can generate the corresponding multi-source appearance in the synthesized videos, which enables full appearance attribute editing for motion transfer as shown in Figure 5. Such capability can achieve rather high flexibility and efficiency in real applications. For example, users can arbitrarily change their clothes and backgrounds in videos without really wearing these clothes or performing in these backgrounds, enabling convenient video re-creation.

Experiments

Dataset

SoloDance Dataset We built a large-scale SoloDance dataset containing 179 solo dance videos with 53,700 frames. Specifically, 143 human subjects were captured with each wearing various clothes and performing complex dances (e.g., modern, street dances) in various backgrounds. Compared to the iPER dataset (Liu et al. 2019b) that only contains 30 subjects performing simple moves (e.g., random actions, A-poses), our dataset offers more appearance variety and motion complexity. We utilized (Cao et al. 2017) and (Gong et al. 2018) to detect body poses and semantic layouts, and further obtained foregrounds and backgrounds for each video. In our experiments, we randomly split the dataset into 153 and 26 videos for training and testing.

iPER Dataset We also evaluated our method on the iPER dataset (2019b). The data preprocessing of the iPER dataset is the same as our SoloDance dataset. Following the original protocol of iPER, we used 164 videos for training and the remaining 42 videos for testing.

Implementation Details

All the frames were resized and cropped to 256x256 sizes to train our models. Since backgrounds are fixed and easy to generate compared to animated human foregrounds, we further cropped the frames to central 192x256 body regions during evaluation to focus on the quality of the synthesized foregrounds. The design of the layout GAN in Stage 1 and the composition GAN in Stage 3 followed (Wang et al. 2018a). The design of our FPNs in Stage 2 followed (Lin et al. 2017) except that we replaced standard convolutions in bottom-up pathways of FPN-A and FPN-M with our LC-DConv to enhance the features. Particularly, the LC-DConv was implemented based on (Dai et al. 2017) by employing layout-constrained sampling locations. Moreover, to enable the supervision of the proposed FTC loss, we utilized (Ilg et al. 2017) to obtain the ground-truth optical flows. We trained each stage for 10 epochs separately with Adam optimizers (Kingma and Ba 2014) (learning rate: 0.0002, β_1: 0.5, β_2: 0.999) on an Nvidia RTX 2080 Ti GPU, where we set $\lambda_1 = 5$ and $\lambda_2 = 0.5$ in Eq. 4 to trade-off the two losses.

Baselines

To evaluate our proposed approach, we made comparisons with state-of-art HVMT methods including a personalized method EDN (Chan et al. 2019), a direct generation method FSV2V (Wang et al. 2019), two feature warping methods LWGAN (Liu et al. 2019b) and SGWGAN (Dong et al. 2018), and an image warping method ClothFlow (Han et al. 2019). In our implementation, we used 3000 frames for each person to train personalized models for EDN, and used the same data as ours to train models for other methods.

Quantitative Results

We utilized both traditional (SSIM and PSNR) and CNN-based metrics (LPIPS (Zhang et al. 2018) and FID (Heusel et al. 2017)) to measure the quality of synthesized frames, which can assess the spatial consistency between synthesized and exemplary images. We also utilized a Temporal Consistency Metric (TCM) (Yao, Chang, and Chien 2017) to evaluate the temporal consistency, which is an essential factor in measuring the quality of videos rather than single frames. Specifically, TCM measures temporal consistency by calculating warping errors between successive synthe-

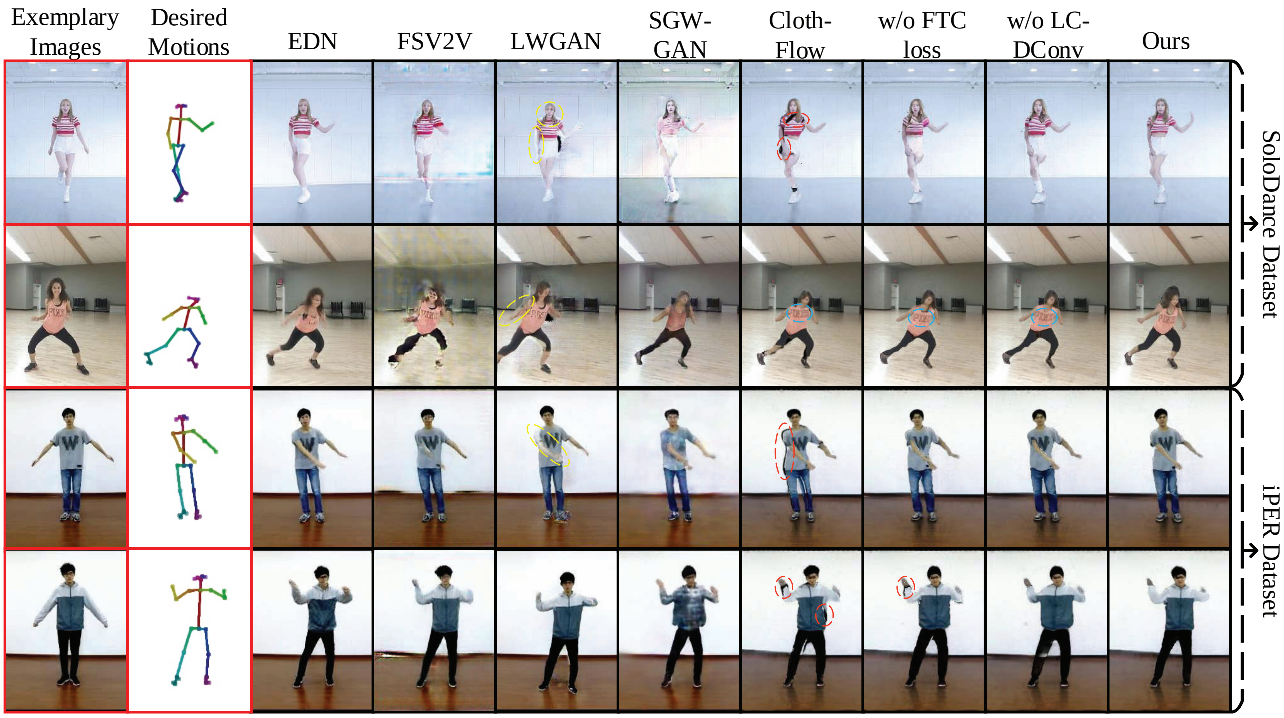

Figure 6: Qualitative comparisons with other methods including EDN (2019), FSV2V (2019), LWGAN (2019b), SGWGAN (2018), ClothFlow (2019), albated variants without FTC loss, LC-DConv. Yellow, blue and red circles point out blurry surfaces, over-stretched clothes patterns, and black chinks (caused by misplacements), respectively. *Please zoom in for a better view.*

sized frames, where each frame is warped by the ground-truth optical flow to compare with its neighbored frame. The quantitative results of all the methods are summarized in Table 1. We can see that our proposed C2F-FWN significantly outperforms all the other methods including the personalized method EDN (2019) for all the metrics (especially for the TCM scores) on both of the two datasets, which indicates that our approach can achieve HVMT with better spatial and temporal consistency.

Qualitative Results

As shown in Figure 6, we randomly visualize some motion transfer video frames synthesized by different methods for qualitative comparisons, where our approach outperforms all the other methods. Specifically, we achieve better spatial consistency with exemplary images than others, especially the direct generation method FSV2V (2019), where our method preserves the exemplary appearance details such as decorative patterns and colors well. Besides, benefiting from our coarse-to-fine flow warping, we can capture the desired motions better than existing warping-based methods, with our warped clothes precisely aligned with the body layouts. However, the feature warping method SGWGAN (2018) can't enable precise feature alignment with the desired motions due to the limited warping capability, which causes poor appearance details. Another feature warping method LWGAN (2019b) results in blurry details on the surface of bodies and clothes (e.g., circled in yellow in Figure 6) because of the low-precision SMPL models. The image warping method ClothFlow (2019) can't warp the exemplary images to align with the desired motions, which results in visual artifacts such as over-stretching and misplacement near the layout boundaries (e.g., circled in blue and red in Figure 6). Although the personalized method EDN (2019) can generate comparable results to us, it often results in blurrier textures. We also show some of our multi-source appearance synthesis results in Figure 5, where we utilized fashion images dissimilar from our training data to extract tops and bottoms. We can see that the multi-source exemplary appearances are also well preserved, enabling flexible appearance attribute editing for HVMT. Videos of the qualitative comparisons and our synthesized results can be found in *our supplementary materials*, where we show that our method can also achieve better temporal consistency.

Ablation Study

We also conducted ablation studies w.r.t. our FTC loss and LC-DConv to demonstrate their effectiveness. Specifically, we implemented two variant models for comparisons. One was trained without our FTC loss, and another only adopted standard convolutions to extract features. As shown in Table 1, our full method outperforms the two variants for all the metrics. As shown in Figure 6, without the two components, the clothes are warped imprecisely (e.g., circled in blue and red in Figure 6), which indicates the importance of the LC-DConv as well as the FTC loss for enhancing our flow warping and improving spatial consistency. Moreover, we observed that the variant without the FTC loss would result in much worse video coherence than our full method, which shows our superiority in improving temporal consistency. **Please refer to *our supplementary video* for more details:** https://youtu.be/THuQN1GXuGI.

Conclusion

In this paper, we propose Coarse-to-Fine Flow Warping Network (C2F-FWN) to achieve both spatial and temporal consistency for HVMT, enabling us to preserve exemplary appearances as well as improve video coherence. Specifically, our coarse-to-fine flow warping can precisely model geometric deformations caused by motions to ensure the spatial consistency, where we further utilize our Layout-Constrained Deformable Convolution (LC-DConv) to enhance features for estimating the transformation flows. To achieve the temporal consistency, we propose a novel Flow Temporal Consistency (FTC) Loss to learn explicit temporal consistency between successive transformation flows, which significantly improves the video coherence. Experimental results tested on our SoloDance dataset and the iPER dataset show our superiority to other methods in terms of both spatial and temporal consistency. Ablation studies w.r.t. our FTC loss and LC-DConv demonstrate their effectiveness in improving our synthesis quality. We also demonstrate that our method can achieve flexible appearance attribute editing provided with alterable multi-source appearance inputs, which shows promising application prospects.

Limitations and Future Work

Although our method works well in most cases, it may fail (e.g., jitters, blurs) due to errors in poses and semantic layouts, which would cause errors in our model inputs and further result in artifacts in our output results. In the future, we can utilize more accurate pose and layout estimation techniques to eliminate these errors. Besides, we currently only provide our model with one single exemplary image, which might suffer from self occlusions and texture missings in cases of extremely large motion changes. Thus, how to attend and aggregate multiple exemplary images for warping is also worth studying in future works.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (U19B2043).

References

Aberman, K.; Shi, M.; Liao, J.; Liscbinski, D.; Chen, B.; and Cohen-Or, D. 2019. Deep Video-Based Performance Cloning. In *Computer Graphics Forum*, volume 38, 219–233. NJ, USA: Wiley Online Library.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7291–7299.

Chan, C.; Ginosar, S.; Zhou, T.; and Efros, A. A. 2019. Everybody dance now. In *Proceedings of the IEEE International Conference on Computer Vision*, 5933–5942.

Chen, Z.; Wang, C.; Yuan, B.; and Tao, D. 2020. PuppeteerGAN: Arbitrary Portrait Animation with Semantic-aware Appearance Transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13518–13527.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Dong, H.; Liang, X.; Gong, K.; Lai, H.; Zhu, J.; and Yin, J. 2018. Soft-gated warping-gan for pose-guided person image synthesis. In *Advances in neural information processing systems*, 474–484.

Fan, L.; Huang, W.; Gan, C.; Ermon, S.; Gong, B.; and Huang, J. 2018. End-to-End Learning of Motion Representation for Video Understanding. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6016–6025.

Gong, K.; Liang, X.; Li, Y.; Chen, Y.; Yang, M.; and Lin, L. 2018. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision*, 770–785.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 26722680.

Han, X.; Hu, X.; Huang, W.; and Scott, M. R. 2019. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE International Conference on Computer Vision*, 10471–10480.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, 6626–6637.

Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; and Brox, T. 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2462–2470.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, 694–711. Springer.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Liang, X.; Gong, K.; Shen, X.; and Lin, L. 2018. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE Trans. pattern analysis and machine intelligence* 41(4): 871–885.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Liu, L.; Xu, W.; Zollhöfer, M.; Kim, H.; Bernard, F.; Habermann, M.; Wang, W.; and Theobalt, C. 2019a. Neural Rendering and Reenactment of Human Actor Videos. *ACM Trans. Graph.* 38(5): 1–14. ISSN 0730-0301.

Liu, W.; Piao, Z.; Min, J.; Luo, W.; Ma, L.; and Gao, S. 2019b. Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, 5904–5913.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* 34(6): 1–16.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2337–2346.

Rocco, I.; Arandjelovic, R.; and Sivic, J. 2017. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6148–6157.

Wang, T.-C.; Liu, M.-Y.; Tao, A.; Liu, G.; Catanzaro, B.; and Kautz, J. 2019. Few-shot Video-to-Video Synthesis. In *Advances in Neural Information Processing Systems*, 5013–5024.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Liu, G.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018a. Video-to-video synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 1152–1164. Curran Associates Inc.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018b. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8798–8807.

Webster, R.; Rabin, J.; Simon, L.; and Jurie, F. 2019. Detecting overfitting of deep generative networks via latent recovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11273–11282.

Wei, D.; Xu, X.; Shen, H.; and Huang, K. 2020. GAC-GAN: A General Method for Appearance-Controllable Human Video Motion Transfer. *IEEE Transactions on Multimedia* 1–1.

Yang, Z.; Zhu, W.; Wu, W.; Qian, C.; Zhou, Q.; Zhou, B.; and Loy, C. C. 2020. TransMoMo: Invariance-Driven Unsupervised Video Motion Retargeting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5306–5315.

Yao, C.-H.; Chang, C.-Y.; and Chien, S.-Y. 2017. Occlusion-aware video temporal consistency. In *Proceedings of the 25th ACM international conference on Multimedia*, 777–785.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.