

GAC-GAN: A General Method for Appearance-Controllable Human Video Motion Transfer

Dongxu Wei, Xiaowei Xu, Haibin Shen, and Kejie Huang

arXiv:1911.10672v2 [cs.CV] 9 Feb 2020

Abstract—Human video motion transfer has a wide range of applications in multimedia, computer vision and graphics. Recently, due to the rapid development of Generative Adversarial Networks (GANs), there has been significant progress in the field. However, almost all existing GAN-based works are prone to address the mapping from human motions to video scenes, with scene appearances are encoded individually in the trained models. Therefore, each trained model can only generate videos with a specific scene appearance, new models are required to be trained to generate new appearances. Besides, existing works lack the capability of appearance control. For example, users have to provide video records of wearing new clothes or performing in new backgrounds to enable clothes or background changing in their synthetic videos, which greatly limits the application flexibility. In this paper, we propose GAC-GAN, a general method for appearance-controllable human video motion transfer. To enable general-purpose appearance synthesis, we propose to include appearance information in the conditioning inputs. Thus, once trained, our model can generate new appearances by altering the input appearance information. To achieve appearance control, we first obtain the appearance-controllable conditioning inputs and then utilize a two-stage GAC-GAN to generate the corresponding appearance-controllable outputs, where we utilize an ACGAN loss and a shadow extraction module for output foreground and background appearance control respectively. We further build a solo dance dataset containing a large number of dance videos for training and evaluation. Experimental results show that, our proposed GAC-GAN can not only support appearance-controllable human video motion transfer but also achieve higher video quality than state-of-art methods.

Index Terms—Motion Transfer, Video Generation, Image Synthesis, Generative Adversarial Networks (GANs).

I. INTRODUCTION

HUMAN video motion transfer (HVMT) aims at synthesizing a video that the person in a target video imitates actions of the person in a source video, which is of great benefit to applications in scenarios such as games, movies and robotics. For example, the animation of virtual characters plays a key role in VR/AR games and movies. Based on HVMT techniques, we can animate the virtual game roles or movie actors freely to perform user-defined mimetic movements, thus rendering plausible visual results [48], [47]. Moreover, the animated visual data can be further utilized as simulated training data to

D. Wei, H. Shen, and K. Huang are with the department of Information Science & Electronic Engineering, Zhejiang University, Hangzhou 310027, China (Email: tracywei@zju.edu.cn; shen_hb@zju.edu.cn; huangkejie@zju.edu.cn).

X. Xiao is with Guangdong Provincial Peoples Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China, 510080. (Email: xiao.wei.xu@foxmail.com)

train robotic agents that work for real-world situations, where real experiences may be hard to obtain [46].

With the recent emergence of Generative Adversarial Networks (GANs) [1] and its variant conditional GANs (cGANs) [2], there are many GAN-based works [15], [16], [17], [18], [19] that achieve great success in HVMT. For ease of discussion, we decompose the video scene into scene appearance (background and human foreground) and human motion in the context of HVMT. Existing works have two limitations. First, only the mapping from human motions to video scenes is addressed while scene appearances are encoded individually in the trained models. Therefore, once trained, each model is specific to the scene appearance of a target video and cannot generalize to other scene appearances. They have to train additional video-specific models with new target videos as the training data to generate new scene appearances. Unfortunately, due to the large cost of manpower and computing resources produced by the data collection and the model training, such approach lacks efficiency for practical applications. Second, existing methods can't control the scene appearance. In particular, background and human foreground appearances are bound together and not allowed to be altered. Therefore, these methods can't synthesize videos with users wearing new clothes or performing in new backgrounds if users have never been in these clothes or backgrounds. However, users expect to alter appearances in their synthetic videos without the efforts of real clothes and background changing. Thus, despite the human motion control, further appearance control is needed to provide high flexibility in practical applications.

In this work, we propose GAC-GAN: a general method for appearance-controllable human video motion transfer. For general-purpose appearance synthesis, we propose to feed our model with appearance conditioning inputs in addition to motion conditioning inputs (body poses) used in other works, allowing the model to learn the mapping from human motions and scene appearances to video scenes. For appearance control, we propose to control output appearances through control of the conditioning inputs. Specifically, we propose a multi-source input selection strategy to first exert appearance control on the conditioning inputs during data preprocessing. Then a two-stage GAC-GAN framework which consists of a layout GAN and an appearance GAN is proposed to generate the corresponding appearance-controllable outputs from the conditioning inputs, where we further apply an elaborate ACGAN loss and a light-weight shadow extraction module to the appearance GAN to achieve control of the output human foreground and background

respectively.

In our experiments, a large solo dance dataset including 148800 frames collected from 124 people is utilized for general-purpose training and evaluation. We first compare our approach against state-of-art video-specific [16], [15] and general-purpose [9] methods through qualitative, quantitative and perceptual evaluations on the test set. The results show that, compared with other methods, our proposed approach can synthesize high-quality motion transfer videos that are perceptually more popular and quantitatively more similar to ground-truth real videos in a general way. Then we apply our method to ordinary and appearance-controllable HVMT tasks for further validation on simulated real-world situations where no ground-truth video is available. The results show that, in addition to the human motion control, our method can further control the appearances of the human foregrounds as well as the surrounding backgrounds flexibly. Moreover, to give a better insight into the proposed GAC-GAN framework, we conduct comprehensive ablation studies for our important components (i.e., multi-source input selection strategy, layout GAN, ACGAN loss and shadow extraction module).

To summarize, our main contributions are as follows:

- We propose GAC-GAN: a general approach enabling appearance-controllable human video motion transfer.
- We achieve higher video quality than state-of-art methods by taking advantage of our novel component designs.
- We construct a large-scale solo dance dataset including a variety of solo dance videos for training and evaluation, which will be released publicly to facilitate future research.

The rest of the paper is structured as follows: Sec.II discusses the related work. Sec.III introduces the problem formulation in our work. In Sec.IV, we describe the proposed GAC-GAN. In Sec.V, we report and discuss our experimental results. Finally, Sec.VI concludes the paper and discusses the future work.

II. RELATED WORK

A. Classic motion transfer.

Early works have attempted to reorder existing video frames [20], [26], [25] to obtain new videos consisting of frames with motions similar to the desired motions, where the results are not temporally coherent and can be easily distinguished from real videos. Later techniques try to animate coarse 3D character models [22], [23], [21] to create rendered motion transfer videos, which results in coarse body silhouettes and unrealistic texture details. Recently, methods [43], [44], [45] estimate detailed 3D characters with controllable body meshes to render plausible video results. However, most of these 3D rendering approaches require massive computation budgets dominated by the production-quality 3D reconstructions, which is inefficient for real-world applications.

B. Image and video generation.

Instead of relying on temporally incoherent video manipulations or computationally expensive 3D reconstructions, current motion transfer works depend more on image and video generation techniques. Traditional generation methods are prone

to deal with syntheses of local textures based on simple hand-crafted features [29]. With the development of deep learning algorithms, variational autoencoder (VAE) [30] and generative adversarial networks (GANs) [1] become two mainstream methods due to their capabilities of synthesizing large-size images. Benefiting from the powerful two-player adversarial training, GAN-based generative models can synthesize images that are less blurry and more realistic than those generated by VAEs, which causes GANs are more exploited in image and video generation works. In the beginning, GAN-based image generation works [31], [32] focus on designing GAN architectures to improve synthetic image resolutions. However, their image results are randomly generated from randomly sampled noises, which is out of user control. Since the emergence of conditional GANs (cGANs) [2], works start to take class labels [35], [36] or descriptive images [3], [34], [33] as extra conditioning inputs to control the output image appearances, which belongs to the same method category as our proposed cGAN-based approach. Besides of image generation, there are also works [37], [38], [39], [40], [41], [42] focus on synthesizing temporally coherent video sequences. For instance, unconditional video generation works [37], [38], [39] try to improve temporal consistency between adjacent synthetic frames based on GANs that consider not only visual quality but also temporal coherence. However, these approaches fail to generate high-quality or long-term video results, with scene appearances are randomly synthesized in an unconditional manner. Besides, video prediction techniques [40], [41], [42] attempt to predict future video sequences based on the currently observed video sequences. Although the synthetic appearances are conditioned on the previous frames, future video motions are unconditionally generated, which is inappropriate for the motion controllable video synthesis that HVMT concerns.

C. GAN-based motion transfer.

Due to the great success of the GAN-based image and video generation approaches mentioned above, many works are developed for motion transfer based on them.

1) *Image-based human pose transfer*: In the recent years, there have been significant efforts which we refer to as image-based methods [9], [11], [14], [13], [8], [12], [10] aiming at synthesizing new pose images given the human appearance of a single input image. The purpose of these image-based works is to impose the input human appearance onto new poses in an image-to-image translation manner [3], which is very similar to the human video motion transfer that we focus on. [9], [11], [14] utilize spatial transformations or surface deformations to transform the input appearance texture into new pose layouts, where the transformed results are rough and refined in detail to generate output images. Similarly, [13], [8], [12] apply such transformations or deformations to appearance features instead of textures, where the transformed features are then decoded to generate new pose images. Furthermore, [10] propose a style discriminator to force the generator to preserve the input appearance style, which gives a new sight from the aspect of discriminator design. Although these image-based methods can achieve general-purpose appearance synthesis, all of them

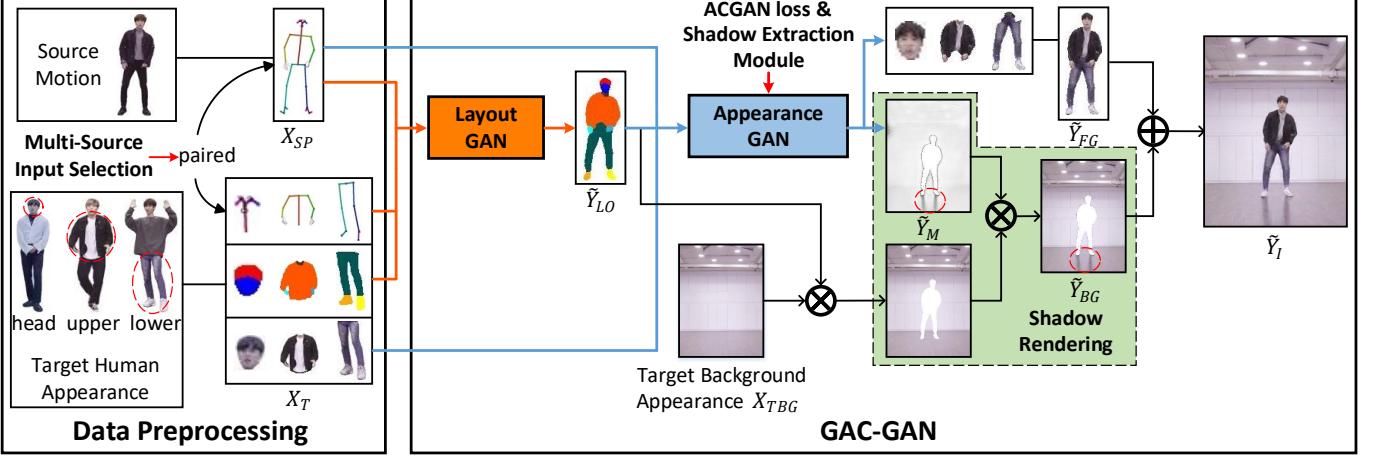


Fig. 1. Overview of our method. In the data preprocessing, we obtain the paired X_{SP} and X_T from the source motion frame and the target appearance frames based on the multi-source input selection strategy. Then the processed inputs are fed into the GAC-GAN which consists of a layout GAN and an appearance GAN to sequentially generate the layout \tilde{Y}_{LO} and the scene appearance \tilde{Y}_I (composed of the synthetic foreground \tilde{Y}_{FG} and the rendered background \tilde{Y}_{BG}), where we further apply an ACGAN loss and a shadow extraction module to the appearance GAN to control foreground and background appearances respectively. In the figure, the orange and the blue arrows represent data flows of the layout GAN and the appearance GAN respectively, \otimes and \oplus represent pixel-wise multiplication and addition operations respectively. In the data preprocessing module, the red circles specify the desired body part appearances. In the GAC-GAN module, the red circles point out how the synthetic shadow map \tilde{Y}_M modulates brightness for the input background image X_{TBG} , which enables shadow rendering.

are designed for still image generation without consideration of temporal coherence, which causes they are not qualified for video synthesis that we concern. Besides, these methods try to generate unseen body views from a single input image, which greatly restricts their performance due to the lack of appearance information, especially when the desired output pose greatly differs from the input pose.

2) *Video-based human motion transfer*: As the video counterpart of the above mentioned image-based pose transfer, video-based motion transfer considers video generation with access to more appearance information contained in a whole video, leading to a higher level of temporal coherence and visual quality. In [16], the authors propose to generate optical flows to warp previously generated frames into temporally consistent new frames. Besides, [15] use a temporal smoothing loss to enforce temporal consistency between adjacent frames. Note that video quality depends not only on temporal coherence but also on appearance details. Thus recent works come up with feeding rendered images of 3D models [17] or transformed images of body parts [18] into their models as input conditions to obtain realistic appearances. Moreover, [19] split their network into two training branches with respect to appearance generation and temporal coherence improvement to account for both sides. Although these works can generate videos with higher quality than image-based methods, an obvious limitation is that they have to train additional models to generate unseen scene appearances, keeping them from general-purpose appearance synthesis required in real-world applications. Besides, none of them can realize controllable appearance synthesis to satisfy user demands for clothes and background changing. Although [18] can support background replacement with user-defined images, they don't allow users to try on different clothes in the synthetic videos.

III. PROBLEM FORMULATION

Before describing our method, we first define the problem to solve: given conditioning input of a source motion video and multiple target appearance videos, we aim at synthesizing a new video with human motion of the source video and combined scene appearance of the target videos. Specifically, the conditioning input is divided into motion conditioning input (source motion) and appearance conditioning input (target appearance), where the target appearance is further divided into human and background appearances. Source motion input is described by the estimated body poses of the source video frames. To enable human appearance control, target human appearance input is decomposed into three user-defined body parts (e.g., head, upper body and lower body) with respect to appearances of face, upper garment and lower garment, each of which is described by the estimated body part poses, layouts and foregrounds of its own target video frames. To enable background appearance control, target background appearance input is described by a user-defined background image. Conditioned on the source motion and the target appearance inputs, we generate the corresponding outputs including body layouts, body part foregrounds and shadow maps, where the body layouts are generated by the layout GAN while the others are generated by the appearance GAN. Then we use the synthetic shadow maps to render shadows on the input background image. Finally, we obtain the synthetic full scenes by composing the synthetic body part foregrounds and the rendered backgrounds together. For the above mentioned inputs and outputs, we give their variable definitions used in this paper as follows:

1) Inputs

- **source motion**: source pose X_{SP}
- **target human appearance** (X_T):
target poses: $X_{TP,H}$, $X_{TP,U}$, $X_{TP,L}$

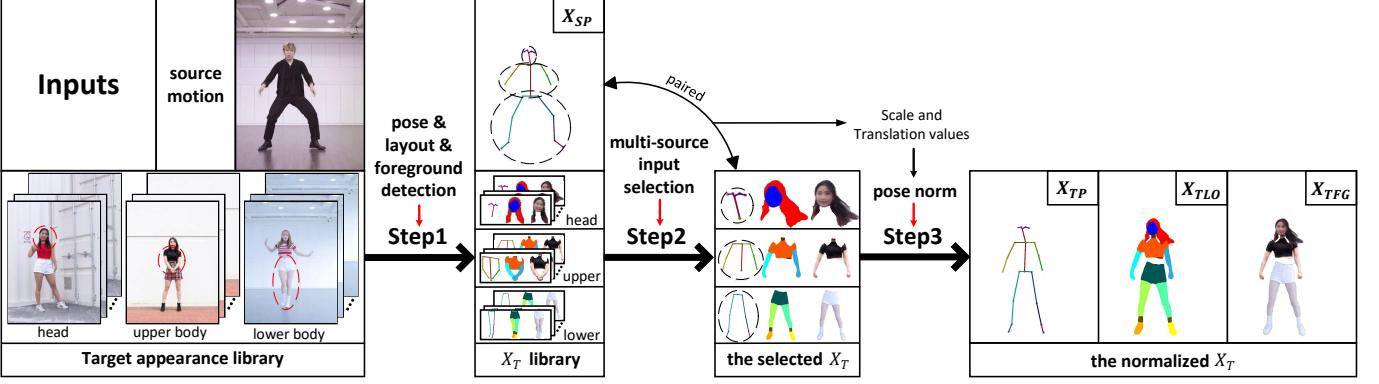


Fig. 2. Illustration of our data preprocessing. In step 1, we detect poses, layouts and foregrounds to obtain the motion condition X_{SP} and the appearance condition (X_T) library, where each red circle specifies a target body part. In step 2, each body part of the source pose X_{SP} is paired with a target body part (pose X_{TP} , layout X_{TLO} , and foreground X_{TFG}) in the X_T library according to body part pose similarity. Then in step 3, we use the computed scale and translation values between body parts of X_{SP} and X_T to transform the body parts of X_T into the same sizes and positions as those of X_{SP} .

target layouts: $X_{TLO,H}, X_{TLO,U}, X_{TLO,L}$
 target foregrounds: $X_{TFG,H}, X_{TFG,U}, X_{TFG,L}$
 • target background appearance: X_{TBG}

2) Outputs

- layout GAN: body layout \tilde{Y}_{LO}
- appearance GAN:
 - body part foregrounds: $\tilde{Y}_{FG,H}, \tilde{Y}_{FG,U}, \tilde{Y}_{FG,L}$
 - shadow map: \tilde{Y}_M
 - background: \tilde{Y}_{BG}
 - full scene: \tilde{Y}_I

where X and \tilde{Y} mean input and output, S and T represent input source and target videos, P, LO, FG, BG, M, I represent pose, layout, foreground, background, shadow map and scene image, H, U, L refer to head, upper body and lower body.

IV. METHOD

In this section, we first give the overview of our proposed method, which is followed by two subsections with respect to our data preprocessing and GAC-GAN framework.

A. Overview

The overview of our method is depicted in Figure 1.

First, we apply **data preprocessing** to the input videos to obtain our conditioning inputs, where we pair each motion conditioning input (source pose X_{SP}) with an optimal appearance conditioning input X_T (target pose X_{TP} , layout X_{TLO} and foreground X_{TFG} of head, upper body and lower body) based on a **multi-source input selection** strategy. Specifically, each body part of X_T is obtained from its own target human appearance source, which can be altered based on user preferences to enable input appearance control.

Next, we feed the motion (X_{SP}) and the appearance (X_T) conditioning inputs into our two-stage **GAC-GAN** which consists of a layout GAN and an appearance GAN, responsible for controllable layout synthesis and appearance synthesis respectively. Because generating appearances directly from body pose points can be extremely hard, we can ease training by dividing our model into these two stages, where the synthetic layout can be regarded as the intermediate representation of

the final appearance result. Specifically, in the first stage, the **layout GAN** is designed to synthesize the foreground layout \tilde{Y}_{LO} whose body pose and body part distribution are consistent with the motion condition (X_{SP}) and the multi-source appearance condition (X_{TLO}) respectively. In the second stage, the **appearance GAN** takes the synthetic layout \tilde{Y}_{LO} as additional motion conditioning input to generate the desired scene appearance \tilde{Y}_I , which is composed of a synthetic foreground \tilde{Y}_{FG} and a rendered background \tilde{Y}_{BG} . As for the foreground, we train the appearance GAN with an **ACGAN loss** to ensure the appearance consistency between the synthetic foreground and the input appearance condition, which therefore enables foreground appearance control in consistency with the input appearance control. As for the background, we implant a light-weight **shadow extraction module** into the appearance GAN to generate a shadow map \tilde{Y}_M that modulates background brightness and renders appearance-irrelevant shadows on X_{TBG} , which therefore enables background appearance control by directly replacing background with arbitrary user-defined images.

B. Data Preprocessing

The main purpose of data preprocessing is to obtain our motion and appearance conditioning inputs. For each frame synthesis, the motion condition is extracted from a source motion frame while the appearance condition is extracted from a target appearance library which contains three kinds of target appearance video frames with respect to head, upper body and lower body. Since video sources of the three body parts are alterable based on user preferences, the multi-source input appearance condition is fully appearance-controllable. With body motion specified by the motion condition, the data preprocessing aims at obtaining the paired appearance condition which contains the maximum appearance information needed for appearance synthesis. Specifically, the data preprocessing consists of the three steps depicted in Figure 2, where the multi-source input selection strategy utilized in step 2 is the key to ensure the obtained appearance condition is optimal for the motion condition. It's noted that there's no restriction on the frame number for the target appearance library, we can

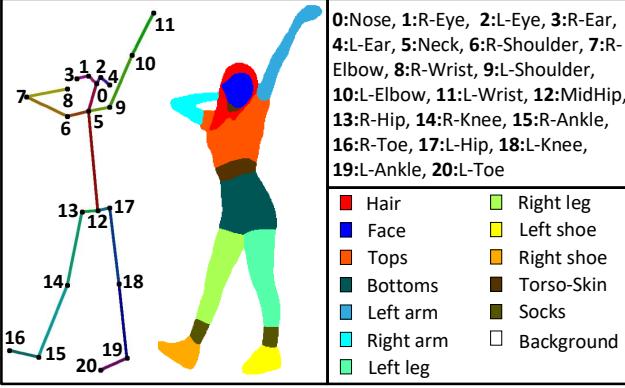


Fig. 3. Illustration of pose and layout detection results. Pose points and semantic labels are distinguished by numbers and colors respectively. "R-" means "right" and "L-" means "left".

obtain the optimal appearance condition no matter how many frames are provided.

1) *Step 1: Detecting Poses, Layouts and Foregrounds:* We utilize [6] and [4] to detect body poses and semantic layouts respectively, where the pose point locations and the layout classes are described in Figure 3. Then we can decompose the full body layouts into body part layouts for the three body part regions. Specifically, head region is a combination of hair and face; upper body region is a combination of tops, torso skin, left arm and right arm; lower body region is a combination of bottoms, left leg, right leg, left shoe, right shoe and socks. Thereafter we can extract foregrounds for each body part by multiplying the full images with the corresponding body part masks derived from the body part layouts. Based on the detections described above, we can obtain the input motion condition from the source motion frame and the appearance condition library from the target appearance library. In particular, the motion condition is the detected source body pose (X_{SP}) of the input source motion frame. The appearance condition library (X_T library) consists of three body part appearance condition libraries, each of which contains target body part poses (X_{TP}), layouts (X_{TLO}) and foregrounds (X_{TFG}) obtained from the corresponding video of the target appearance library.

2) *Step 2: Multi-Source Input Selection:* Since human appearances vary significantly with body poses, we propose a multi-source input selection strategy to select the optimal X_T based on the body part pose similarity. For each body part of the motion condition (X_{SP}), we select the paired body part appearance condition from the corresponding body part appearance condition library, where the pose of the selected body part appearance condition is the most similar to the body part pose of the motion condition within the body part appearance condition library. Thus we obtain the selected X_T which consists of three body part appearance conditions. Each body part appearance condition is composed of a body part pose, layout and foreground, containing the maximum appearance information needed for body part appearance synthesis. Specifically, pose similarity of each body part is denoted as the average cosine similarity between the corresponding source and target body part pose vectors:

$$Sim = \frac{1}{N} \sum_{i=1}^N \frac{\vec{V}_S^i \cdot \vec{V}_T^i}{|\vec{V}_S^i| |\vec{V}_T^i|} \quad (1)$$

where Sim is the body part pose similarity. \vec{V}_S^i and \vec{V}_T^i represent the i-th body part pose vectors of the source pose X_{SP} and the target pose X_{TP} respectively, $|\vec{V}_S^i|$ and $|\vec{V}_T^i|$ represent vector lengths of \vec{V}_S^i and \vec{V}_T^i respectively. N is the number of body part pose vectors, which equals to 5, 7, 8 for pose vectors of head, upper body, lower body. In particular, head pose vectors are $\vec{P}_0 \vec{P}_1, \vec{P}_0 \vec{P}_2, \vec{P}_1 \vec{P}_3, \vec{P}_2 \vec{P}_4, \vec{P}_0 \vec{P}_5$. Upper body pose vectors are $\vec{P}_5 \vec{P}_6, \vec{P}_6 \vec{P}_7, \vec{P}_7 \vec{P}_8, \vec{P}_5 \vec{P}_9, \vec{P}_9 \vec{P}_{10}, \vec{P}_{10} \vec{P}_{11}, \vec{P}_5 \vec{P}_{12}$. Lower body pose vectors are $\vec{P}_{12} \vec{P}_{13}, \vec{P}_{13} \vec{P}_{14}, \vec{P}_{14} \vec{P}_{15}, \vec{P}_{15} \vec{P}_{16}, \vec{P}_{12} \vec{P}_{17}, \vec{P}_{17} \vec{P}_{18}, \vec{P}_{18} \vec{P}_{19}, \vec{P}_{19} \vec{P}_{20}$. In the above description, $P_0 \sim P_{20}$ represent pose points marked as numbers as shown in Figure 3.

3) *Step 3: Pose Normalization:* Although body parts of the selected X_T have the most similar poses with those of X_{SP} , sizes and positions of different parts are not compatible with each other and therefore needed to be normalized to form a whole body spatially consistent with X_{SP} . In practice, we apply a pose normalization to transform each body part of X_T into the same size and position as the corresponding part of X_{SP} , where the scale values and the translation distances of different parts are computed separately by analyzing the differences between body parts of X_{SP} and X_{TP} in vector lengths and point locations:

$$\begin{aligned} Scale &= \frac{\sum_{i=1}^{N_v} |\vec{V}_S^i|}{\sum_{i=1}^{N_v} |\vec{V}_T^i|} \\ Translation &= \frac{1}{N_p} \sum_{j=1}^{N_p} (P_S^j - P_T^j) \end{aligned} \quad (2)$$

where N_v is the number of body part pose vectors, N_p is the number of body part pose points, P_S and P_T represent source and target body part pose points respectively. For head, $N_p = 6$, pose points are $P_0 \sim P_5$. For upper body, $N_p = 8$, pose points are $P_5 \sim P_{12}$. For lower body, $N_p = 9$, pose points are $P_{12} \sim P_{20}$.

Thus we obtain the transformed body part pose points, layouts and foregrounds. Then the pose points of different parts are connected to compose a new target pose X_{TP} while the body part layouts are processed into a one-hot tensor X_{TLO} with each channel represents a body part as shown in Figure 3. Similarly, the body part foregrounds are also processed into a tensor X_{TFG} which consists of body part channels consistent with X_{TLO} . By separating different body parts by different channels, we can eliminate the loss of appearance information caused by the overlap between body parts that come from different video frames. Moreover, since the obtained body parts are inherently misaligned, we can eliminate the difference between single-source and multi-source appearance inputs, which benefits our training because only single-source inputs are available during training due to the lack of ground truths for multi-source appearance outputs.

C. GAC-GAN

Given appearance is fully controllable in the conditioning input, the GAC-GAN is designed to generate the corresponding fully controllable appearance output. As shown in Figure 4, our GAC-GAN has two stages: a layout GAN and an appearance GAN, described in detail in the following subsections. It's

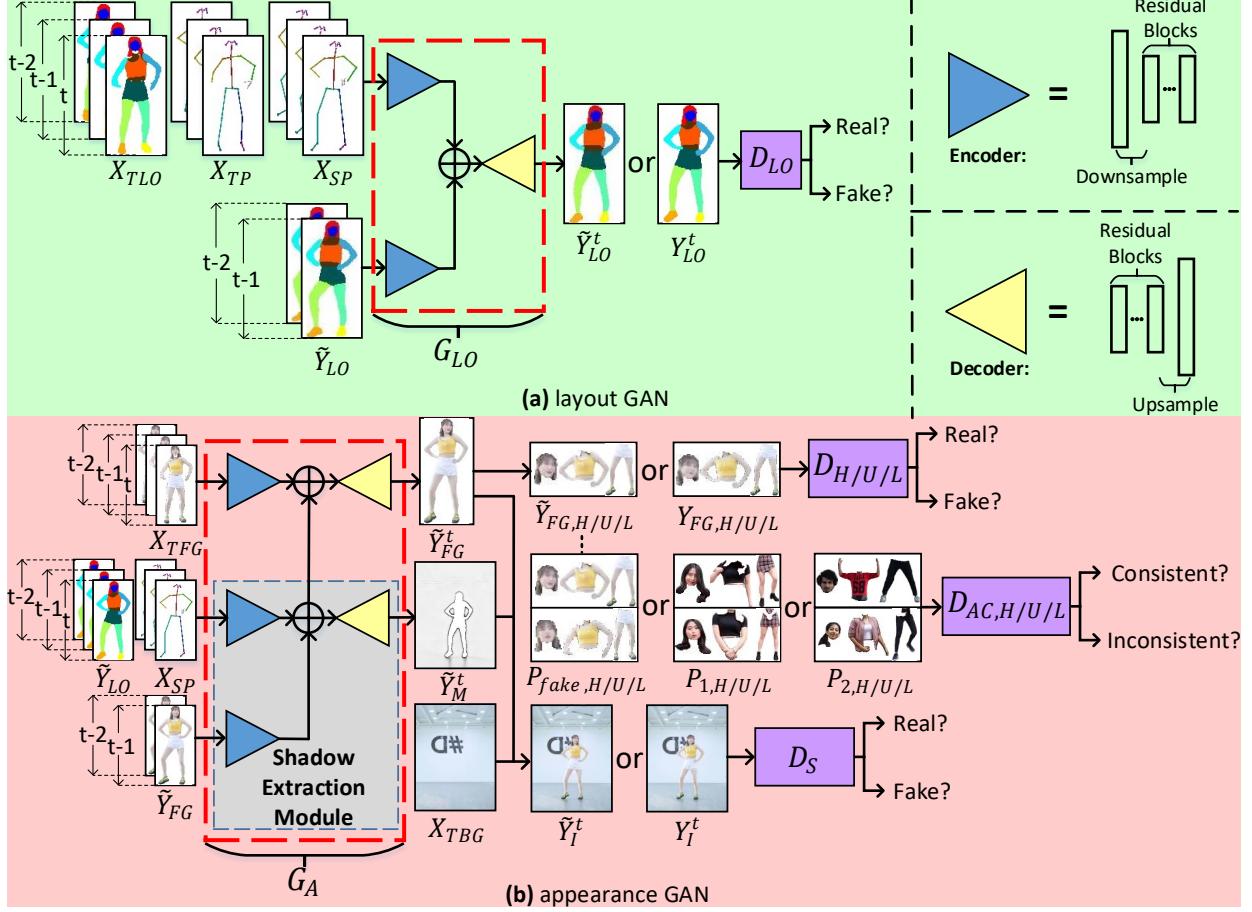


Fig. 4. Illustration of the GAC-GAN. (a) and (b) depict frameworks of the layout GAN and the appearance GAN respectively, where encoder and decoder architectures are also drawn above. In (b), foregrounds and discriminators of the three body parts are drawn in the same blocks annotated by $H/U/L$ for simplicity, which are separated in practice.

noted that, because videos are generated frame by frame, we present the generation of the frame at time t as an example in the following discussions for convenience.

1) Layout GAN: The layout GAN aims at synthesizing the desired multi-source body layout with body part distributions consistent with the multi-source appearance condition. By taking the synthetic layout as additional motion condition, we can describe the human motion at a more accurate pixel level compared to other works [15], [16], [17], [18], [19] that use sparse body pose points as motion conditions.

Network Architectures: Our layout GAN is made of a layout generator G_{LO} and a layout discriminator D_{LO} as shown in Figure 4(a). Specifically, the generator G_{LO} consists of two encoders and one decoder. The first encoder learns to encode features for the concatenation of three consecutive source poses, target poses and target layouts: $X_{LO}|_{t-2}^t = [X_{SP}|_{t-2}^t, X_{TP}|_{t-2}^t, X_{TLO}|_{t-2}^t]$. The second encoder learns to encode features for the concatenation of two previously generated layouts: $\tilde{Y}_{LO}|_{t-2}^{t-1}$. Then the two kinds of features are summed and fed into the decoder to generate the desired layout \tilde{Y}_{LO}^t . Here we include features of the concatenated consecutive frames to improve temporal consistency. Besides, the discriminator D_{LO} is designed to be multi-scale [3] to determine whether the generated layout is real or fake.

Objective Function: To train the layout GAN, we design the objective like this:

$$L_{LO} = L_{GAN}^{LO} + \lambda_{SS} L_{SS}^{LO} + \lambda_T L_T^{LO} + \lambda_{FM} L_{FM}^{LO} \quad (3)$$

L_{GAN}^{LO} is the adversarial loss of the layout GAN, which is given by:

$$\begin{aligned} L_{GAN}^{LO} = & E[\log D_{LO}(Y_{LO}, X_{LO}) \\ & + \log[1 - D_{LO}(\tilde{Y}_{LO}, X_{LO})]] \end{aligned} \quad (4)$$

where Y_{LO} is the real layout map with respect to \tilde{Y}_{LO} . L_{SS}^{LO} is the structural sensitive loss adapted from [5] and weighted by λ_{SS} , which is used to minimize the difference between Y_{LO} and \tilde{Y}_{LO} at both the pixel level and the structure level. It can be derived like this:

$$\begin{aligned} L_{SS}^{LO} &= L_{joint} \cdot L_{pixel}, \\ L_{joint} &= \frac{1}{2n} \sum_{i=1}^n \|C_{i,real} - C_{i,fake}\|_2^2 \end{aligned} \quad (5)$$

where the pixel-wise softmax loss L_{pixel} is weighted by the joint structure loss L_{joint} , which is an L2 loss used to measure the structural difference between the real and the generated layout maps. $C_{i,real}$ and $C_{i,fake}$ represent center points of the real and the generated layout maps, respectively, which are computed by averaging coordinate values of the i -th layout regions for the two layout maps. Specifically, when i ranges from 1 to n ($n = 9$), the i -th region represents: head, tops,

bottoms, left arm, right arm, left leg, right leg, left shoe and right shoe. As shown in Figure 3, all the regions have their class labels except for the head, which is a merged region of face and hair.

L_T^{LO} weighted by λ_T is the temporal loss used to minimize temporal difference between the real and the generated layout sequences, which can be derived like this:

$$L_T^{LO} = E[\log D_{LO}^T(S_{LO}) + \log[1 - D_{LO}^T(\tilde{S}_{LO})]] \quad (6)$$

where D_{LO}^T is the temporal discriminator of the layout GAN, trained to determine whether a layout sequence is real or fake. S_{LO} and \tilde{S}_{LO} are the real and the generated layout sequences, which are obtained by concatenating three consecutive Y_{LOS} and \tilde{Y}_{LOS} sampled by the sampling operator presented in vid2vid [16].

L_{FM}^{LO} is the discriminator feature matching loss presented in pix2pixHD [7] and weighted by λ_{FM} , which is used to improve synthesis quality.

2) *Appearance GAN*: Provided with the additional synthetic motion condition that specifies the desired body layout, the appearance GAN aims at synthesizing the desired foreground and background appearances, which are added together to compose the full scene appearance.

As for the foreground, since the appearance is already controllable in the input appearance condition, we can synthesize the corresponding controllable foreground appearance by ensuring the appearance consistency between the synthetic and the input appearances. Therefore, we propose an **ACGAN loss** to supervise not only visual quality but also appearance consistency during training. Besides, since ground truths for multi-source appearance outputs don't exist, we utilize three part-specific ACGAN losses with respect to head, upper body and lower body to supervise different body parts separately rather than supervise them as a whole, which helps to alleviate inner relevance between body parts that come from the same videos in our training data.

As for the background, we implant a light-weight **shadow extraction module** into the appearance GAN to generate the shadow map that modulates background brightness and renders background shadow rather than directly generate the background appearance from scratch [15], [16], [17], [18], [19]. The reasons are manifold: 1) Since video backgrounds are fixed and can be regarded as still images, patterns of the backgrounds are much fewer than those of the foregrounds in the training data. A deep learning model could easily get overfitted if trained with a few kinds of background appearances for a large number of training steps. 2) Besides, an overfitted model may tend to remember the relevance between co-occurred foreground and background appearances, which may cause failures when synthesizing new human foregrounds. 3) Compared to generating the fixed background appearance which can be easily described by a still image, generation of the appearance-irrelevant background shadow is more worth studying, which enables background appearance control by adding shadows to alterable user-defined background images.

Then we describe the architectures in detail to explain the above mentioned functionalities. As shown in Figure 4(b), the appearance GAN is made of an appearance generator G_A , a scene discriminator D_S , three standard body part discriminators

D_H , D_U , D_L and three appearance-consistency body part discriminators $D_{AC,H}$, $D_{AC,U}$, $D_{AC,L}$.

Generator: Specifically, G_A consists of three encoders and two decoders. The first encoder learns to encode the target foreground appearance features with $X_A^1|_{t-2} = X_{TFG}|_{t-2}^t$ as its input. The second encoder learns to encode the source motion features with $X_A^2|_{t-2} = [X_{SP}|_{t-2}^t, \tilde{Y}_{LO}|_{t-2}^t]$ as its input. The third encoder learns to encode features for previously generated foregrounds $\tilde{Y}_{FG}|_{t-2}^{t-1}$. Then the three kinds of features are summed and fed into the first decoder to generate \tilde{Y}_{FG}^t , which is the desired foreground appearance at time t . Meanwhile, features of the second and the third encoders are summed and fed into the second decoder to generate the shadow map \tilde{Y}_M^t , which is output by a sigmoid layer into the same size as the input background image X_{TBG} . Thus, the second decoder, the second and the third encoders form our **shadow extraction module**, which is light-weight because it only requires one additional decoder on the basis of foreground synthesis modules. By multiplying X_{TBG} with \tilde{Y}_M^t , the background brightness is modulated pixel by pixel to achieve shadow rendering. Since the generation has no relation to background appearance, the shadow map \tilde{Y}_M^t is identical to any X_{TBG} and therefore supports shadow rendering for arbitrary images, which enables background appearance control. Then the synthetic foreground and the rendered background are added together to compose the full image \tilde{Y}_I^t , which is the desired video scene at time t .

Discriminators: In addition, we design multiple multi-scale discriminators (D_H , D_U , D_L , $D_{AC,H}$, $D_{AC,U}$, $D_{AC,L}$ and D_S) for the three part-specific ACGAN losses and one scene GAN loss. Specifically, each part-specific **ACGAN loss** is used for the supervision of a specific body part and is made of a standard GAN loss and an appearance consistency loss, aiming at supervising visual quality and appearance consistency respectively. As for the visual quality, we decompose the generated and the real foreground appearances into the three body parts and feed them as the fake and the real samples into their corresponding standard body part discriminators D_H , D_U and D_L , forcing the generator G_A to synthesize more realistic body part appearances. As for the appearance consistency, we further apply three appearance-consistency (AC) body part discriminators $D_{AC,H}$, $D_{AC,U}$ and $D_{AC,L}$ to ensure appearances of the generated body parts are consistent with their input appearance conditions. Specifically, we obtain three kinds of body part appearance pairs as training samples for each D_{AC} as shown in Figure 4(b): 1) consistent pair P_1 : two body parts from the same person, labeled as "true"; 2) inconsistent pair P_2 : two body parts from different persons, labeled as "false"; 3) fake pair P_{fake} : body part of the generated \tilde{Y}_{FG} and the corresponding part of the input appearance condition X_{TFG} , labeled as "false" when updating discriminator and labeled as "true" when updating generator. In company with the progress of D_{ACs} that distinguish inconsistent body part appearances well, G_A learns to generate more consistent body part appearances during adversarial training. The **scene GAN loss** is designed to force the appearance generator G_A to focus on details at part boundaries and compose the full scene harmoniously, where we feed \tilde{Y}_I and Y_I as the fake and the

real samples to the scene discriminator D_S for training.

Objective Function: To train the appearance GAN, we design the objective like this:

$$\begin{aligned} L_A = & L_{ACGAN}^H + L_{ACGAN}^U + L_{ACGAN}^L + L_{GAN}^S \\ & + \lambda_T L_T^A + \lambda_{FM} L_{FM}^A + \lambda_{VGG} L_{VGG}^A \end{aligned} \quad (7)$$

$L_{ACGAN}^{H/U/L}$ are ACGAN losses of different body parts, each of which is summed by a standard GAN loss L_{GAN} and an appearance-consistency loss L_{AC} . Since all of them have the same design, we only give the derivation of L_{ACGAN}^H as an example:

$$L_{ACGAN}^H = L_{GAN}^H + \lambda_{AC} L_{AC}^H \quad (8)$$

$$\begin{aligned} L_{GAN}^H = & E[\log D_H(Y_{FG,H}, X_{A,H}) \\ & + \log[1 - D_H(\tilde{Y}_{FG,H}, X_{A,H})]] \end{aligned} \quad (9)$$

$$\begin{aligned} L_{AC}^H = & E[\log D_{AC,H}(P_{1,H}) \\ & + \log[1 - D_{AC,H}(P_{2,H})] \\ & + \log[1 - D_{AC,H}(P_{fake,H})]] \end{aligned} \quad (10)$$

where λ_{AC} is the weight of L_{AC} , $Y_{FG,H}$ represents head region of the real foreground, $X_{A,H}$ represents the conditioning input obtained by concatenating head regions of X_A^1 and X_A^2 , $P_{1,H}$, $P_{2,H}$ and $P_{fake,H}$ are consistent, inconsistent and fake head appearance pairs respectively.

L_{GAN}^S is the scene GAN loss, derived as follows:

$$L_{GAN}^S = E[\log D_S(Y_I, X_I) + \log[1 - D_S(\tilde{Y}_I, X_I)]] \quad (11)$$

where Y_I is the real scene image with respect to \tilde{Y}_I , $X_I = [X_A^1, X_A^2, X_{TBG}]$.

L_T^A is the temporal loss weighted by λ_T to improve temporal consistency, which can be derived as follows:

$$L_T^A = E[\log D_A^T(S_I) + \log[1 - D_A^T(\tilde{S}_I)]] \quad (12)$$

where D_A^T is the temporal discriminator of the appearance GAN, trained to determine whether an image sequence is real or fake. S_I and \tilde{S}_I are the real and the generated image sequences, which are obtained similarly to the layout sequences by concatenating three consecutive Y_I s and \tilde{Y}_I s.

L_{FM}^A is the discriminator feature matching loss weighted by λ_{FM} , L_{VGG}^A is the VGG loss [27], [28], [7] weighted by λ_{VGG} .

V. EXPERIMENTS

A. Solo Dance Dataset

We construct a large solo dance dataset with 124 dance videos collected from 58 males and 66 females, including a variety of human identities and clothing styles that allow for appearance generalization. Our dataset covers four main dance types (modern, jazz, rumba, tap) with the dancer in each video performing a dance different from others. Each video is an individual dance clip captured at a 30fps frame rate, containing 1200 continuous frames with the corresponding appearances and motions. To satisfy the setting that single persons perform difficult movements in stationary backgrounds, only solo dance videos with fixed viewpoints are included in the dataset.

After the videos are collected, we automatically extract backgrounds for each video by stitching detached background regions of different frames. Then we detect poses, layouts and foregrounds for each frame, where we crop and resize all the

frames to central 192x256 regions and manually rectify ones with bad detection results for better data quality. Next, we divide each processed video sequence into two halves, where the first half is used to extract X_{SPS} and the second is used to obtain the paired X_{TS} . Therefore, we obtain 600 available conditioning inputs and the corresponding ground-truth frames for each of the 124 videos. In our experiments, we use 100 videos for training and the remaining 24 videos for testing.

B. Experimental Setup

1) Our Method: The design of encoders and decoders follows pix2pixHD [7], where the numbers of convolutional filters are decreased to half of the original pix2pixHD to reduce the model size. All the discriminators that distinguish single frames (standard and AC discriminators) follow the multi-scale PatchGAN architecture [3], and each of them has three spatial scales to model different image resolutions. All the temporal discriminators that distinguish sequences rather than single frames follow the design of [16], and each of them has three time scales to ensure both short-term and long-term temporal consistency.

During the training stage, the layout GAN and the appearance GAN are trained separately with Adam optimizers (learning rate: 0.0002, batch size: 4) on 4 Nvidia RTX 2080 Ti GPUs for 10 epochs, where we set $\lambda_{AC} = 5$ and $\lambda_{SS} = \lambda_T = \lambda_{FM} = \lambda_{VGG} = 10$ in the objective functions. Since frames (layouts and foregrounds) at time -1 and -2 don't exist, we directly replace them with two same-size zero tensors to first generate the frame at time 0, which is then taken as input together with the zero tensor at time -1 to generate the frame at time 1. By doing this during training, the layout GAN and the appearance GAN learn to handle the generation of the first two frames.

2) Other Methods: We also implement the following methods for comparisons:

- **Video-based methods:**

We compare our method with two state-of-art video-based methods vid2vid [16] and EDN [15], both are video-specific with each model can only generate videos with the same scene appearance. In our implementation, each of their models is trained with 3000 frames of one specific video.

- **Image-based methods:**

Since video-based methods are video-specific, we implement a state-of-art image-based method PoseWarp [9] as a general-purpose baseline, which is trained on the same data as ours in a general way.

- **w/o input selection:**

To evaluate the effectiveness of our input selection strategy which enables input appearance control, we implement a model trained with body part appearance conditions selected randomly with no extra computation.

- **w/o layout GAN:**

To evaluate the effectiveness of our layout GAN that provides more accurate motion conditions, we implement a model with only the appearance GAN, which is fed with only 2D poses as motion conditions.

- **w/o ACGAN loss:**

To evaluate the effectiveness of our ACGAN loss which

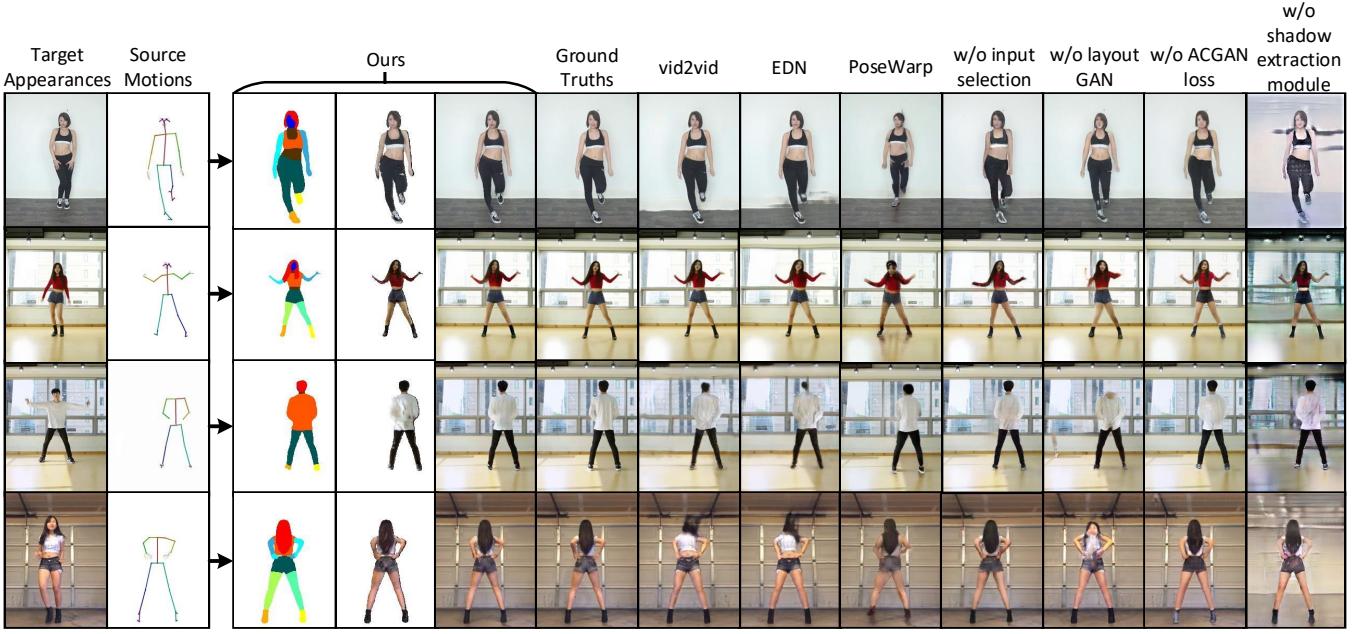


Fig. 5. Qualitative comparison results on HVMT tasks (please zoom in for a better view). From left to right: input target appearances, input source motions, our generated results (layouts, foregrounds, full scenes), ground-truth frames, results of vid2vid [16], results of EDN [15], results of PoseWarp [9], results of the four ablated variants with respect to input selection strategy, layout GAN, ACGAN loss and shadow extraction module.

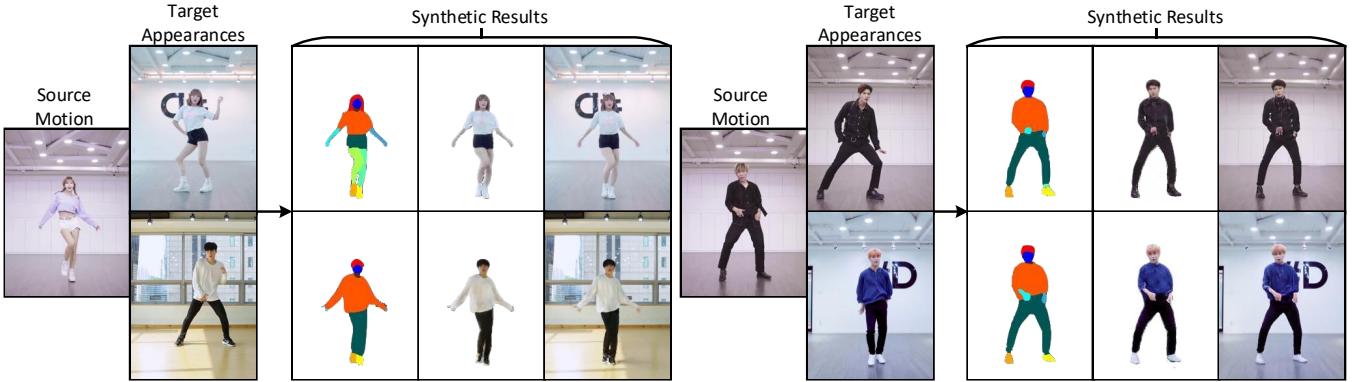


Fig. 6. Examples of ordinary HVMT (please zoom in for a better view). Each synthetic result (layout, foreground, full scene) is generated to have the same motion as its input source motion image and the same appearance as its input target appearance image.

enables foreground appearance control, we implement a model whose appearance GAN is trained without ACGAN loss.

- **w/o shadow extraction module:**

To evaluate the effectiveness of our shadow extraction module which enables background appearance control, we implement a model that generates backgrounds from scratch with fixed background images included in its input appearance conditions.

C. Qualitative Results

To assess the quality of our synthetic results, we test different methods on our test set and compare their synthetic frames with ground-truth video frames. It's noted that ground truths are available here because the motion and the appearance conditions of each synthetic frame are obtained from the same person as has been stated in the description of our dataset

(Sec.V-A). As shown in Figure 5, we randomly visualize some synthetic frames generated by our method and other methods to make qualitative comparisons. Based on the proposed GAC-GAN, we can synthesize motion transfer videos with realistic appearance and body pose details, which are consistent with the input target appearances and source motions. In contrast, the image-based method PoseWarp [9] can't preserve the target appearances well with body poses and locations are not consistent with the desired source motions. Although the two video-based methods vid2vid [16] and EDN [15] perform well when synthesizing appearances of frequent poses (e.g., front bodies in the first two rows of Figure 5), they render bad visual results when synthesizing appearances of infrequent poses (e.g., backside bodies in the last two rows of Figure 5). We think the main reason is that infrequent poses are less explored during training due to the imbalance between numbers of frequent and infrequent poses in their training data, which contains only one video sequence for each video-

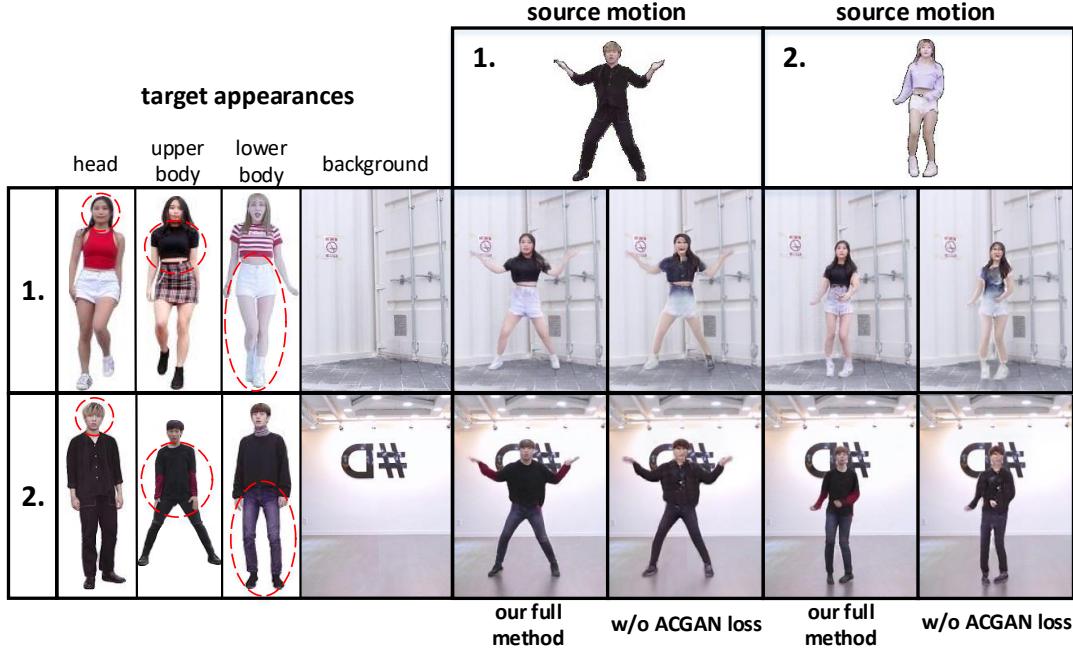


Fig. 7. Examples of multi-source appearance control (please zoom in for a better view). Each synthetic image is generated based on five inputs in terms of body motion, appearances of head, upper body, lower body and background. We also show the results generated by the variant model without ACGAN loss, allowing for comparisons with our full method.

specific model. However, the quality of our results is not influenced by such imbalance because we provide our model with optimal appearance inputs that contain the maximum texture information needed for appearance synthesis. Besides, our model is trained with access to more infrequent poses contained in the whole dataset, leading to better results than EDN and vid2vid when synthesizing unseen infrequent pose appearances. Please refer to our supplementary material for the video version of the qualitative comparison results.

Then we test our method on tasks which have no ground-truth frame for a better understanding of the appearance-controllable human video motion transfer that we realize:

• Ordinary HVMT:

Transfer one person's motion to another person without further appearance control on body part foregrounds and surrounding backgrounds, which has the same test setting as the qualitative comparisons shown in Figure 5 except for the absence of ground-truth frames. As shown in Figure 6, our one-time trained model can generate high-quality motion transfer video frames with details of the source motions and the target appearances are well preserved.

• Appearance-Controllable HVMT:

For HVMT with multi-source appearance control, we let our model synthesize videos with appearances of body parts and backgrounds coming from different appearance sources, where the synthetic appearances are naturally composed and the synthetic motions are consistent with the source motions as shown in Figure 7. For further evaluation on background appearance control based on our synthetic shadow maps, we add shadows to different backgrounds and fuse them with the synthetic foregrounds to achieve controllable background replacement, where

detailed shadows are rendered in harmony with the human motions as shown in Figure 8.

For a full and animated version of the synthetic visual results, please refer to our supplementary materials.

D. Quantitative Results

We also make a quantitative assessment to analyze differences between synthetic and ground-truth video frames by four metrics: Structural Similarity (SSIM), Peak Signal to Noise Ratio (PNSR), Learned Perceptual Image Patch Similarity (LPIPS) [24] and Video Fréchet Inception Distance (VFID) [16]. In particular, SSIM and PSNR are classic metrics that measure the pixel-level image similarity between synthetic results and ground truths, which are simple and based on shallow functions. LPIPS is a newly invented metric that accounts for similarity measurement between two images, which is computed based on features extracted by deep models. VFID is also a deep metric with a video recognition CNN model performing as its feature extractor, which measures temporal consistency in addition to visual quality. It's noted that SSIM and PSNR are similarity metrics while LPIPS and VFID are distance metrics, which means higher values are better for the former while the opposite for the latter. All the comparison results are summarized in the first five rows of Table I. We can see that our method outperforms other methods for all the metrics, which indicates that our synthetic results have not only higher visual quality but also better temporal consistency.

E. Human Perceptual Results

For human perceptual assessment, we conduct a human subjective study by performing preference tests on the Amazon

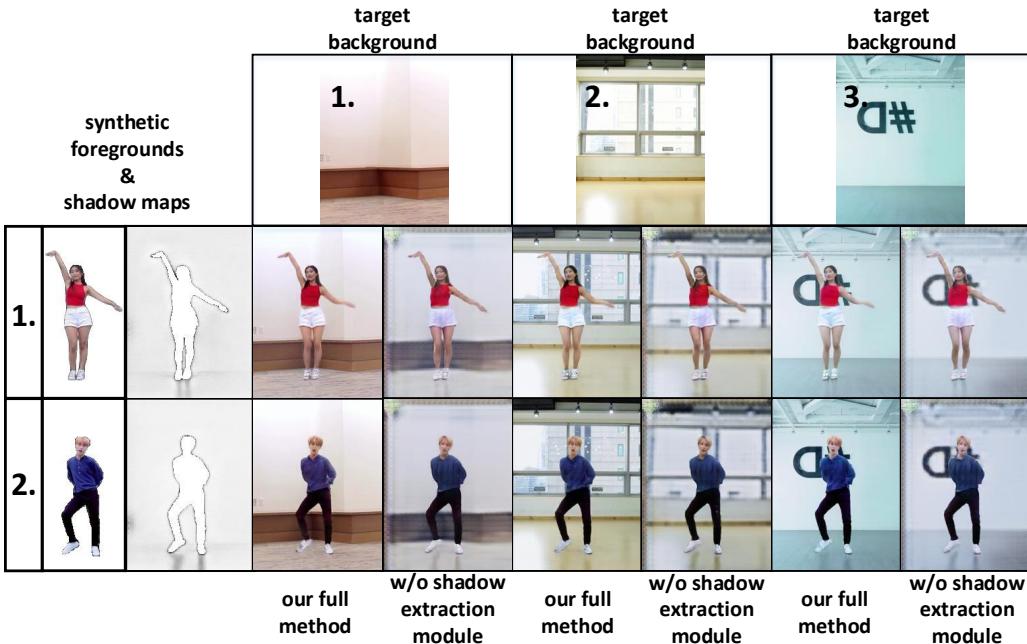


Fig. 8. Examples of background appearance control (please zoom in for a better view). The input backgrounds are modulated by the synthetic shadow maps to fuse with the synthetic foregrounds. We also show the results generated by the variant model without shadow extraction module, allowing for comparisons with our full method.

TABLE I

QUANTITATIVE AND PERCEPTUAL COMPARISON RESULTS. SSIM AND PSNR ARE SIMILARITY METRICS, THE HIGHER THE BETTER. LPIPS AND VFID ARE DISTANCE METRICS, THE LOWER THE BETTER. PREFERENCE SCORE IS DENOTED AS THE PROPORTION OF PERCEPTUALLY PREFERRED VIDEOS GENERATED BY OUR METHOD.

	vid2vid [16]	EDN [15]	PoseWarp [9]	w/o input selection	w/o layout GAN	w/o ACGAN loss	w/o shadow extraction module	Ours
SSIM	0.8834	0.8711	0.8380	0.8652	0.8545	0.8613	0.8580	0.8947
PSNR	26.9316	26.5653	23.5423	24.7923	22.5398	24.1372	24.0782	27.9458
LPIPS	0.0352	0.0363	0.0537	0.0413	0.0436	0.0394	0.0419	0.0341
VFID	3.9752	4.3410	7.0721	5.1426	5.3624	4.9845	5.2187	3.9689
Preference Score	69.2%	73.1%	93.8%	76.2%	81.5%	77.7%	80.8%	—

Mechanical Turk (AMT). Particularly, each question is an A/B test where we show turkers two videos generated by our method and a compared method and let them choose which video looks more realistic in consideration of visual quality and temporal consistency. After gathering 10 answers for 13 videos generated by different methods, we summarize the average human preference scores in the last row of Table I. The results indicate that videos generated by our method are also perceptually preferred to those generated by others, which is consistent with our qualitative and quantitative results. It's noted that each preference score in the table represents the proportion of perceptually preferred videos generated by our full method when compared to one of the compared methods. Since the comparison between our full method and itself is missing, the preference score of our full method is left blank.

F. Ablation Studies

We also compare our full method with the above mentioned four variants with respect to ablations of our input selection strategy, layout GAN, ACGAN loss and shadow extraction module. As can be seen from the quantitative and the perceptual

results shown in the 5-8th columns of Table I, our full method outperforms all the variants significantly, which indicates that videos generated by our full method have higher visual quality and better temporal consistency than those generated by the variants without our important components. We also make comparisons on qualitative results as shown in the last four columns of Figure 5. The 10th and the 12th columns indicate that, without the selected optimal appearance inputs and the elaborate ACGAN loss, the model can't preserve human appearances well, which results in blurry faces and bodies, proving that both the two components can improve human appearance details. The 11th column indicates that, without the layout GAN to provide additional motion conditioning inputs that describe human motions more accurately, the model even can't generate the desired body motions, let alone satisfactory appearances, which proves the effectiveness of our two-stage framework design. The last column shows that, without shadow rendering, the model fails in both background and foreground synthesis, which proves that our shadow extraction module can also improve foreground synthesis quality. Moreover, we make further qualitative comparisons to demonstrate the importance

of our ACGAN loss and shadow extraction module in the multi-source foreground and background appearance control respectively. As shown in Figure 7, without training with the proposed ACGAN loss, the model renders bad body part appearances which are mixed up together and inconsistent with the input appearance conditions. As shown in Figure 8, without the synthetic shadow maps to achieve shadow rendering, the model can only generate background appearances from scratch, which results in blurry backgrounds as well as unrealistic foregrounds.

VI. CONCLUSION

In this paper, we present GAC-GAN for general-purpose and appearance-controllable human video motion transfer. To synthesize videos with controllable appearances, we propose a multi-source input selection strategy to first obtain controllable input appearance conditions. Moreover, given such appearance-controllable inputs, we propose a two-stage GAN framework trained with the ACGAN loss and implanted with the shadow extraction module to enable the compatible synthesis of the appearance-controllable outputs. Extensive experiments on our large-scale solo dance dataset show that our proposed method can not only enable appearance control in a general way but also achieve higher video quality than state-of-art methods. We also conduct comprehensive ablation studies with respect to our input selection strategy, layout GAN, ACGAN loss and shadow extraction module. The results show that our full method achieves higher performance than all the ablated variants, which proves the effectiveness of our important components. Although our method performs well in most cases, challenges and open problems remain: 1) Since the GAC-GAN is a deep learning model, our method may fail when the GAN model is tested on unseen domains that are too different from the training domains (e.g., generate videos for CG characters rather than real humans). 2) Because the quality of the synthetic outputs highly depends on the conditioning inputs which are obtained based on pose and layout estimation techniques, texture artifacts may occur when these estimations fail. In the future, we may also explore the potential of synthesizing more complex videos where multiple people dance together rather than solo dance videos. Besides, video synthesis with movable camera views is also worth studying, requiring further consideration of background motions. Both of them are promising extensions to our accomplished work.

REFERENCES

- [1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [2] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *arXiv preprint arXiv:1411.1784*, 2014.
- [3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
- [4] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, “Instance-level human parsing via part grouping network,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 770–785.
- [5] X. Liang, K. Gong, X. Shen, and L. Lin, “Look into person: Joint body parsing & pose estimation network and a new benchmark,” *IEEE Trans. pattern analysis and machine intelligence*, vol. 41, no. 4, pp. 871–885, 2018.
- [6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.
- [7] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional gans,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.
- [8] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, “Soft-gated warping-gan for pose-guided person image synthesis,” in *Advances in Neural Information Processing Systems*, 2018, pp. 474–484.
- [9] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, “Synthesizing images of humans in unseen poses,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8340–8348.
- [10] M. Wang, G.-Y. Yang, R. Li, R.-Z. Liang, S.-H. Zhang, P. M. Hall, and S.-M. Hu, “Example-guided style-consistent image synthesis from semantic labeling,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1495–1504.
- [11] N. Neverova, R. Alp Guler, and I. Kokkinos, “Dense pose transfer,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 123–138.
- [12] D. Liang, R. Wang, X. Tian, and C. Zou, “Pegan: Partition-controlled human image generation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8698–8705.
- [13] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, “Deformable gans for pose-based human image generation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3408–3416.
- [14] M. Zanfir, A.-I. Popa, A. Zanfir, and C. Sminchisescu, “Human appearance transfer,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5391–5399.
- [15] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, “Everybody dance now,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5933–5942.
- [16] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, “Video-to-video synthesis,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2018, pp. 1152–1164.
- [17] L. Liu, W. Xu, M. Zollhöfer, H. Kim, F. Bernard, M. Habermann, W. Wang, and C. Theobalt, “Neural rendering and reenactment of human actor videos,” *ACM Trans. Graph.*, vol. 38, no. 5, pp. 139:1–139:14, Oct. 2019. [Online]. Available: <http://doi.acm.org/10.1145/3333002>
- [18] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, “Dance dance generation: Motion transfer for internet videos,” *arXiv preprint arXiv:1904.00129*, 2019.
- [19] K. Aberman, M. Shi, J. Liao, D. Liscinski, B. Chen, and D. Cohen-Or, “Deep video-based performance cloning,” in *Computer Graphics Forum*, vol. 38, no. 2. Wiley Online Library, 2019, pp. 219–233.
- [20] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, “Recognizing action at a distance,” in *null*. IEEE, 2003, p. 726.
- [21] J. Lee and S. Y. Shin, “A hierarchical approach to interactive motion editing for human-like figures,” in *Siggraph*, vol. 99, 1999, pp. 39–48.
- [22] C. Hecker, B. Raabe, R. W. Enslow, J. DeWeese, J. Maynard, and K. van Prooijen, “Real-time motion retargeting to highly varied user-created morphologies,” in *ACM Transactions on Graphics (TOG)*, vol. 27, no. 3. ACM, 2008, p. 27.
- [23] G. K. Cheung, S. Baker, J. Hodgins, and T. Kanade, “Markerless human motion transfer,” in *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004*. IEEE, 2004, pp. 373–378.
- [24] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
- [25] A. Schödl, R. Szeliski, D. H. Salesin, and I. Essa, “Video textures,” in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 2000, pp. 489–498.
- [26] A. Schödl and I. A. Essa, “Controlled animation of video sprites,” in *Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*. ACM, 2002, pp. 121–127.
- [27] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *European conference on computer vision*. Springer, 2016, pp. 694–711.

- [28] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Advances in neural information processing systems*, 2016, pp. 658–666.
- [29] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *International journal of computer vision*, vol. 40, no. 1, pp. 49–70, 2000.
- [30] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [31] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a laplacian pyramid of adversarial networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, p. 1486?494.
- [32] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," 09 2016.
- [33] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in neural information processing systems*, 2017, pp. 700–708.
- [34] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
- [35] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2642–2651.
- [36] T. Miyato and M. Koyama, "cGANs with projection discriminator," in *International Conference on Learning Representations*, 2018. [Online]. Available: <https://openreview.net/forum?id=ByS1VpgRZ>
- [37] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Advances in neural information processing systems*, 2016, pp. 613–621.
- [38] M. Saito, E. Matsumoto, and S. Saito, "Temporal generative adversarial nets with singular value clipping," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2830–2839.
- [39] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1526–1535.
- [40] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," in *4th International Conference on Learning Representations, ICLR 2016*, 2016.
- [41] X. Liang, L. Lee, W. Dai, and E. P. Xing, "Dual motion gan for future-flow embedded video prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1744–1752.
- [42] E. L. Denton *et al.*, "Unsupervised learning of disentangled representations from video," in *Advances in neural information processing systems*, 2017, pp. 4414–4423.
- [43] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black, "Clothcap: Seamless 4d clothing capture and retargeting," *Acm Transactions on Graphics*, vol. 36, no. 4, pp. 1–15, 2017.
- [44] V. Leroy, J. S. Franco, and E. Boyer, "Multi-view dynamic shape refinement using local temporal integration," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [45] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll, "Detailed, accurate, human shape estimation from clothed 3d scan sequences," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 5484–5493.
- [46] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on Robot Learning*, 2017, pp. 1–16.
- [47] A. Shysheya, E. Zakharov, K.-A. Aliev, R. Bashirov, E. Burkov, K. Iskakov, A. Ivakhnenko, Y. Malkov, I. Pasechnik, D. Ulyanov *et al.*, "Textured neural avatars," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2387–2397.
- [48] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt, "Monoperfcap: Human performance capture from monocular video," *ACM Transactions on Graphics (ToG)*, vol. 37, no. 2, pp. 1–15, 2018.



Dongxu Wei (S'19) received his B.S. degree in electronic science and technology from Harbin Institute of Technology. He is currently pursuing a Ph.D degree with the college of Information Science & Electronic Engineering of Zhejiang University since 2017, advised by Prof. Haibin Shen. His research interests include image and video understanding, image and video synthesis, and computational photography.



Xiaowei Xu (S'14-M'17) received the B.S. and Ph.D. degrees in electronic science and technology from Huazhong University of Science and Technology, Wuhan, China, in 2011 and 2016 respectively. He worked as a post-doc researcher at University of Notre Dame, IN, USA from 2016 to 2019. He is now an AI researcher at Guangdong Provincial People's Hospital. His research interests include deep learning, and medical image segmentation. He was a recipient of DAC system design contest special service recognition reward in 2018 and outstanding

contribution in reviewing, Integration, the VLSI journal in 2017. He has served as TPC members in ICCD, ICCAD, ISVLSI and ISQED. He has published more than 50 papers in international peer-reviewed conference proceedings and journals.



Haibin Shen, male, born in 1967, Ph.D., is a professor of Zhejiang University, the deputy director of HPEC (High Performance Embedded Computing) Key Lab of Ministry of Education and a member of the second level of 151 talents project of Zhejiang Province. His research interests are focused on learning algorithm, processor architecture and modeling. He has long-term working experience in system and IC both in university and industry. He has published over 30 SCI&EI indexed papers in academic journals, and was granted more than 20 patents for inventions.

He was also in charge of more than 10 national projects, including those from National Science and Technology Major Project and National High Technology Research and Development Project (863 Project), and participated in the projects of Chinese Nature Science Foundation and the formulation of national standards. His research achievement has been used by many authority organizations, and he was awarded the first prize of Electronic Information Science and Technology Award from Chinese Institute of Electronics.



Kejie Huang (M'13-SM'18) received his Ph.D degree from the Department of Electrical Engineering, National University of Singapore (NUS), Singapore, in 2014. He has been a principal investigator at College of Information Science & Electronic Engineering, Zhejiang University (ZJU) since 2016. Prior to joining ZJU, he spent five years in the IC design industry including Samsung and Xilinx, two years in the Data Storage Institute, Agency for Science Technology and Research (A*STAR), and another three years in Singapore University of (SUTD), Singapore. He has authored or coauthored more than 30 scientific papers in international peer-reviewed journals and conference proceedings. He holds four granted international patents, and another eight pending ones. His research interests include low power circuits and systems design using emerging non-volatile memories, architecture and circuit optimization for reconfigurable computing systems and neuromorphic systems, machine learning, and deep learning chip design. He currently serves as the Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMSPART II: EXPRESS BRIEFS.