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

Dongxu Wei, Xiaowei Xu, Haibin Shen, and Kejie Huang, *Senior Member, IEEE*

***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 encoded individually in the trained models. Therefore, each trained model can only generate videos with a specific scene appearance, and 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 General Appearance-Controllable GAN (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 Appearance-Consistency GAN (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 on our solo dance dataset and iPER dataset 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 Virtual Reality (VR) / Augmented Reality (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 [1], [2]. Moreover, the animated visual data can be further utilized as simulated training data to train robotic agents that work for real-world situations, where real experiences may be hard to obtain [3].

With the recent emergence of Generative Adversarial Networks (GANs) [4] and its variant conditional GANs (cGANs) [5], there are many GAN-based works [6]–[10] 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, besides the human motion control, further appearance control is needed to provide high flexibility in practical applications.

In this work, we propose a two-stage GAN-based framework named "GAC-GAN" to address the limitations described above. Specifically, "G" refers to "General", which means the proposed GAC-GAN is a general method that can generate appearances for arbitrary human subjects. "AC" refers to "Appearance-Controllable", which means our approach can provide the flexibility to control the synthetic scene appearances. For general-purpose appearance synthesis, we propose to feed our model with appearance conditioning inputs in addition to motion conditioning inputs (e.g., body poses) used in other works, allowing the model to learn the mapping from

This work was supported by the National Natural Science Foundation of China (U19B2043). *(Corresponding author: Kejie Huang.)*

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. Xu is with Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China, 510080. (Email: xiao.wei.xu@foxmail.com)

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes video versions of the qualitative results. This material is 84.2 MB in size.

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 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 Appearance-Consistency GAN (ACGAN) loss and a light-weight shadow extraction module to the appearance GAN to achieve control of the output human foreground and background respectively. Compared with single-stage frameworks employed by previous methods [6]–[10], our two-stage framework has two advantages: 1) With the help of body layouts generated in the first stage, we can precisely separate and compose different body parts to enable multi-source appearance generation and thus achieve appearance control, which is incapable for single-stage methods that directly generate appearances from body pose points. 2) Benefiting from the two-stage design, we can divide the appearance synthesis task into two easier subtasks, which can ease training as well as improve performance.

In our experiments, a large solo dance dataset including 148800 frames collected from 124 people and iPER dataset [11] are utilized for general-purpose training and evaluation. We first compare our approach against state-of-art video-specific [6], [7] and general-purpose [11], [12] 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. In Sec.VI, we discuss our limitations and possible future directions. Finally, Sec.VII concludes the paper.

II. Related Work

A. Classic Motion Transfer

Early works have attempted to reorder existing video frames [13]–[15] to obtain new videos consisting of frames with facial or bodily 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 [16]–[18] to create rendered motion transfer videos, which results in coarse body silhouettes and unrealistic texture details. Recently, methods [19]–[21] 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 [22]. With the development of deep learning algorithms, variational autoencoder (VAE) [23] and generative adversarial networks (GANs) [4] 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 to be more exploited in image and video generation works. Unconditional GAN-based image generation works [24], [25] 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) [5], works manage to take class labels [26], [27] or descriptive images [28], [29] or both of them [30] as extra conditioning inputs to control the output image appearances, which belongs to the same method category as our proposed cGAN-based approach. Besides image generation, there are also works [31]–[36] focus on synthesizing temporally coherent video sequences. For instance, unconditional video generation works [31]–[33] 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 randomly synthesized in an unconditional manner. Besides, video prediction techniques [34]–[36] 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.

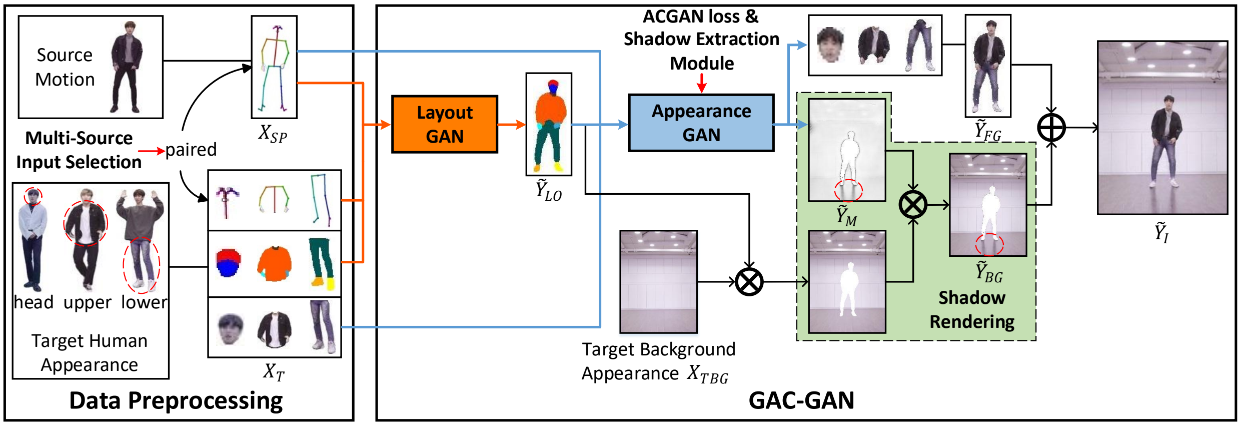

Fig. 1. Overview of our method. In the data preprocessing, we obtain the paired X_{SP} and X_T from the source motion frame and the target appearance frames based on the multi-source input selection strategy. Then the processed inputs are fed into the GAC-GAN which consists of a layout GAN and an appearance 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.

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 [11], [12], [37]–[42] 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 [28], which is very similar to the human video motion transfer that we focus on. [12], [37], [38] 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, [39], [40] apply such transformations or deformations to appearance features instead of textures, where the transformed features are then decoded to generate new pose images. Moreover, a recent work [11] proposes to warp appearances based on 3D mesh correspondences rather than 2D transformations, which achieves superior performance. Instead of blending or warping existing images and features, [41] designs a multistage GAN framework conditioned on different pose priors to directly generate pose images. Furthermore, [42] proposes 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 are designed for still image generation without consideration of temporal coherence, which causes them to be 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 [7], the authors propose to generate optical flows to warp previously generated frames into temporally consistent new frames. Besides, [6] uses 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 [8] or transformed images of body parts [9] into their models as input conditions to obtain realistic appearances. Moreover, [10] splits the 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 [9] 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

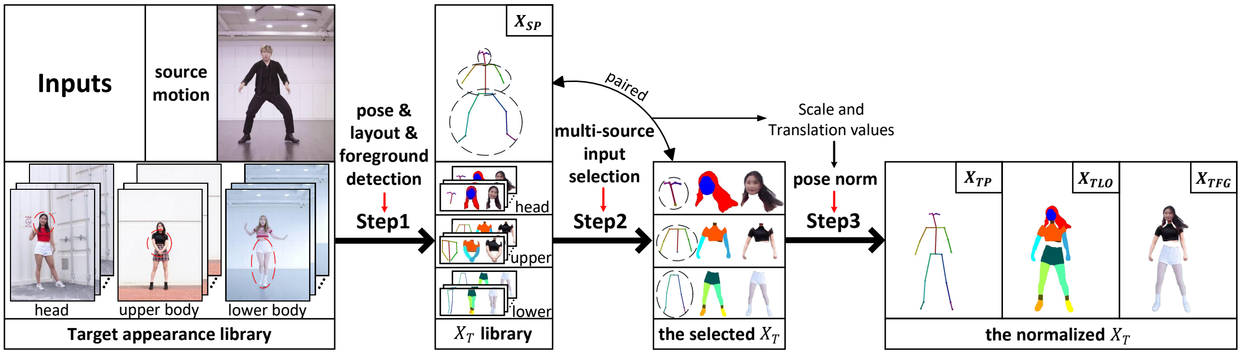

Fig. 2. Illustration of our data preprocessing. In step 1, we detect poses, layouts and foregrounds to obtain the motion condition X_{SP} and the appearance condition (X_T) library, where each red circle specifies a target body part. In step 2, each body part of the source pose X_{SP} is paired with 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}.

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: $X_{TP,H}$, $X_{TP,U}$, $X_{TP,L}$
     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. 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 obtain the optimal appearance condition no matter how many frames are provided.

1) Step 1: Detecting Poses, Layouts and Foregrounds: We utilize [43] and [44] 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 obtain 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 selected body part appearance condition has the largest pose similarity with the body part motion condition within the 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:

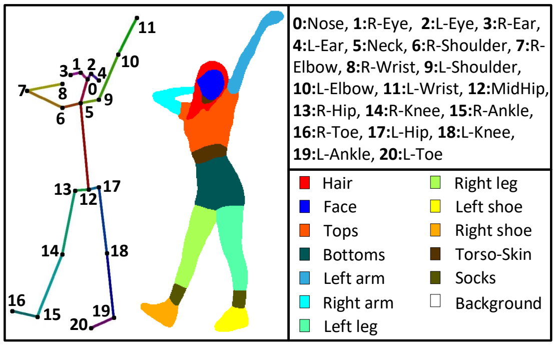
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
0:Nose, 1:R-Eye, 2:L-Eye, 3:R-Ear, 4:L-Ear, 5:Neck, 6:R-Shoulder, 7:R-Elbow, 8:R-Wrist, 9:L-Shoulder, 10:L-Elbow, 11:L-Wrist, 12:MidHip, 13:R-Hip, 14:R-Knee, 15:R-Ankle, 16:R-Toe, 17:L-Hip, 18:L-Knee, 19:L-Ankle, 20:L-Toe
Hair
Face
Tops
Bottoms
Left arm
Right arm
Left leg
Right leg
Left shoe
Right shoe
Torso-Skin
Socks
Background

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".

$$Sim = \frac{1}{N}\sum_{i=1}^{N}\frac{\overrightarrow{V_S^i}\cdot\overrightarrow{V_T^i}}{|\overrightarrow{V_S^i}|\,|\overrightarrow{V_T^i}|} \qquad (1)$$

where Sim is the body part pose similarity, $\overrightarrow{V_S^i}$ and $\overrightarrow{V_T^i}$ represent the i-th body part pose vectors of the source pose X_{SP} and the target pose X_{TP} respectively, $|\overrightarrow{V_S^i}|$ and $|\overrightarrow{V_T^i}|$ represent vector lengths of $\overrightarrow{V_S^i}$ and $\overrightarrow{V_T^i}$ 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 $\overrightarrow{P_0P_1}$, $\overrightarrow{P_0P_2}$, $\overrightarrow{P_1P_3}$, $\overrightarrow{P_2P_4}$, $\overrightarrow{P_0P_5}$. Upper body pose vectors are $\overrightarrow{P_5P_6}$, $\overrightarrow{P_6P_7}$, $\overrightarrow{P_7P_8}$, $\overrightarrow{P_5P_9}$, $\overrightarrow{P_9P_{10}}$, $\overrightarrow{P_{10}P_{11}}$, $\overrightarrow{P_5P_{12}}$. Lower body pose vectors are $\overrightarrow{P_{12}P_{13}}$, $\overrightarrow{P_{13}P_{14}}$, $\overrightarrow{P_{14}P_{15}}$, $\overrightarrow{P_{15}P_{16}}$, $\overrightarrow{P_{12}P_{17}}$, $\overrightarrow{P_{17}P_{18}}$, $\overrightarrow{P_{18}P_{19}}$, $\overrightarrow{P_{19}P_{20}}$. 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:

$$\begin{aligned} Scale &= \frac{\sum_{i=1}^{N_v}|\overrightarrow{V_S^i}|}{\sum_{i=1}^{N_v}|\overrightarrow{V_T^i}|} \\ Translation &= \frac{1}{N_p}\sum_{j=1}^{N_p}(P_S^j - P_T^j) \end{aligned} \qquad (2)$$

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 representing 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 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 [6]–[10] 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 [28] to determine whether the generated layout is real or fake.

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

$$L_{LO} = L_{GAN}^{LO} + \lambda_{SS} L_{SS}^{LO} + \lambda_{T} L_{T}^{LO} + \lambda_{FM} L_{FM}^{LO} \tag{3}$$

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

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

where Y_{LO} is the real layout map with respect to $\tilde{Y}_{LO}$.
L_{SS}^{LO} is the structural sensitive loss adapted from [45] 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:

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

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 [7] which minimizes the difference between the real and the generated sequences to improve temporal consistency, as can be given by:

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

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_{LO}s and $\tilde{Y}_{LO}$s sampled by the sampling operator presented in vid2vid [7].
L_{FM}^{LO} is the discriminator feature matching loss presented in pix2pixHD [46] and weighted by λ_{FM}, which is used to stabilize training as well as 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

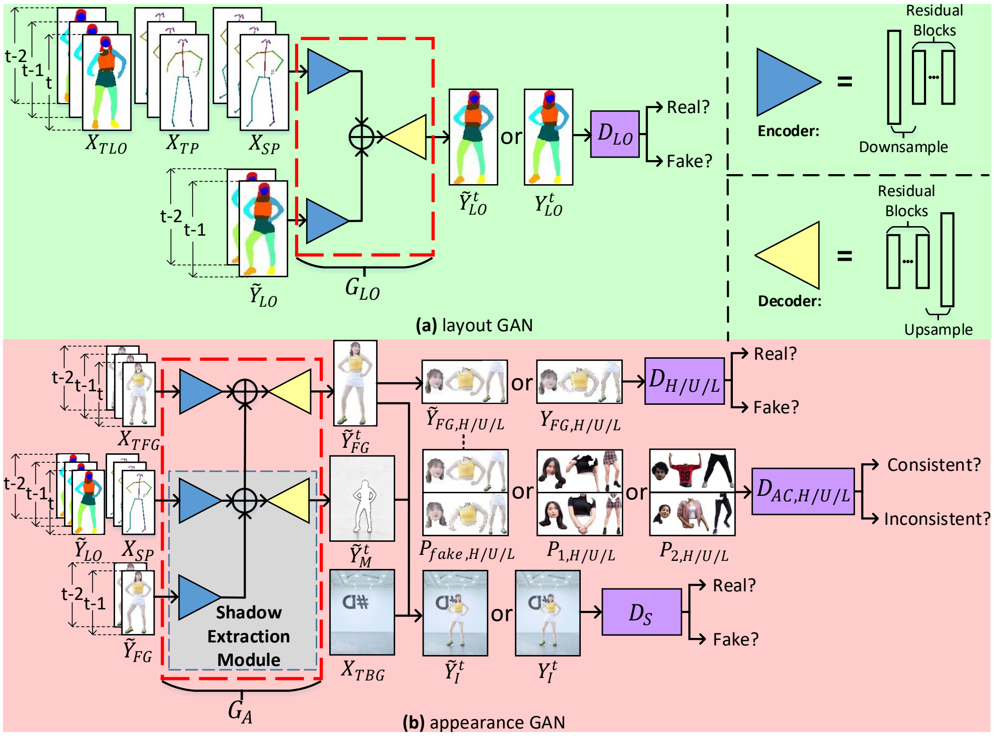

Fig. 4. Illustration of the GAC-GAN. (a) and (b) depict frameworks of the layout GAN and the appearance GAN respectively, where encoder and decoder architectures are also drawn above. In (b), foregrounds and discriminators of the three body parts are drawn in the same blocks annotated by $H/U/L$ for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

background appearance from scratch [6]–[10]. 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 $X^1_A|^t_{t-2} = X_{TFG}|^t_{t-2}$ as its input. The second encoder learns to encode the source motion features with $X^2_A|^t_{t-2} = [X_{SP}|^t_{t-2}, \tilde{Y}_{LO}|^t_{t-2}]$ as its input. The third encoder learns to encode features for previously generated foregrounds $\tilde{Y}_{FG}|^{t-1}_{t-2}$. Then the three kinds of features are summed and fed into the first decoder to generate $\tilde{Y}^t_{FG}$, 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 $\tilde{Y}^t_M$, which is output by a sigmoid layer into the same size as the input background image X_{TBG}. Particularly, the second and the third encoders are reusable across foreground generation and shadow map generation. Thus, our **shadow extraction module** is light-weight because it only requires the second decoder in addition to the structures needed for foreground generation. By multiplying X_{TBG} with $\tilde{Y}^t_M$, the background brightness is modulated pixel by pixel to achieve shadow rendering. Since the generation has no relation to background appearance, the shadow map $\tilde{Y}^t_M$ 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 $\tilde{Y}^t_I$, 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_{AC}s 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:

$$\begin{aligned} L_A =& L^H_{ACGAN} + L^U_{ACGAN} + L^L_{ACGAN} + L^S_{GAN} \\ &+ \lambda_T L^A_T + \lambda_{FM} L^A_{FM} + \lambda_{VGG} L^A_{VGG} \end{aligned} \tag{7}$$

$L^{H/U/L}_{ACGAN}$ 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^H_{ACGAN} as an example:

$$L^H_{ACGAN} = L^H_{GAN} + \lambda_{AC} L^H_{AC} \tag{8}$$

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

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

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^1_A and X^2_A, $P_{1,H}$, $P_{2,H}$ and $P_{fake,H}$ are consistent, inconsistent and fake head appearance pairs respectively.

L^S_{GAN} is the scene GAN loss, derived as follows:

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

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

Similar to L^{LO}_T, L^A_T weighted by λ_T is the temporal loss [7] used to improve temporal consistency, which is given by:

$$L^A_T = E[log D^T_A(S_I) + log[1 - D^T_A(\tilde{S}_I)]] \tag{12}$$

where D^T_A 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_Is and $\tilde{Y}_I$s.

L^A_{FM} is the discriminator feature matching loss [46] weighted by λ_{FM}, and L^A_{VGG} is the VGG loss [46]–[48] weighted by λ_{VGG}. Both can be used to stabilize training while improving synthesis quality.

V. Experiments

A. Datasets

1) 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 various dance types with 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_{SP}s and the second is used to obtain the paired X_Ts. Since both X_{SP}s and X_Ts come from the same human subjects, we can obtain 600 pairs of conditioning inputs and ground-truth frames which enable supervision for each video. In our experiments, we use 100 videos for training and the remaining 24 videos for testing.

2) Impersonator (iPER) Dataset: Besides the experiments conducted on our own dataset, we also test and compare our GAC-GAN with other methods on the iPER dataset proposed in [11], which consists of videos with people performing simple actions rather than complex dance moves. Following the original protocol in [11], we use 164 videos for training and the remaining 42 videos for testing. To obtain our conditioning inputs, we apply data preprocessing to these videos in the same way as the preprocessing of our own dataset.

B. Experimental Setup

1) Our Method: The design of encoders and decoders follows pix2pixHD [46], 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 [28], 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 [7], 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. Except λ_{SS} and λ_{AC}, all the other hyper-parameters including λ_T, λ_{FM} and λ_{VGG} follow the settings ($\lambda_T = \lambda_{FM} = \lambda_{VGG} = 10$) presented in [46] and [7]. As for λ_{SS} and λ_{AC}, we train several variant models with these two parameters set to different values and choose the optimal ones ($\lambda_{SS} = 10$, $\lambda_{AC} = 5$) to trade off the effects of our loss functions better. To stabilize the training, we employ the Least Squares GAN (LSGAN) loss [49] which can overcome the vanishing gradient problem rather than employ the regular sigmoid cross entropy GAN loss in our implementation. We further adopt the Two-Timescale Update Rule (TTUR) presented in [50] to balance the learning speed between generators and discriminators, which can also make the training more stable. We also note that, since frames (layouts and foregrounds) at time -1 and -2 don't exist, we directly replace them with two 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 [7] and EDN [6], 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 two state-of-art image-based methods PoseWarp [12] and LWGAN [11] as general-purpose baselines, which are trained on the same data as ours in a general way.
- **Without 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.
- **Without 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.
- **Without ACGAN Loss**:
  To evaluate the effectiveness of our ACGAN loss which enables foreground appearance control, we implement a model trained without ACGAN loss.
- **Without 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 two image-based methods PoseWarp [12] and LWGAN [11] can't preserve the target appearances well with body poses and locations not very consistent with the desired source motions. Although the two video-based methods vid2vid [7] and EDN [6] 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-specific model. Although including more infrequent poses in the training video can eliminate such imbalance, it requires more manpower cost for data collection and cleaning, which would further limit their flexibility and efficiency. 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 well-preserved details of the source motions and the target appearances.
- **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 of 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.

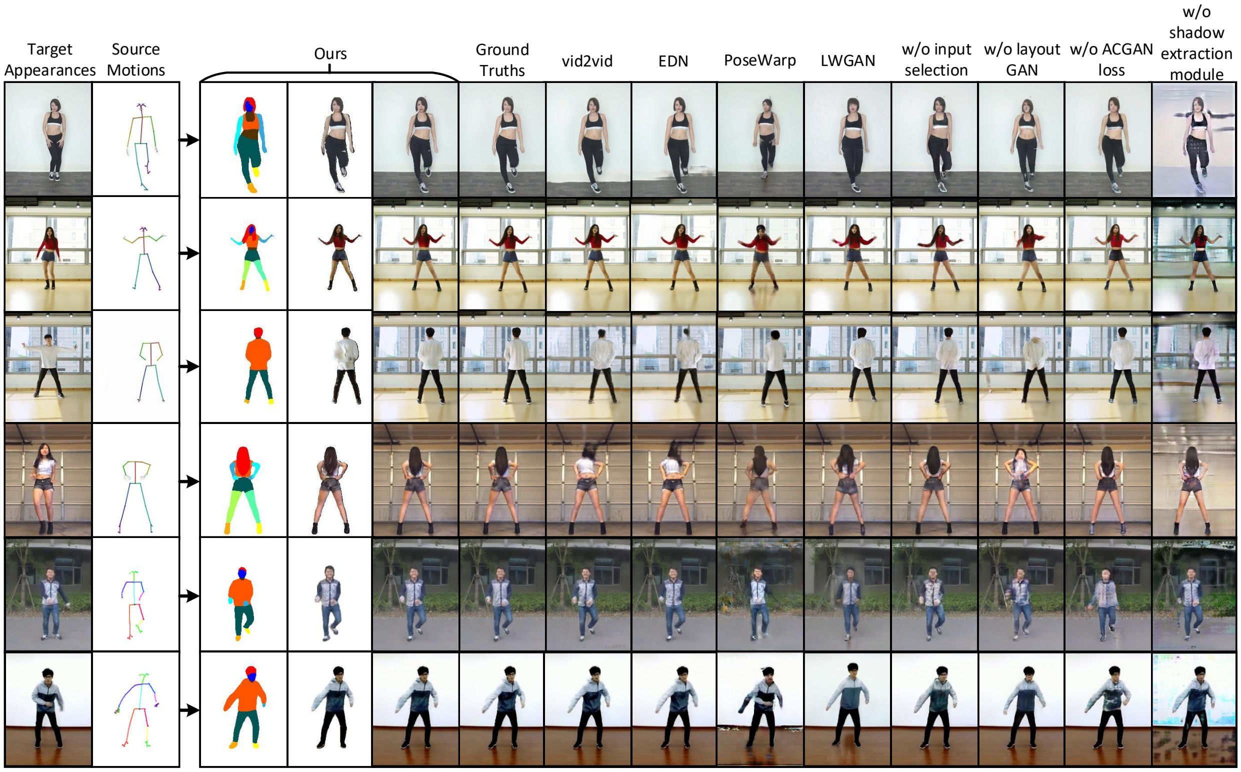

Fig. 5. Qualitative comparison results on HVMT tasks (please zoom in for a better view). The first four rows are results tested on our solo dance dataset. The last two rows are results tested on iPER [11] dataset. Columns from left to right are: input target appearances, input source motions, our generated results (layouts, foregrounds, full scenes), ground-truth frames, results of vid2vid [7], results of EDN [6], results of PoseWarp [12], results of LWGAN [11], results of the four ablated variants with respect to input selection strategy, layout GAN, ACGAN loss and shadow extraction module.

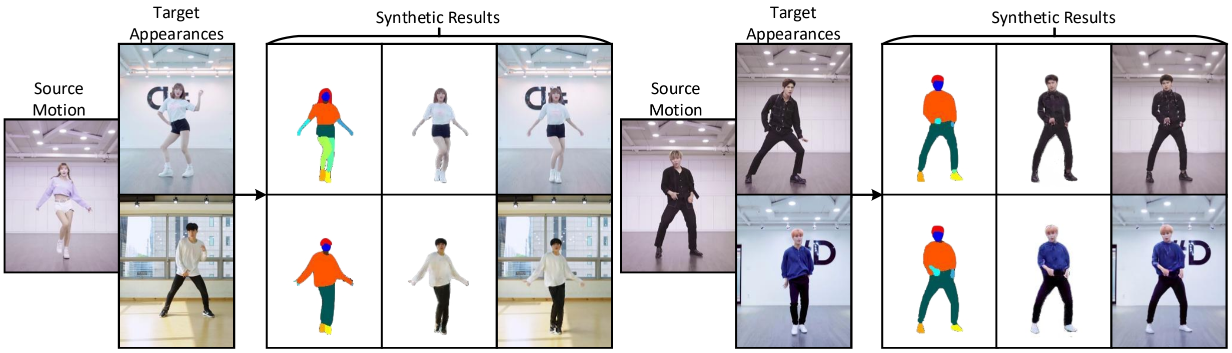

Fig. 6. Examples of ordinary HVMT (please zoom in for a better view). Each synthetic result (layout, foreground, full scene) is generated to have the same motion as its input source motion image and the same appearance as its input target appearance image.

For a full and animated version of our HVMT results, please refer to the video examples in 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) [51] and Video Fréchet Inception Distance (VFID) [7]. In particular, SSIM and PSNR are classic metrics that measure the pixel-level image similarity, which are simple and based on shallow functions. LPIPS is a newly invented metric 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 left portion 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 than those generated by other methods. Besides, we observe that video-based methods which employ temporal consistency designs can obtain better quantitative results than image-based methods in most cases, causing the similarity among results

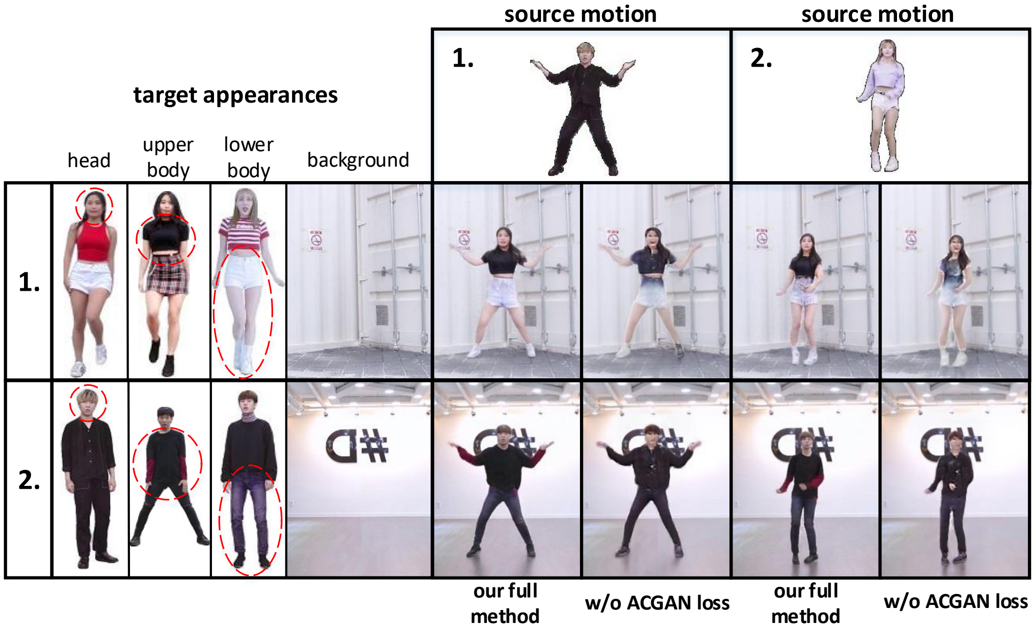

Fig. 7. Examples of multi-source appearance control (please zoom in for a better view). Each synthetic image is generated based on five inputs in terms of body motion, appearances of head, upper body, lower body and background. We also show the results generated by the variant model without the ACGAN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

TABLE I
QUANTITATIVE AND PERCEPTUAL COMPARISON RESULTS ON OUR DATASET AND IPER [11] DATASET. SSIM AND PSNR ARE SIMILARITY METRICS, THE HIGHER THE BETTER. LPIPS AND VFID ARE DISTANCE METRICS, THE LOWER THE BETTER. PREFERENCE IS DENOTED AS THE PREFERENCE RATIO OF VIDEOS GENERATED BY OUR METHOD TO VIDEOS GENERATED BY OTHERS.

Datasets	Metrics	vid2vid [7]	EDN [6]	PoseWarp [12]	LWGAN [11]	w/o input selection	w/o layout GAN	w/o ACGAN loss	w/o shadow extraction module	Ours
Ours	SSIM	0.8834	0.8711	0.8380	0.8538	0.8652	0.8545	0.8613	0.8580	**0.8947**
	PSNR	26.9316	26.5653	23.5423	24.1451	24.7923	22.5398	24.1372	24.0782	**27.9458**
	LPIPS	0.0352	0.0363	0.0537	0.0391	0.0413	0.0436	0.0394	0.0419	**0.0341**
	VFID	3.9752	4.3410	7.0721	4.5528	5.1426	5.3624	4.9845	5.2187	**3.9689**
	Preference (ours/others)	0.692/0.308	0.731/0.269	0.938/0.062	0.769/0.231	0.762/0.238	0.815/0.185	0.777/0.223	0.808/0.192	—
iPER [11]	SSIM	0.8688	0.8522	0.8254	0.8498	0.8631	0.8473	0.8597	0.8511	**0.8724**
	PSNR	26.9289	25.9380	22.9176	24.0757	24.4592	22.6384	23.9643	23.7256	**27.1205**
	LPIPS	0.0416	0.0434	0.0623	0.0457	0.0475	0.0511	0.0496	0.0507	**0.0409**
	VFID	4.5685	4.7561	7.5910	4.9174	5.3457	5.8542	5.4426	5.4139	**4.4481**
	Preference (ours/others)	0.619/0.381	0.652/0.348	0.881/0.119	0.719/0.281	0.757/0.243	0.814/0.186	0.771/0.229	0.781/0.219	—

of video-based methods.

E. Human Perceptual Results

For human perceptual assessment, we conduct a human subjective study by performing preference tests on the Amazon 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 preference results (our dataset: 130 results, iPER dataset: 210 results) for videos generated by different methods, we summarize the average human preferences in the 6th and the 11th rows 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 value in the table represents the preference ratio of videos generated by our full method to videos generated by a counterpart. Since the comparison between our full method and itself is missing, the preference value 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 right portion 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 11th and the 13th columns show that, without the selected optimal appearance inputs and the elaborate ACGAN loss, the variants can't preserve human appearances well and thus obtain blurry faces and bodies, which indicates that both the two components can improve

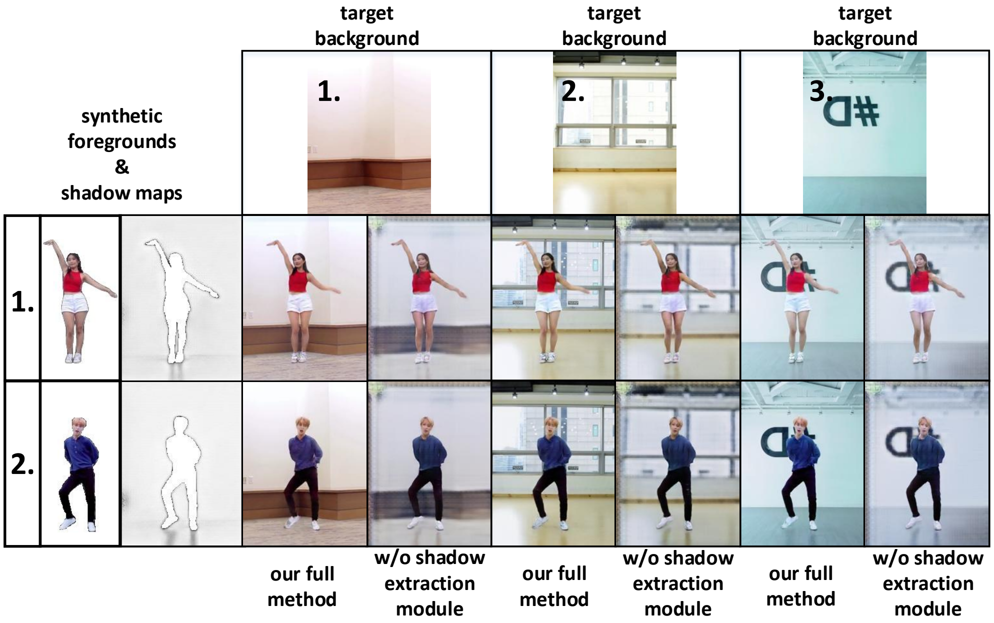
target background
target background
target background
1.
2.
3.
synthetic foregrounds & shadow maps
1.
2.
our full method
w/o shadow extraction module
our full method
w/o shadow extraction module
our full method
w/o shadow extraction module

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 the shadow extraction module, allowing for comparisons with our full method.

human appearance details. The 12th column shows that, without the layout GAN to provide additional motion conditioning inputs that describe human motions more accurately, the single-stage variant 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 the shadow rendering, the variant fails in both background and foreground synthesis, which indicates 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 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. Discussion

Despite the fact that the proposed GAC-GAN can achieve general-purpose and appearance-controllable human video motion transfer with higher video quality than state-of-art methods, we still have the following limitations to be addressed in our future works.

1) Visual artifacts may occur in our generated results. The main reason is that the layout generation highly depends on the conditioning inputs which are obtained based on pose and layout estimation techniques. Therefore, visual artifacts may occur in the generated layouts when these estimations fail, resulting in artifacts in the final video results. We can alleviate this by employing better pose and layout estimation algorithms in the future. Besides, sharp human motion changes such as rapidly turning faces or moving limbs over faces may violate the temporal consistency around the face region learned by our model, which may cause our generated layouts to include jittering around the face region and thus further result in face jittering in our final results. A possible solution is to employ a separate face generator that only takes facial landmarks as its conditioning inputs to ensure the generated face appearances independent of the face layouts which might suffer from the jittering. We can also adopt different loss weights for different layout regions in stage 1, and make the weight of the face region larger than other regions to strengthen the supervision of the face layouts, which may also alleviate the jittering problem.

2) Due to the lack of ground-truth frames for the supervision of multi-source generation, there would be a quality decline for multi-source generation when compared with single-source generation. Although the experimental results show that employing separate ACGAN losses for different body parts can help to eliminate such quality discrepancy, the effect of the proposed ACGAN loss is determined by how well the Appearance-Consistency (AC) discriminators distinguish inconsistent appearances, which is restricted by the relatively insufficient appearance variety of our training data. Therefore, we can include more videos of different human subjects wearing different clothes in our dataset to improve the performance of multi-source generation in the future. Besides, we can split the full appearance generator into multiple smaller body part generators to generate different body parts separately rather than generate them as a whole, which can further alleviate the inner relevance of different body parts during training and thus improve the multi-source generation.

3) Due to the lack of background appearances with lighting conditions that are very different from indoor environments,

our shadow extraction module currently doesn't account for shadow rendering for such backgrounds, which also remains an open challenge. In the future, we can train a separate shadow extraction module on datasets containing various backgrounds to address this problem, where we can further include backgrounds in the conditioning inputs to allow the module to adapt to different background appearances.

4) Due to the nature of deep neural networks, our method may fail when the trained model is tested on unseen domains that are too different from the training domains. For example, if we want to generate videos for Computer-Generated (CG) characters such as cartoon characters, we need to include various cartoon video sequences together with the estimated poses and layouts in our training data, and then train additional models to learn their computer-generated texture and body shape patterns, which look very different from real humans and even lead to the failure of pose and layout estimations.

Furthermore, we may also explore the potential of synthesizing more complex videos where multiple people dance together rather than solo dance videos in the future. 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.

VII. 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 and iPER 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.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[6] 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.

[7] 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.

[8] 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

[9] 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.

[10] K. Aberman, M. Shi, J. Liao, D. Liscbinski, 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.

[11] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao, "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*, 2019, pp. 5904–5913.

[12] 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.

[13] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *null*. IEEE, 2003, p. 726.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[24] 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. 1486–1494.

[25] Y. Guo, Q. Chen, J. Chen, Q. Wu, Q. Shi, and M. Tan, "Auto-embedding generative adversarial networks for high resolution image synthesis," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2726–2737, 2019.

[26] 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.

[27] T. Miyato and M. Koyama, "cGANs with projection discriminator," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=ByS1VpgRZ

[28] 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.

[29] 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.

[30] W. Wang, X. Alameda-Pineda, D. Xu, E. Ricci, and N. Sebe, "Learning how to smile: Expression video generation with conditional adversarial recurrent nets," *IEEE Transactions on Multimedia*, 2020.

[31] 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.

[32] Y. Yan, B. Ni, W. Zhang, J. Xu, and X. Yang, "Structure-constrained motion sequence generation," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1799–1812, 2018.

[33] 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.

[34] 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.

[35] 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.

[36] X. Chen and W. Wang, "Uni-and-bi-directional video prediction via learning object-centric transformation," *IEEE Transactions on Multimedia*, 2019.

[37] 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.

[38] 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.

[39] 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.

[40] D. Liang, R. Wang, X. Tian, and C. Zou, "Pcgan: Partition-controlled human image generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8698–8705.

[41] Y. Liu, W. Chen, L. Liu, and M. S. Lew, "Swapgan: A multistage generative approach for person-to-person fashion style transfer," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2209–2222, 2019.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2794–2802.

[50] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018.

[51] 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.

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 a 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 Technology and Design (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.