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

**Dongxu Wei** and **Haibin Shen** and **Kejie Huang** [1]



**Abstract.** Due to the rapid development of Generative Adversarial Networks (GANs), there has been significant progress in the field of human video motion transfer which has a wide range of applications in computer vision and graphics. However, existing GAN-based works only support motion-controllable video synthesis while appearances of different video components are bound together and uncontrollable, which means one person can only appear with the same clothing and background. Besides, most of these works are person-specific and require to train an individual model for each person, which is inflexible and inefficient. Therefore, we propose appearance composing GAN: a general method enabling control over not only human motions but also video appearances for arbitrary human subjects within only one model. The key idea is to exert layout-level appearance control on different video components and fuse them to compose the desired full video scene. Specifically, we achieve such appearance control by providing our model with optimal appearance conditioning inputs obtained separately for each component, allowing controllable component appearance synthesis for different people by changing the input appearance conditions accordingly. In terms of synthesis, a two-stage GAN framework is proposed to sequentially generate the desired body semantic layouts and component appearances, both are consistent with the input human motions and appearance conditions. Coupled with our ACGAN loss and background modulation block, the proposed appearance composing GAN can achieve general and appearance-controllable human video motion transfer. Moreover, we build a solo dance dataset containing a large number of dance videos for training and evaluation. Experimental results show that, when applied to motion transfer tasks involving a variety of human subjects, our proposed appearance composing GAN achieves appearance-controllable synthesis with higher video quality than state-of-art methods based on only one-time training. Example synthetic video results are available online: https://youtu.be/8fhr5bcFM6Y.

## 1 Introduction

Human video motion transfer 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. With the recent emergence of Generative Adversarial Networks (GANs) [10] and its variant conditional GANs (cGANs) [18], there have been many works [4, 24, 17, 28, 1] for video-based human motion transfer that achieve great success. However, these works have two main weaknesses: 1) For each video, only the mapping between human motions and video frames is learned while the video appearance is learned individually in a person-specific model, which means additional models are required to be trained to generate new appearances. 2) Appearances of different video components (e.g. head, upper body, lower body and background) are learned and generated as a whole, which causes the video appearance is not decomposable or controllable.

In this work, the proposed cGAN-based appearance composing GAN addresses the above two weaknesses from two aspects correspondingly: 1) In addition to motion conditioning inputs used in other works, we provide our model with separate appearance conditioning inputs with respect to head, upper body and lower body, allowing to generate new video appearances by changing the appearance conditions accordingly. Specifically, the appearance condition of each body component is obtained separately during data preprocessing using our elaborate input selection strategy, which is optimally designed based on a pose similarity matching mechanism. By doing this, we can minimize the loss of texture information for appearance synthesis when the input body pose is different from the desired output body pose. 2) Compared to other works that generate the video appearance as a whole, we decompose the full video scene into different video components, which can be generated separately with appearances come from different target videos to achieve multi-source appearance control. In particular, we utilize a layout GAN to firstly synthesize the desired semantic layout, which is better than the sparse body pose points in describing human motion. Then the synthetic layout is taken as dense motion conditioning input by an appearance GAN, which aims at synthesizing the layout-specified foreground and background. By selecting input appearance conditions for different components separately from different target videos, the layout GAN and the appearance GAN can generate the desired multi-source layout and appearance respectively.

To enable the appearance GAN to achieve such multi-source synthesis, the key is to separate different components from each other to alleviate their inner relevance. Therefore, synthesis of each body component foreground is supervised separately during training by a component-specific ACGAN loss, which is carefully designed for not only appearance quality but also appearance consistency between generated components and input appearance conditions. Moreover, different from previous works that directly synthesize the whole background, we embed a light-weight background modulation block to synthesize a modulation map instead, which can effectively modulate the background brightness to render natural shadows. Since the map is generated from motion conditioning inputs and irrelevant to background appearance, it supports controllable background replacement with arbitrary images. Based on the proposed GAN framework, we can control appearances of not only the background but also the body component foregrounds. By fusing them to compose the full video scene, we achieve the appearance-controllable human video

[1] Zhejiang University, China, email: huangkejie@zju.edu.cn

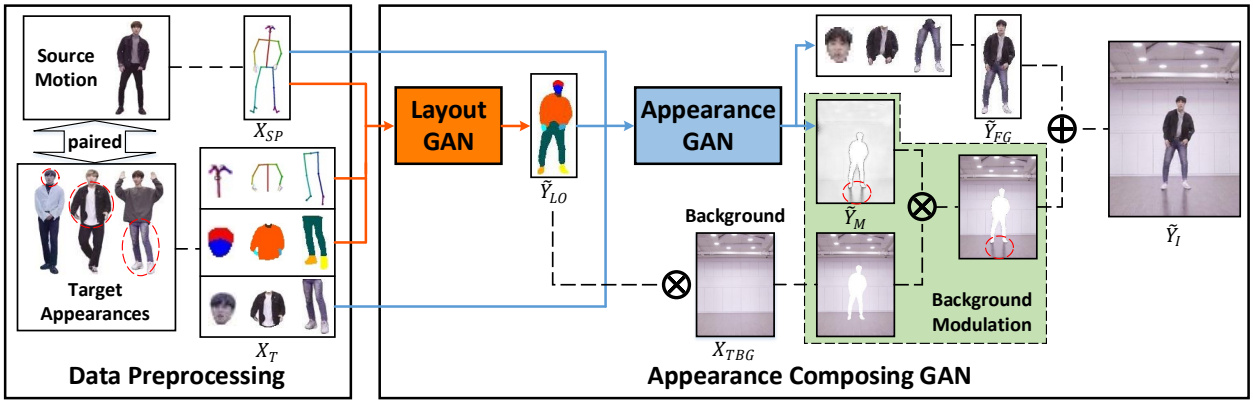


**Figure 1.** Overview of our method. In the data preprocessing, we obtain the paired $X_{SP}$ and $X_T$ from the source motion frame and the target appearance frames. Then the processed inputs are fed into the appearance composing GAN to sequentially generate the layout map $\tilde{Y}_{LO}$ and the scene image $\tilde{Y}_I$, where $\tilde{Y}_I$ is composed of the synthetic foreground $\tilde{Y}_{FG}$ and the modulated background $X_{TBG}$. 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 operation respectively.

motion transfer.

In our experiments, a large solo dance dataset including 148800 frames collected from 124 people is utilized for training and evaluation. Firstly, we compare our approach against state-of-art person-specific [24, 4] and general-purpose [2] methods through qualitative, quantitative and perceptual evaluations on the dataset. 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 tasks of multi-source appearance synthesis and controllable background replacement, where the results show that our method can achieve flexible appearance control on body components as well as surrounding backgrounds. Moreover, to give a better insight into the appearance composing GAN framework, we further conduct comprehensive ablation studies with respect to our input selection strategy, layout GAN, AC GAN loss and background modulation block.

To summarize, our main contributions are as follows:

- We propose appearance-composing 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 effective input selection strategy and novel GAN framework design.
- We construct a large-scale solo dance dataset including a variety of solo dance videos for training and evaluation, which will be released together with our codes to facilitate future research.

## 2 Related Work

Early works have attempted to manipulate existing video frames [8, 20, 21] or animate 3D character models [11, 5, 14] to create human motion transfer videos. However, they result in either low-quality synthetic videos or massive computation budgets dominated by production-quality 3D reconstructions.

Recently, there have been significant efforts which we refer to as image-based methods [2, 19, 26, 22, 6, 15, 23] 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 appearance onto new poses in an image-to-image translation manner [12]. [2, 19, 26] utilize spatial transformation or surface warping 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, [22, 6, 15] apply such transformations to appearance features instead of textures, where the transformed features are then decoded to obtain new pose images. Furthermore, [23] propose a style consistency discriminator to force the generator to preserve the input appearance style, which gives a new sight from the aspect of discriminator design. However, all of these image-based methods are designed for still image synthesis without consideration of temporal consistency, which causes they are not qualified for human motion 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.

As the video counterpart of the above mentioned image-based pose transfer, video-based motion transfer considers temporal coherence with access to more appearance information contained in a whole video, leading to a higher level of temporal consistency and visual quality. In [24], the authors propose to generate optical flows to warp previously generated frames into temporally consistent new frames. Besides, [4] use a temporal smoothing loss to enforce temporal coherence between adjacent frames. Note that video quality depends not only on temporal consistency but also on appearance details. Thus recent works come up with feeding rendered images of 3D models [17] or transformed images of body parts [28] into their models as input conditions to obtain realistic appearances. Moreover, [1] split their network into two training branches responsible for appearance generation and temporal coherence improvement respectively. Although these works can generate videos with higher visual quality than image-based methods, an obvious limitation is that they have to train an individual model for each person with less feasibility of application. Besides, most of them generate the full video scene directly with body components and backgrounds bound together, keeping them from controllable appearance synthesis.

## 3 Method

### 3.1 Overview

Before describing our method, we first define the problem to solve: given inputs of a source video and multiple target videos, we aim at synthesizing a new video with human motion of the source video and combined appearance of the target videos.

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 each motion conditioning input (source pose) is paired with an optimal appearance conditioning input (target pose, layout and foreground). Benefiting from our multi-source input selection strategy, the appearance input of each body component can be selected separately from its own target appearance source. Next, we feed the two conditioning inputs into our appearance composing GAN which consists of a layout GAN and an appearance GAN, responsible for controllable layout synthesis and appearance synthesis respectively. By controlling the motion and the appearance inputs through input video sources, the layout GAN can generate the desired multi-source body layout, which is taken as layout-level motion conditioning input by the appearance GAN to generate the corresponding multi-source appearance. To decompose the synthesis of the whole body into individual body components, the appearance GAN generates foregrounds of different components separately. For background appearance synthesis, a background modulation block is applied in the appearance GAN to generate a modulation map that renders natural shadows instead of generating the whole background. Because the generation has no relation to background appearance, it supports background replacement with arbitrary images. Finally, the synthetic component foregrounds and the modulated background are added together to compose the full video scene.

Before further discussions, we give variable definitions used in this paper as follows, where $X$ means input and $\tilde{Y}$ means output, $S$ and $T$ represent input source and target videos, $P$, $LO$, $FG$, $BG$, $M$, $I$ represent pose, layout, foreground, background, modulation map and full scene image, $H$, $U$, $L$ refer to head, upper body and lower body.

1. **Inputs**
   - **source motion**: source pose $X_{SP}$
   - **target human appearance** ($X_T$):
     target poses ($X_{TP}$): $X_{TP,H}$, $X_{TP,U}$, $X_{TP,L}$
     target layouts ($X_{TLO}$): $X_{TLO,H}$, $X_{TLO,U}$, $X_{TLO,L}$
     target foregrounds ($X_{TFG}$): $X_{TFG,H}$, $X_{TFG,U}$, $X_{TFG,L}$
   - **target background appearance**: $X_{TBG}$
2. **Outputs**
   - **layout GAN**: synthetic body layout $\tilde{Y}_{LO}$
   - **appearance GAN**:
     synthetic foregrounds: $\tilde{Y}_{FG,H}$, $\tilde{Y}_{FG,U}$, $\tilde{Y}_{FG,L}$
     synthetic background modulation map: $\tilde{Y}_M$
     synthetic full scene: $\tilde{Y}_I$

## 3.2 Data Preprocessing

The main purpose of data preprocessing is to select the optimal appearance conditioning input ($X_T$) which contains the most target appearance information needed for appearance synthesis with respect to the desired source motion ($X_{SP}$), where each component has its own appearance inputs selected from the corresponding target video frames to enable multi-source control. The preprocessing consists of the following three steps, aiming at minimizing the loss of appearance information for appearance synthesis in unseen body poses. It's noted that there's no restriction on the input frame number, we can obtain optimal appearance inputs no matter how many frames are provided.

### 3.2.1 *Step 1: Obtaining Poses, Layouts and Foregrounds*

We utilize [3] and [9] to detect body poses and semantic layouts respectively. Thereafter we can extract foregrounds for each body component by multiplying the images with corresponding channels of the one-hot layout maps. Hence we can obtain the source motion inputs $X_{SP}$ from the source video and the target appearance inputs $X_T$ from the target videos, which are paired with each other in step 2.

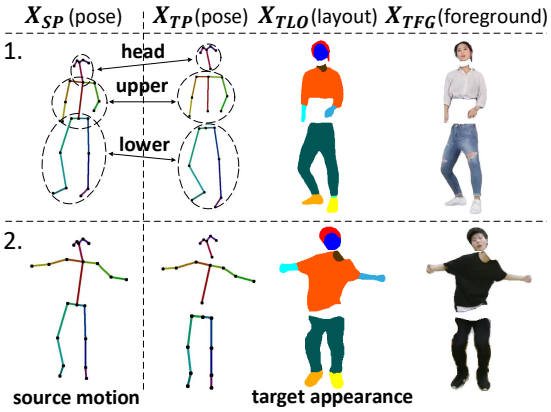


**Figure 2.** Multi-source input selection. We show two examples of pose similarity matching in our multi-source input selection. Each component of each source pose is paired with a target component (pose, layout and foreground) according to component pose similarity.

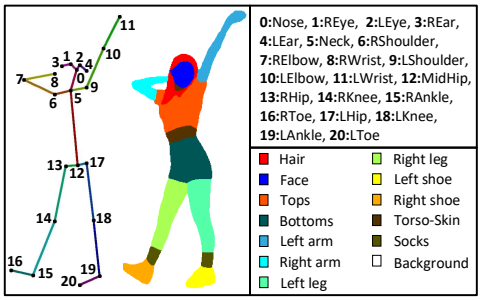


**Figure 3.** Illustration of pose and layout detection results. Pose points and semantic labels are distinguished by numbers and colors respectively.

### 3.2.2 *Step 2: Multi-Source Input Selection*

Since human appearance varies significantly with body pose, we propose a pose similarity matching mechanism depicted in Figure 2 to pair each component of each $X_{SP}$ with the $X_T$ whose corresponding component has the most similar component pose. Provided with different target videos as appearance sources, we can easily obtain multi-source input appearance conditions. Specifically, each component pose similarity is denoted as the average cosine similarity between the corresponding source and target component pose vectors:

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

where $Sim$ is the component pose similarity, $N$ is the number of component pose vectors, $\overrightarrow{V_S}$ and $\overrightarrow{V_T}$ represent component pose vectors of the source pose $X_{SP}$ and the target pose $X_{TP}$ respectively. For vectors of head, upper and lower body, $N = 5, 7$ and $8$ respectively, with vector point locations are shown in Figure 3.

### 3.2.3 *Step 3: Pose Normalization*

Given that $X_{SP}$ and $X_T$ are paired for each frame, we apply a pose normalization to transform each component of $X_T$ into the same size

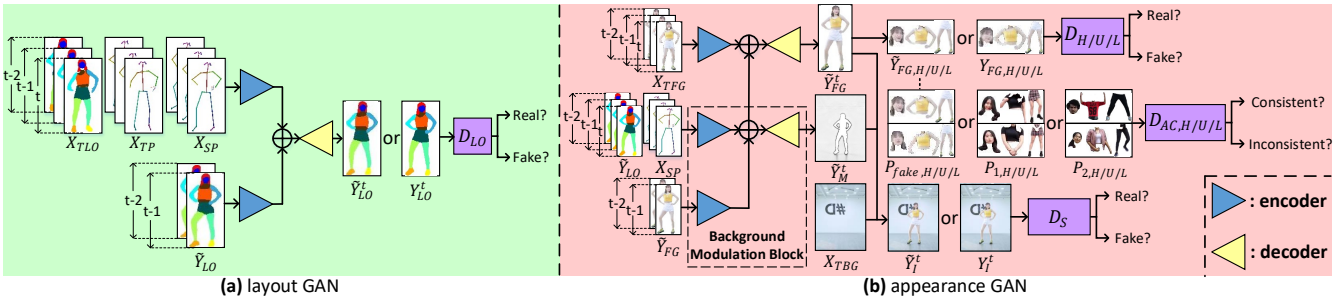


**Figure 4.** Illustration of the appearance composing GAN. (a) and (b) depict frameworks of the layout GAN and the appearance GAN respectively. In (b), foregrounds and discriminators of the three body components are drawn together ($H/U/L$) for simplicity, which are separated in practice.

and position as the corresponding component of $X_{SP}$, where the scale values and the translation distances of different components are computed separately by analyzing the corresponding component pose vector lengths and point locations like this:

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

where $N_v$ is the number of component pose vectors, $N_p$ is the number of component pose points, $P_S$ and $P_T$ represent source and target component pose points respectively.

Thus we obtain the transformed component pose points, semantic layouts and foregrounds. Then the pose points of different components are connected to compose a new target pose $X_{TP}$ while the semantic layouts are processed into a one-hot tensor $X_{TLO}$ with each channel represents a body part as shown in Figure 3. Similarly, the foregrounds are also processed into a tensor $X_{TFG}$ 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 components which come from different target video frames.

## 3.3 Appearance Composing GAN

Our appearance composing GAN has two stages: a layout GAN and an appearance GAN, both of which are trained in a single-source manner due to the lack of ground truth data for supervision when source poses and target appearances come from different persons. Although it results in a difference between single-source training and multi-source testing, we can eliminate this difference through our elaborate model designs described in the following. 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.

### *3.3.1 Layout GAN*

The layout GAN aims at synthesizing the multi-source layout with each component consistent with its appearance input. By taking the generated layout as an additional motion condition, the subsequent appearance GAN can achieve layout-specified appearance synthesis.

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 concatenate consecutive input frames together to improve temporal consistency. In addition, the discriminator $D_{LO}$ is designed to be multi-scale [12] to determine whether the generated layout is real or fake.

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

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

$L_{GAN}^{LO}$ 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 $\tilde{Y}_{LO}$.

$L_{FM}^{LO}$ is the discriminator feature matching loss presented in pix2pixHD [25] and weighted by $\lambda_{FM}$.

$L_T^{LO}$ is the temporal loss used to improve temporal consistency, which is presented in vid2vid [24] and weighted by $\lambda_T$.

$L_{SS}^{LO}$ is the structural sensitive loss adapted from [16] and weighted by $\lambda_{SS}$, which is used to measure the difference between $Y_{LO}$ and $\tilde{Y}_{LO}$ at both the pixel level and the structure level.

### *3.3.2 Appearance GAN*

The function of the appearance GAN embraces three aspects: foreground decomposition, background modulation and scene composition.

- Foreground decomposition refers to separate foreground syntheses for different body components, which is achieved by applying our component-specific ACGAN losses during training. By separating syntheses of different components during training, we can eliminate the difference between single-source training and multi-source testing, which enables straightforward application to multi-source appearance synthesis tasks without adaptation.
- Background modulation refers to pixel-level background brightness control dominated by the generated modulation map, which avoids background generation from scratch and enables controllable background replacement.
- Scene composition refers to fusion of the synthetic component foregrounds and the modulated background. To ensure a harmonious composition, a scene discriminator is designed to coordinate with the appearance generator during training.

As shown in Figure 4(b), the appearance GAN is made of an appearance generator $G_A$, a scene discriminator $D_S$, three standard component discriminators $D_H$, $D_U$, $D_L$ and three appearance-consistency component discriminators $D_{AC,H}$, $D_{AC,U}$, $D_{AC,L}$.

Specifically, $G_A$ consists of three encoders and two decoders: The first encoder learns to encode the target appearance features with $X_A^1|_{t-2}^t = X_{TFG}|_{t-2}^t$ as its input. The second encoder learns to encode the source motion features with $X_A^2|_{t-2}^t = [X_{SP}|_{t-2}^t, \tilde{Y}_{LO}|_{t-2}^t]$ 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 human appearance at time $t$. Meanwhile, features of the second and the third encoders are summed and fed into the second decoder to generate the background modulation map $\tilde{Y}_M^t$, which is output by a sigmoid layer with the same size as the target background $X_{TBG}$. By multiplying $X_{TBG}$ with $\tilde{Y}_M^t$, the background brightness is modulated pixel by pixel to achieve shadow rendering. Then the foreground and the background are added together to compose the full image $\tilde{Y}_I^t$, which is the desired video scene at time $t$.

To achieve a better generation at the component level as well as the scene level, we train our generator with multiple multi-scale discriminators responsible for syntheses of the individual components and the full scene.

As for component level synthesis, we utilize three standard component discriminators $D_H$, $D_U$ and $D_L$ to force the generator $G_A$ synthesize more realistic head, upper and lower body. In practice, we decompose the generated and the real foreground into the three body components and feed them as the fake and the real samples into their corresponding standard discriminators. Inspired by [23], we further apply three appearance-consistency (AC) component discriminators $D_{AC,H}$, $D_{AC,U}$ and $D_{AC,L}$ to ensure appearances of the generated components are consistent with their input appearance conditions. Specifically, we obtain three kinds of component foreground pairs as training samples for each $D_{AC}$ as shown in Figure 4(b): 1) consistent pair $P_1$: two components from the same person, labeled as "true"; 2) inconsistent pair $P_2$: two components from different persons, labeled as "false"; 3) fake pair $P_{fake}$: component of the generated $\tilde{Y}_{FG}$ and its corresponding component appearance condition in $X_{TFG}$, labeled as "false" when updating discriminator and labeled as "true" when updating generator. In company with the progress of $D_{AC}$s which distinguish inconsistent component appearances well, $G_A$ learns to generate more consistent component appearances during adversarial training.

As for scene level synthesis, we utilize a scene discriminator $D_S$ to force the appearance generator focus on details at component boundaries to synthesize the full scene better, where $\tilde{Y}_I$ and $Y_I$ are fed as the fake and the real samples for training.

To train the appearance GAN, the objective is designed like this:

$$\begin{aligned} L_A =& L_{ACGAN}^H + L_{ACGAN}^U + L_{ACGAN}^L + L_{GAN}^S \\ &+ \lambda_{FM} L_{FM}^A + \lambda_{VGG} L_{VGG}^A + \lambda_T L_T^A \end{aligned} \tag{5}$$

$L_{ACGAN}^{H/U/L}$ are ACGAN losses of different body components, each of which is summed by a standard adversarial 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:

$$L_{ACGAN}^H = L_{GAN}^H + \lambda_{AC} L_{AC}^H \tag{6}$$

$$\begin{aligned} L_{GAN}^H =& E[log D_H(Y_{FG,H}, X_{A,H}) \\ &+ log[1 - D_H(\tilde{Y}_{FG,H}, X_{A,H})]] \end{aligned} \tag{7}$$

$$\begin{aligned} L_{AC}^H =& E[log D_{AC,H}(P_{1,H}) \\ &+ log[1 - D_{AC,H}(P_{2,H})] \\ &+ log[1 - D_{AC,H}(P_{fake,H})]] \end{aligned} \tag{8}$$

where $\lambda_{AC}$ is the weight of $L_{AC}$, $Y_{FG,H}$ represents head region of the real foreground, $X_{A,H}$ represents head region of the conditioning input $X_A = [X_A^1, X_A^2]$, $P_{1,H}$, $P_{2,H}$ and $P_{fake,H}$ are consistent, inconsistent and fake head component pairs respectively.
$L_{GAN}^S$ is the GAN loss for the full scene, derived as follows:

$$L_{GAN}^S = E[log D_S(Y_I, X_I) + log[1 - D_S(\tilde{Y}_I, X_I)]] \tag{9}$$

where $Y_I$ is the real scene image, $X_I = [X_A, X_{TBG}]$.
$L_{FM}^A$ is the discriminator feature matching loss weighted by $\lambda_{FM}$, $L_{VGG}^A$ is the VGG loss [13, 7, 25] weighted by $\lambda_{VGG}$, $L_T^A$ is the temporal loss weighted by $\lambda_T$ to improve temporal consistency.

# 4 Experiments

## 4.1 Dataset

We construct a large dataset with 124 dance videos collected from 58 males and 66 females, including a variety of human identities, dance categories and wearing styles. 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.

In detail, we first extract backgrounds automatically for each video by stitching background regions of different frames. Then we cut each video into a sequence with 1200 frames to detect poses, layouts and foregrounds, where we crop and resize all the frames to central 192x256 regions and manually rectify ones with bad detection results for better data quality. Then we divide each processed video sequence into two halves, where the first half is used to extract $X_{SP}$ and the second is used to obtain the paired $X_T$. Therefore, we obtain 600 available conditioning inputs and the corresponding ground truth frames for each of the 124 videos. In practice, we use 100 videos for training and the remaining 24 videos for testing.

## 4.2 Experimental Setup

### *4.2.1 Our Method*

The design of encoders and decoders follows pix2pixHD [25], where the numbers of convolutional filters are decreased to half of the original pix2pixHD to reduce the model size. The layout GAN and the appearance GAN are trained separately with Adam optimizers (learning rate: 0.0002, batch size: 4) on 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.

### *4.2.2 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 [24] and EDN [4], both are person-specific with each model can only generate videos with the same 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 person-specific, we implement a state-of-art image-based method PoseWarp [2] as a general-purpose baseline, which is trained on the same data as ours for comparisons on the generalization ability.
- **w/o input selection**:
  To evaluate the effectiveness of our input selection strategy, we implement a model trained with appearance conditioning inputs selected randomly with no extra computation.

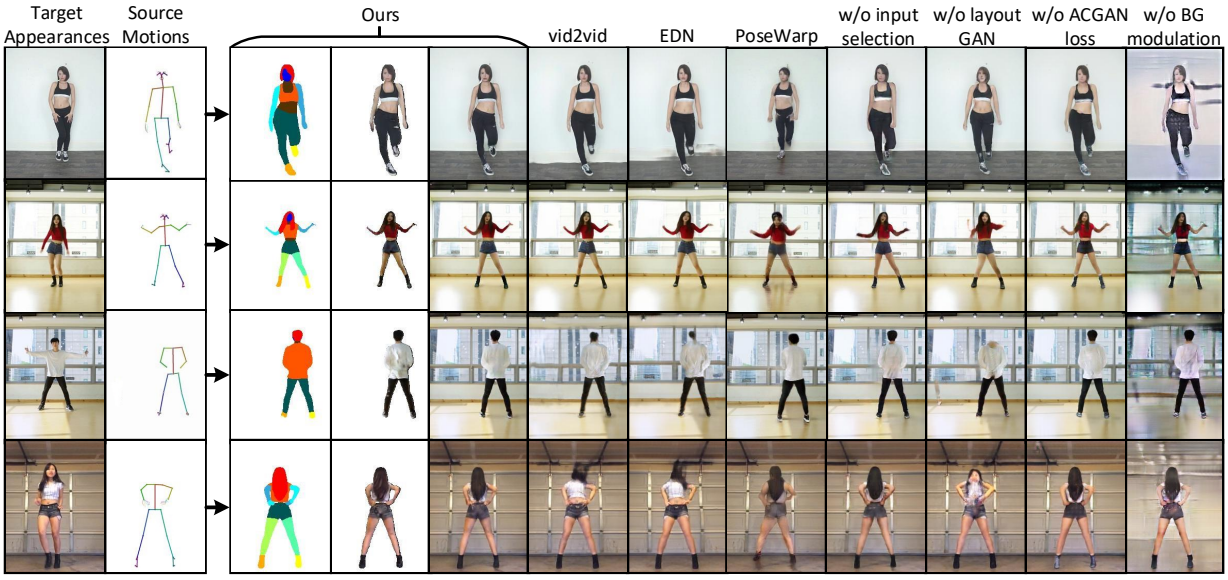


**Figure 5.** Qualitative comparison results on motion transfer tasks (please zoom in for a better view). From left to right: input target appearances, input source motions, our generated results (layouts, foregrounds, full scenes), results of vid2vid [24], results of EDN [4], results of PoseWarp [2], results of the four ablated variants with respect to input selection strategy, layout GAN, ACGAN loss and background modulation block.

- **w/o layout GAN**:
  To evaluate the effectiveness of our layout GAN, we implement a model with only the appearance GAN, which is fed with 2D poses as source motion conditions without synthetic body layouts.
- **w/o ACGAN loss**:
  To evaluate the effectiveness of our ACGAN loss that separates syntheses of different components, we implement a model whose appearance GAN is trained without AC discriminators.
- **w/o background modulation**:
  To evaluate the effectiveness of our background modulation block, we implement a model that generates backgrounds from scratch with fixed background images included in its input conditions.

## 4.3 Qualitative Results

To assess the quality of our synthetic results, we randomly visualize some synthetic frames and compare them with those generated by other methods as shown in Figure 5. Based on the proposed appearance composing 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 [2] can't preserve the target appearances well with body poses and locations are not consistent with the desired source poses. Although the two video-based methods vid2vid [24] and EDN [4] perform well when synthesizing appearances in frequent poses (e.g. front bodies in the first two rows of Figure 5), they render bad visual results when synthesizing appearances in 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 in their training data, which contains only one video sequence for each person-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 most texture information needed for appearance synthesis. Besides, our model is trained with access to more infrequent poses contained in the whole multi-person dataset, leading to better results than EDN and vid2vid when synthesizing unseen infrequent pose appearances.

We also validate our method on appearance-controllable motion transfer tasks. For multi-source appearance synthesis, we let our model synthesize videos with component appearances come from different people, where we obtain realistic results as shown in Figure 6(a). For controllable background replacement, we modulate different backgrounds and fuse them with the synthetic foregrounds to compose full scene images. As shown in Figure 6(b), the full scenes are naturally composed and rendered with detailed shadows in harmony with the body poses.

For a better view of our synthetic visual results, please refer to the online video examples.

## 4.4 Quantitative Results

We also make a quantitative assessment to evaluate differences between synthetic and ground truth video frames by three metrics: Structural Similarity (SSIM), Learned Perceptual Image Patch Similarity (LPIPS) [27] and Video Fréchet Inception Distance (VFID) [24], where SSIM and LPIPS are used to measure single frames while VFID that considers temporal consistency is used to measure video sequences. Results summarized in the first three columns of Table 1 show that our method outperforms others for all the metrics.

## 4.5 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 10 answers for 13 videos generated by different methods, we summarize the average human preference scores in the last column of Table 1. The results indicate that videos generated by our method are also perceptually preferred to those generated by others.

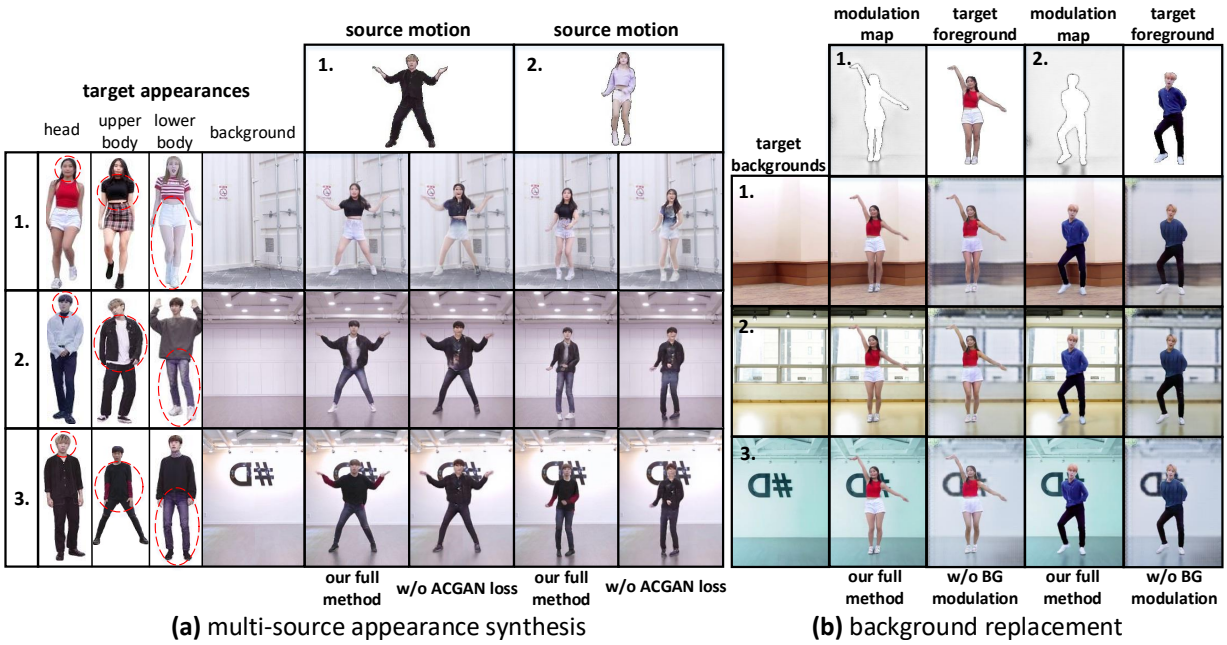


**Figure 6.** Examples of appearance controllable motion transfer tasks (please zoom in for a better view). (a) shows the results of multi-source appearance synthesis, each image is synthesized based on five inputs in terms of pose, head, upper body, lower body and background. (b) shows the results of controllable background replacement, backgrounds are modulated by the synthetic modulation maps to fuse with the foregrounds. Both (a) and (b) contain the results generated by the corresponding ablated variants, allowing for comparisons with our full method.

**Table 1.** Quantitative and perceptual comparison results in motion transfer tasks. SSIM is a similarity metric, 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.

| | SSIM | LPIPS | VFID | Preference Score |
|---|---|---|---|---|
| vid2vid [24] | 0.8834 | 0.0352 | 3.9752 | 69.2% |
| EDN [4] | 0.8711 | 0.0363 | 4.3410 | 73.1% |
| PoseWarp [2] | 0.8380 | 0.0537 | 7.0721 | 93.8% |
| w/o input selection | 0.8652 | 0.0413 | 5.1426 | 76.2% |
| w/o layout GAN | 0.8545 | 0.0436 | 5.3624 | 81.5% |
| w/o ACGAN loss | 0.8613 | 0.0394 | 4.9845 | 77.7% |
| w/o BG modulation | 0.8580 | 0.0419 | 5.2187 | 80.8% |
| Ours | **0.8947** | **0.0341** | **3.9689** | — |

## 4.6 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 background modulation block. The quantitative and the perceptual results are shown in the 4-7th rows of Table 1, where our full method outperforms all the variants significantly. We also make comparisons on qualitative results as shown in the last four columns of Figure 5. The 9th and the 11th 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. The 10th column indicates that, without the layout GAN to achieve layout-level appearance control, the model even can't generate the desired body poses, let alone satisfactory appearances. The last column shows that, without background modulation, the model fails in both background and foreground synthesis. Moreover, we make further qualitative comparisons to demonstrate the importance of our ACGAN loss and background modulation block in multi-source appearance synthesis and background replacement respectively. As shown in Figure 6(a), without training with the proposed ACGAN loss, the model renders bad component appearances which are mixed up together and inconsistent with the input appearance conditions. As shown in Figure 6(b), without the synthetic background modulation maps, the model renders blurry backgrounds as well as unrealistic foregrounds.

## 5 Conclusion

In this paper, we present appearance composing GAN for general-purpose appearance-controllable human motion transfer. In order to synthesize controllable appearances for arbitrary people, we propose a multi-source input selection strategy to first obtain controllable input appearance conditions. Moreover, to enable compatible appearance synthesis given such multi-source inputs, we propose a two-stage GAN framework to separate syntheses of different components at a layout level, where we further employ our ACGAN loss and background modulation block for appearance enhancement. Extensive experiments on our large-scale solo dance dataset show that our proposed method can not only enable general-purpose appearance control but also achieve higher video quality than state-of-art methods. In the future, we may 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] Kfir Aberman, Mingyi Shi, Jing Liao, D Liscbinski, Baoquan Chen, and Daniel Cohen-Or. Deep video-based performance cloning. In *Computer Graphics Forum*, volume 38, pages 219–233. Wiley Online Library, 2019.

[2] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8340–8348, 2018.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[4] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5933–5942, 2019.

[5] German KM Cheung, Simon Baker, Jessica Hodgins, and Takeo Kanade. Markerless human motion transfer. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, pages 373–378. IEEE, 2004.

[6] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. In *Advances in Neural Information Processing Systems*, pages 474–484, 2018.

[7] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in neural information processing systems*, pages 658–666, 2016.

[8] Alexei A Efros, Alexander C Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *null*, page 726. IEEE, 2003.

[9] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 770–785, 2018.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] Chris Hecker, Bernd Raabe, Ryan W Enslow, John DeWeese, Jordan Maynard, and Kees van Prooijen. Real-time motion retargeting to highly varied user-created morphologies. In *ACM Transactions on Graphics (TOG)*, volume 27, page 27. ACM, 2008.

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

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

[14] Jehee Lee and Sung Yong Shin. A hierarchical approach to interactive motion editing for human-like figures. In *Siggraph*, volume 99, pages 39–48, 1999.

[15] Dong Liang, Rui Wang, Xiaowei Tian, and Cong Zou. Pcgan: Partition-controlled human image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8698–8705, 2019.

[16] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):871–885, 2018.

[17] Lingjie Liu, Weipeng Xu, Michael Zollhöfer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. Neural rendering and reenactment of human actor videos. *ACM Trans. Graph.*, 38(5):139:1–139:14, October 2019.

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

[19] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 123–138, 2018.

[20] Arno Schödl and Irfan A Essa. Controlled animation of video sprites. In *Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 121–127. ACM, 2002.

[21] Arno Schödl, Richard Szeliski, David H Salesin, and Irfan Essa. Video textures. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 489–498. ACM Press/Addison-Wesley Publishing Co., 2000.

[22] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3416, 2018.

[23] Miao Wang, Guo-Ye Yang, Ruilong Li, Run-Ze Liang, Song-Hai Zhang, Peter M Hall, and Shi-Min Hu. Example-guided style-consistent image synthesis from semantic labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1495–1504, 2019.

[24] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1152–1164. Curran Associates Inc., 2018.

[25] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[26] Mihai Zanfir, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human appearance transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5391–5399, 2018.

[27] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[28] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Dance dance generation: Motion transfer for internet videos. *arXiv preprint arXiv:1904.00129*, 2019.