IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 4, 2021, accepted January 14, 2021, date of publication January 22, 2021, date of current version February 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3053617

Supervised Video-to-Video Synthesis for Single Human Pose Transfer

HONGYU WANG[1,2], (Graduate Student Member, IEEE), MENGXING HUANG[1,3], (Member, IEEE), DI WU[1,2], YUCHUN LI[1,3], AND WEICHAO ZHANG[1,3]

[1]State Key Laboratory of Marine Resource Utilization in South China Sea, Hainan University, Haikou 570228, China
[2]College of Computer and Cyber Security, Hainan University, Haikou 570228, China
[3]College of Information and Communication Engineering, Hainan University, Haikou 570228, China

Corresponding authors: Mengxing Huang (huangmx09@163.com) and Di Wu (hainuwudi@163.com)

This work was supported in part by the Key Research and Development Program of Hainan Province under Grant ZDYF2019020, in part by the National Natural Science Foundation of China under Grant 61462022 and Grant 71161007, in part by the National Key Research and Development Program of China under Grant 2018YFB140235, and in part by the Education Department of Hainan Province under Grant Hnky201922.

ABSTRACT In this paper, we focus on human pose transfer in different videos, i.e., transferring the dance pose of a person in given video to a target person in the other video. Our methods can be summed up in three stages to tackle this challenging scenario. Firstly, we extract the frames and pose masks from the source video and target video. Secondly, we use our model to synthesize the frames of target person with the given dance pose. Thirdly, we refine the generated frames to improve the quality of outputs. Our model is built on three stages: 1) human pose extraction and normalization. 2) a GAN based on cross-domain correspondence mechanism to synthesize dance-guided person image in target video by consecutive frames and pose stick images. 3) coarse-to-fine generation strategy which includes two GANs: a GAN used to reconstruct human face in target video, the other generates smoothing frame sequences. Finally, we compress the sequential frames generated from our model into video format. Compared with previous works, our model manifests better person appearance consistency and time coherence in video-to-video synthesis for human motion transfer, which makes the generated video look more realistic. The qualitative and quantitative comparisons represent our approach performs significant improvements over the state-of-the-art methods. Experiments on synthetic frames and ground truth validate the effectiveness of the proposed method.

INDEX TERMS Generative adversarial network (GAN), image-to-image translation, video-to-video synthesis, pose-guided person image generation.

I. INTRODUCTION

With the development of various generative adversarial networks (GANs), variational autoencoder (VAE), and conditional GANs (CGANs), pose-guided person image generation has been widely studied recently. However, human pose transfer in videos is still an important and challenging research domain in computer vision. Being able to synthesize novel realistic videos of a person in a melodious dance from the online music videos will not only enrich our lifestyle but also have a great application prospect in e-commerce business, short video production, virtual clothes try-on and automatic fashion design, etc. Due to the popularization of short videos on the Internet, people have more and more opportunities to become Internet celebrities. Nevertheless, making a music video or short video is usually a time-consuming and complex work, which always requires the high cost and a professional team. Generally, it is a meaningful task to create videos by artificial intelligence.

Given a music video of a singer or dancer, we can visualize how to transfer the dance in the video to ourselves. In our daily life, if someone wants to become specialized in dance, then he or she needs to spend a lot of time to study and imitate the movements of dance instructors. It takes years to improve a trainee from amateur to professional level. How about to only use a deep generative model to synthesize a personalized dancing video without day-by-day practice? It can show your outstanding performance in dance or transfer the dance steps from your favorite stars to your body. Turning you from a green hand into a dance master immediately.

The associate editor coordinating the review of this manuscript and approving it for publication was Claudio Cusano.

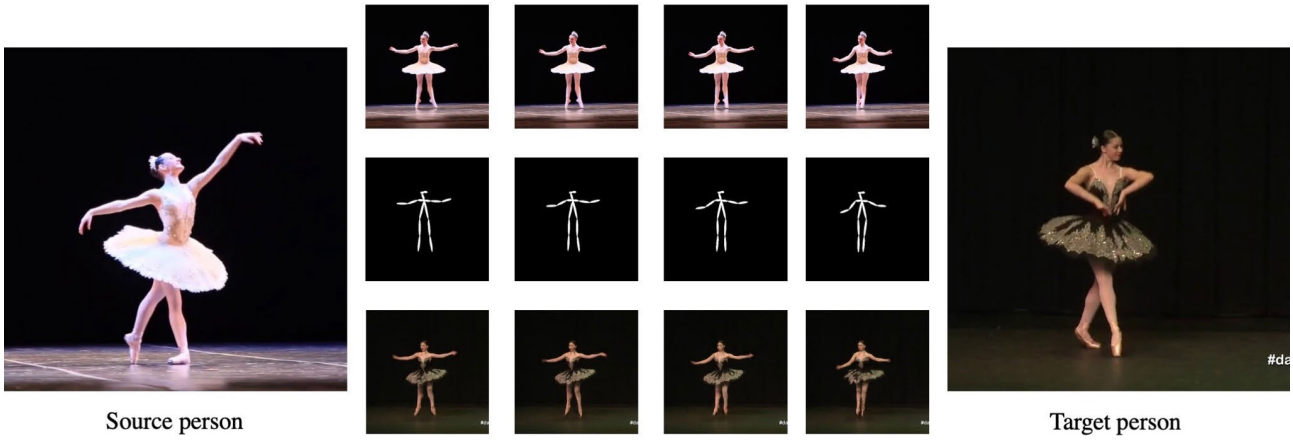

FIGURE 1. Given a dance video (the central top row) of a source person (left) and body pose frames (the central second row), our model can generate a new video (the central bottom row) of the target person (right) with the frames of source pose. The results show that our proposed method can not only produce frames with visual human appearance but also retain the details of target video, such as texture, style, color, clothes, and background.

For the purpose of accomplishing this challenging work (see Fig. 1), we propose a method to tackle with human motion transfer in different videos. Our work is inspired by video-to-video synthesis research [1]–[5], similarly to their methods, we use consecutive human motion of target video to guide training our model. In our method, the pose stick figures are viewed as an intermediate representation during transferring human pose between two given videos. Because the key-point based pose can retain the body positions of different persons rather than their appearance, in addition, the sequential frames of pose label map can preserve the movement trajectory of the human beings superiorly. Open Pose skeleton estimator [6] is utilized for extracting keypoints of human pose and represent them as multi-channel label maps in order to adapt to our model. Then the continuous frames of pose stick feature map and the target person images in corresponding video are fed into the generator to produce realistic images. The well-trained generator is used to synthesize the images of target person under the source pose. Finally, we convert the sequence of generated images into a video.

Differentially from previous human motion transfer methods which almost use CycleGAN [7] or pix2pixHD [8] as global conditional generative adversarial networks (conditional GANs) [9], we present a new human pose transfer (HPT) [10] general framework based on cross-domain correspondence network (CoCosNet) [11] for pose-guided person image generation. Specifically, the generator comprises two sub-networks and learns cross-domain correspondence and image translation simultaneously by end-to-end learning. After training, the coherent pose label maps of the source person are input into the generator to synthesize the person images with target pose in a frame-by-frame manner. To solve the blurred person face images, we design a face GAN to enhance the image quality and facial reality. Compared with several popular methods, the quantitative results exhibit the superior performance of our model. Furthermore, our method also achieves compelling results in ablation studies. In summary, our contributions can be depicted as follows:

- We propose a novel three-stage framework to address the task of human motion transfer on Internet videos.
- An end-to-end video generation network learns to translate continuous frames with the cross-domain correspondence, which outperforms other methods.
- We collect two datasets: a high-quality single-person video dataset which we use mobile phones to record ourselves, and a series of short videos we download from YouTube including various dance types.

II. RELATED WORK

A. IMAGE-TO-IMAGE TRANSLATION

Conditional image-to-image translation [12] aims to learn a mapping function that transforms an input image to another image with a target domain. Variations of Conditional Generative Adversarial Networks (CGANs) [9] have become widely-used models for image translation between different domains owing to their remarkable effectiveness. Pix2pix [13] introduces an encoder-decoder with skip connections following U-Net [14] architecture, and its modified version pix2pixHD [8] uses a coarse-to-fine generator to synthesize high-resolution images, which composes of a global generator network and a local enhancer network. Different from these methods, CycleGAN [7] requires unpaired images to learn domain transfer with a cycle consistency loss, analogously, DiscoGAN and DualGAN [15], [16] also use reconstruction consistency to study cross-domain mapping with unsupervised learning. InstaGAN [17] extends the previously proposed CycleGAN model by taking into account per-instance segmentation masks. Specially, it introduces a context preserving loss to learn identity function. CoCosNet [11] presents an end-to-end framework for exemplar-based image translation, and learns the dense semantic correspondence for cross-domain images by weakly supervised learning. A novel generative model named

swapping autoencoder [82] shows excellent performance in image manipulation task, which encodes each image into two disentangled components and map the swapped features via an unsupervised manner.

B. VIDEO-TO-VIDEO SYNTHESIS

The aim of video-to-video synthesis (vid2vid) [1], [3] is to convert an input semantic video to an output convincing video. Generally speaking, video restoration [18]–[23], including super-resolution [24]–[31], deblurring [32]–[37], dehazing [38]–[44], blending [45], [46] and future video prediction [47]–[53] can be considered as different research directions of the video-to-video synthesis issues. A routine approach is to represent source video as consecutive frames in order, and then generate target video from the model-processed images according to the time sequence. Few-shot vid2vid [3] takes a semantic video and some images of target domain for video generation, which enhances the domain generalization capability by using attention mechanisms.

Compared with images, videos have one more dimension: temporal information, so videos can be understood as a sequence of images in time orientation. Analogously, video style transfer [54]–[57] can be regarded as image-to-image translation at the time series.

C. POSE-GUIDED PERSON IMAGE GENERATION

In early research, PG2 [58] proposed a two-stage model to generate person images and refine the results via a coarse-to-fine module. Ma *et al.* [59] further improved their previous work by disentangling and encoding the image factors, such as foreground, background and pose. For the geometric variability and spatial displacement challenges, Soft-Gated Warping-GAN [60] demonstrates the excellent performance. Instead of Open Pose [6], Natalia *et al.* develop a model based on Dense Pose [61] for mapping pixels from input images and pose labels to a common surface-based coordinate system. Similarly, Li *et al.* [62] proposed to estimate dense and intrinsic 3D appearance flow to guide person image generation. Zhu *et al.* [63] introduce cascaded Pose-Attentional Transfer Blocks into generator to transform the source data. On this basis, Men *et al.* [64] put forward a new network architecture with style block connections and a human parser to separate the attributes and encode them respectively. In order to deal with person image spatial transformation problems, Ren *et al.* [65] combine flow-based operations with attention mechanisms and the model consists of a Global Flow Field Estimator and a Local Neural Texture Renderer. Furthermore, [66], [67] also use an unsupervised manner to tackle this task via end-to-end training. Different from these methods, 3D body mesh recovery component is proposed to disentangle the pose and shape, moreover, an innovative Liquid Warping Block (LWB) [68] is applied to preserve the diverse texture, style, color, facial details, cloth fabrics and other source information.

III. MATHOD OVERVIEW

Given two different videos, one is the reference video with a source person, and the other is the training video that needs to transfer the pose of a target person, we aim at generating a new video of the target person with motions as same as the source person. We consider the human motion transfer task as follows. Representing the input videos as a sequence of frames, $\{F_s \in \mathbb{R}^{3\times W\times H}\}$ is the frames of the source person, while $\{F_t \in \mathbb{R}^{3\times W\times H}\}$ is the frames of the target person. The images are center-cropped to the resolution of 512×512 and resized to 256×256. Then we use human pose estimator (HPE) to extract the 18-channel heat map that encodes the locations of 18 joints of a human body, i.e., the corresponding keypoint-based pose $P_s \in \mathbb{R}^{K\times W\times H}$ and $P_t \in \mathbb{R}^{K\times W\times H}(K = 18)$. After that, the images of the target pose P_t and the frames of target person F_t are fed into the model to train the generator. After end-to-end training, the source pose P_s and the frames of target person F_t are input into the network to synthesize images with the target person appearance but under the pose P_s. Finally, we splice the successive images into a video. In the following, we will describe each module of our model in detail.

A. POSE LABEL GENERATION

We use a pre-trained pose detector $\mathcal{P}$ [6], [69] to extract the 18 human keypoints which are encoded into a 18-channel binary heatmap. Due to open pose detector only extracts the keypoints coordinates of human body, we should visualize these keypoints and link the joints so that we can obtain the pose skeleton labels for the human pose transfer task. This open-source method has been widely used in [58]–[60], [63], [64]. In order to create pose label images that encode human body position, we fill each channel with pixels around the radius of corresponding keypoint and draw the lines between nearby coordinates. Then the generated pose skeleton labels are used as semantic images to guide generative adversarial network to output the person images under the target pose.

B. GLOBAL POSE NORMALIZATION

We have observed that the body proportions of persons are not always the same in different videos. This is caused by the distance between camera lens and captured person is always different in each video. Some cases maybe occur during human motion transfer, for example, the position of source person is visibly upper or lower than the target person in a frame. If we directly input the pose label images to our model, the results are not always satisfactory. Therefore, it is essential to find an appropriate transformation in terms of translation between the source and target pose.

For the sake of translating the source pose reasonably, we need to caculate a suitable translation value. The translation factor is computed according to the maximum and minimum head positions in the condition frames. We describe the pose normalization as the translation of pose coordinates in the y direction. When given a frame where the source

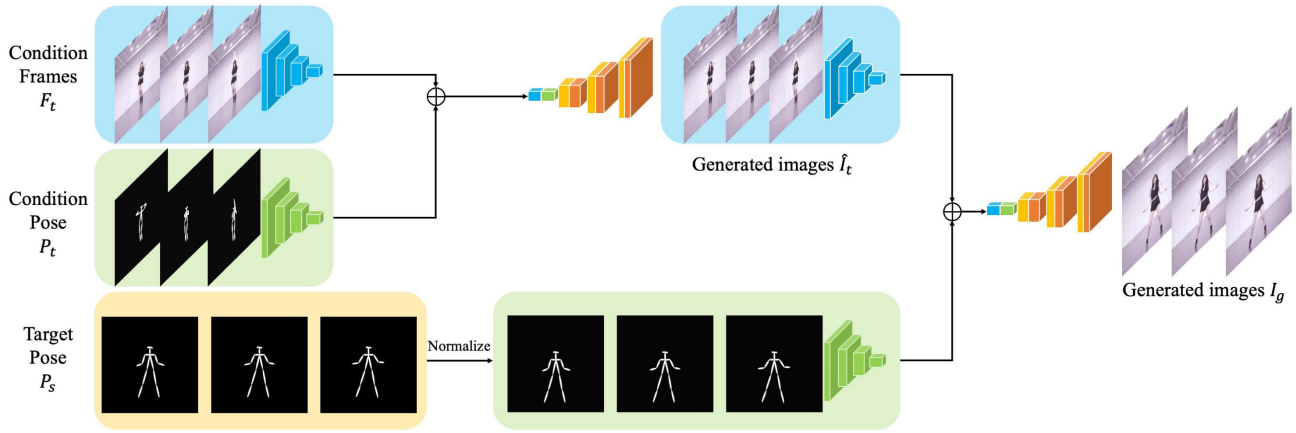

FIGURE 2. The overview of our framework architecture. With the input frames F_t, its pose frames P_t extracted by the pose detector, and the pose frames of the source person P_s, the goal of our model is to generate new consecutive frames in pose P_s. The pipeline of the generator contains two stages. In the first stage, the network is trained to output the frames under the source pose P_t, then the normalized target pose P_s and the generated frames are fed into the well-trained generator to transform the human pose during the second stage.

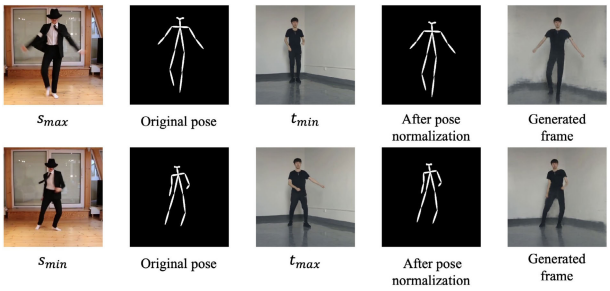

FIGURE 3. The pose normalization component is used to adjust the position of pose keypoints so that the translated pose coordinates appear in appropriate position.

person is at a random location, then the translation $\mathcal{N}$ for the source pose is determined by:

$$\mathcal{N} = t_{\min} + \frac{L - s_{\min}}{s_{\max} - s_{\min}} (t_{\max} - t_{\min}) \tag{1}$$

where $s_{\min}$ and $s_{\max}$ are the source person's farthest and closest positions in one video, $t_{\min}$ and $t_{\max}$ are the target person's farthest and closest positions in the other video. These symbols are all the y coordinates of person's head and represent the farthest and closest distances from the person to the camera respectively. From the visual perspective, you can feel that when the person's head is at the farthest position, the person in the video is always far away from you, and the person is small in the frame, while the head is at the closest position, the person looks large in the frame and close to you. Location L is the average coordinates in y direction of source person's head extracted by pose detector (Open pose represents the 18 keypoints of human body as paired coordinates in both x and y directions). We define the head coordinates as the position of source person in the images.

Different from [2] uses the average of the left and right ankle coordinates to determine the position of person, we select head coordinates to translate the source person's pose vertically. As you can see the 1st row in Fig. 3, when the original pose skeleton is higher in frame, the pose normalization module will translate the pose coordinates downwards in vertical direction, similarly in the 2nd row, if the original pose skeleton is lower in frame, it will translate the pose coordinates upwards in vertical direction. The goal of pose normalization is to find a transformation factor in terms of translation between a source pose and a target pose.

C. HUMAN POSE TRANSFER NETWORK ARCHITECTURE

1) CROSS-DOMAIN CORRESPONDENCE NETWORK

Our goal is to learn the translation from the pose domain P to the image domain I. To be specific, given a pose image $x_P \in P_t$ and its person image $y_I \in I_t$, the generated person image $z_I \in I_g$ should conform to the appearance as I_t but under the pose P_s. As shown in Fig. 2, we adopt an indirect two-stage method to accomplish human pose transfer task due to directly input the source pose figures and target person images into the model will cause the background information lost. In the first stage, the frames of target person F_t and its extracted pose figures P_t are fed into the generative adversarial network (GAN) to learn the mapping between the pose figures and person foreground. For the background texture, we use the spatially-adaptive denormalization (SPADE) block [83] to restore it by projecting the spatially structural style information to different activation locations [11]. In the second stage, we input the pose figures from source person P_s and the frames of target person F_t into the well-trained model to generate the person images I_g under the source person's pose P_s.

In order to match the features between the pose labels and person images, we add a domain alignment of pose and appearance in our model. We convert the input images which from two different domains to a shared domain S where the both input semantics can be represented. Two domain adaptors without weight sharing are used to regulate the person images and the pose skeletons to a shared domain S. The domain adaptors comprise several Conv-InstanceNorm-LeakReLU blocks and the spatial size of features in S is 64×64 [11]. Concretely, the condition pose labels P_t and

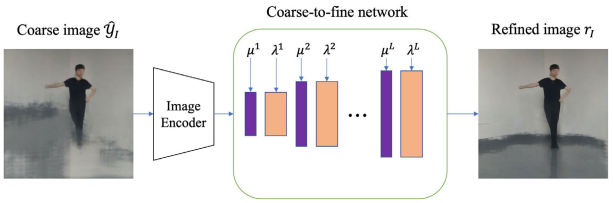

FIGURE 4. The illustration of the coarse-to-fine network. The background of coarse image is warped randomly since the pose figure has no information on it. We employ modulation parameters to control the projection of the style information to different activation locations and the refined image contains the background texture.

the target person frames F_t are sent to the feature pyramid network (FPN) to extract the features and transform them to the representations $x_S \in \mathbb{R}^{WH\times C}$ and $y_S \in \mathbb{R}^{WH\times C}$ in S (W, H are the spatial size, C is the channel-wise dimension). The transformation can be defined as:

$$x_S = \mathcal{T}_{P\to S}\left(x_P; \theta_{\mathcal{T},P\to S}\right), \tag{2}$$

$$y_S = \mathcal{T}_{I\to S}\left(y_I; \theta_{\mathcal{T},I\to S}\right). \tag{3}$$

where θ indicates the learnable parameter.

In order to match and correspond the features within shared domain S better, we denote a correlation matrix $\mathcal{M}_{corr} \in \mathbb{R}^{WH\times WH}$ of which each element is a pairwise feature correlation.

$$\underset{u,v}{\mathcal{M}_{corr}} = \frac{[x_S(u) - \bar{x}_S(u)]^T\,[y_S(v) - \bar{y}_S(v)]}{\|x_S(u) - \bar{x}_S(u)\|\,\|y_S(v) - \bar{y}_S(v)\|} \tag{4}$$

where $x_S(u) - \bar{x}_S(u)$ and $y_S(v) - \bar{y}_S(v) \in \mathbb{R}^C$ represent the channel-wise centralized features of x_S and y_S in position u and v. $\mathcal{M}_{corr}$ manifests a higher semantic similarity between x_S and y_S.

While learning the match by indirect supervised training, we observe that when the network focuses on the correct corresponding regions, the generator performs well. For this reason, we match these intermediate representations from the pose images and person images according to $\mathcal{M}_{corr}$ and obtain the coarse image $\hat{y}_I \in \mathbb{R}^{WH\times 3}$. As in [11], the blurred image $\hat{y}_I$ is acquired by selecting the most correlative pixels in y_I and computing their weighted average.

$$\hat{y}_I(u) = \sum_v \underset{v}{softmax}(\sigma\,\mathcal{M}_{corr}(u,v)) \cdot y_I(v) \tag{5}$$

where σ is the coefficient that adjusts the sharpness of the softmax, $\hat{y}_I$ is the coarse person image without pose transfer, and *softmax* function is used as a filter to output the suitable semantic similarity weight between pose image and person image calculated by the correlation matrix $\mathcal{M}_{corr}$.

2) COARSE-TO-FINE NETWORK

The spatially-adaptive denormalization (SPADE) is used to conserve the structural information in image synthesis. In the SPADE residual block [83], the image is projected onto an embedding space and then convolved to produce the modulation parameters λ^i and μ^i (see in Fig. 4), and they characterize the style and texture information of the original image y_I. The coarse-to-fine network has L layers with the style information progressively injected [11]. For refining the details of output,

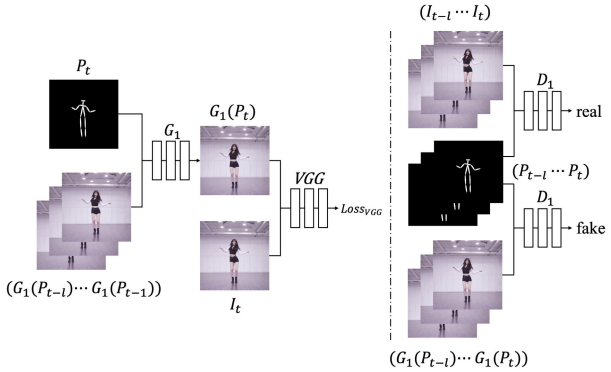

FIGURE 5. The architecture of temporally coherent sequence network. The purpose of this module is to distinguish true temporally interframe sequence from the incoherent sequence.

such as texture, style and color etc., we combine the positional normalization (PN) and spatially-variant denormalization. Γ represents the projection of original image y_I and ξ_Γ is the mapping parameter, the denormalization parameter λ^i and μ^i can be depicted as:

$$\lambda^i, \mu^i = \Gamma_i\left(y_I; \xi_\Gamma\right) \tag{6}$$

We use two convolutional layers to actualize the mapping Γ, λ^i and μ^i have the same spatial size as y_I. With the texture modulation for each normalization layer [11], the refined image r_I can be denoted as:

$$r_I = \mathcal{G}\left(\Phi, \Gamma_i\left(y_I; \xi_\Gamma\right); \theta_\mathcal{G}\right) \tag{7}$$

where $r_I \in \hat{I}_t$ is the refined person image before pose transfer, Φ denotes the vector of coarse image $\hat{y}_I$, and $\theta_\mathcal{G}$ is the learnable parameter.

For transferring the source person's pose to the target person, the pose figures P_s and refined images of target person $\hat{I}_t$ are fed into the well-trained generator to produce the images of target person I_g (see in Fig. 2) under the source person's pose P_s. In our method, video synthesis is based on the consecutive frame sequences. In order to make the synthesized video more realistic and natural, we add two components to improve the quality of generated images.

3) COHERENT INTERFRAME CONSISTENCY

We notice that in the process of video-to-video synthesis, it is necessary to set an appropriate image interval to generate coherent video sequences. If the time interval is too long, the actions of the person in the video are incoordinate, while too short time interval will produce information redundancy, which increases the training time. Inspired by [2], [70], we consider that the video frames are produced sequentially, and use pair $(P_t, G_1(P_{t-l})\cdots G_1(P_{t-1}))$ to train the generator G_1 to output $G_1(P_t)$. As shown in Fig. 5, our network predicts l consecutive frames where the first generated sequence $G_1(P_{t-l})$ is conditioned on its paired pose label image P_{t-l} and a placeholder image z (the first frame in the video, which is not produced). The last generated sequence $G_1(P_t)$ is dependent on its corresponding pose label image P_t and the previous $l-1$ frames output $(G_1(P_{t-l})\cdots G_1(P_{t-1}))$. The discriminator D_1 is used to distinguish the real temporally coherent sequence $(P_{t-l}\cdots P_t, I_{t-l}\cdots I_t)$ from the fake

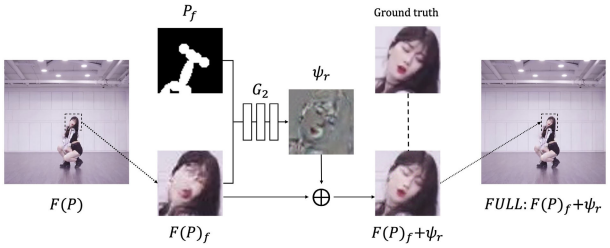

FIGURE 6. The overview of our face refinement network. The residual map ψ_r is used to enhance the facial details of target person.

sequence $(P_{t-l} \cdots P_t, G_1(P_{t-l}) \cdots G_1(P_t))$. The temporal smoothing loss $\mathcal{L}_{ts}(G_1, D_1)$ is given by:

$$\begin{aligned}\mathcal{L}_{ts}(G_1, D_1) = \mathbb{E}_{(P,I)}\left[\log D_1\left(P_{t-l} \cdots P_t, I_{t-l} \cdots I_t\right)\right] \\ + \mathbb{E}_P\left[\log\left(1 - D_1\left(P_{t-l} \cdots P_t, G_1\left(P_{t-l}\right)\right.\right.\right. \\ \left.\left.\left. \cdots G_1\left(P_t\right)\right)\right)\right] \end{aligned} \tag{8}$$

4) FACE REFINEMENT NETWORK

We capture and extract the head information of the target person in the frames, and crop the face images into the size of 64×64. To preserve the facial details of target person, such as the hairstyle, appearance features, facial expressions, we use a GAN to reconstruct the face region. Fig. 6 shows that the face refinement network emphasizes the content integrity and style consistency, which can generate realistic outputs and overcome detail deficiency.

The local face enhancement GAN and global generator network are separate. First, the global generator produces the coarse results which combine the target person and full background. Then we input the cropped images centered around the face $F(P)_f$ and the corresponding pose label maps in the same size P_f to generator G_2 which produces a residual $\psi_r = G_2(P_f, F(P)_f)$. The final result contains a residual map with coarse face region $F(P)_f + \psi_r$ to refine both the full image and the original face. Finally, the optimized face image patch will be embedded into the corresponding region of the original image. Simultaneously, we use a discriminator to guide the training of the model, which tries to differentiate the real face pairs (P_f, I_f) from the fake face pairs $(P_f, F(P)_f + \psi_r)$. The objective function of face enhancement learning can be formulated as follows:

$$\begin{aligned}&\mathcal{L}_{face}(G_2, D_2) \\ &\quad = \mathbb{E}_{(P_f, I_f)}\left[\log D_2\left(P_f, I_f\right)\right] \\ &\qquad + \mathbb{E}_{P_f}\left[\log\left(1 - D_2\left(P_f, F(P)_f + \psi_r\right)\right)\right]\end{aligned} \tag{9}$$

where P_f is the face region of the pose label map P, I_f is the face region of ground truth I_t. The perceptual loss [77] is used to refine the face region I_f of I_t.

IV. GLOBAL NETWORK LOSS

A. DOMAIN ALIGNMENT LOSS

To ensure the transformed features x_S and y_S from different domains into the same domain completely, we set up a loss function. This domain alignment loss is used to align the embedding features of paired images (x_P, r_I). Formally this loss is defined by:

$$\mathcal{L}_{\mathcal{D}}^{\ell_1} = \left\|\mathcal{T}_{P\to S}(x_P) - \mathcal{T}_{I\to S}(r_I)\right\|_1 \tag{10}$$

where x_P is the conditional pose image, and r_I is the refined person image before pose transfer. $\mathcal{T}$ is the domain transformation operation.

B. RECONSTRUCTION LOSS

For the reconstruction term, we argue that the generator should be capable of avoiding distinct color distortions, so that the results are not significantly different from the source images in human visual perception. Therefore, we introduce $\mathcal{L}_{rec}$ to enforce the L_1 distance constraint between the generated images z_I and source images y_I. The reconstruction loss is computed as:

$$\mathcal{L}_{rec} = \left\|z_I - y_I\right\|_1 \tag{11}$$

C. PERCEPTUAL LOSS

To make the generated images look more natural and smooth in RGB color space, we add a perceptual loss, which has been confirmed as available in image generation tasks [63], [64], [77]. The perceptual loss can be written as:

$$\mathcal{L}_{per}^{\ell_1} = \frac{1}{W_\zeta H_\zeta C_\zeta}\sum_{x=1}^{W_\zeta}\sum_{y=1}^{H_\zeta}\sum_{z=1}^{C_\zeta}\left\|\Psi_\zeta\left(I_g\right)_{x,y,z} - \Psi_\zeta\left(I_t\right)_{x,y,z}\right\|_1 \tag{12}$$

where Ψ_ζ is the output feature from layer ζ of pre-trained VGG-19 network on ImageNet [71], in addition, $W_\zeta, H_\zeta, C_\zeta$ are spatial width, height and depth of feature Ψ_ζ. I_g is the pose transformed image and I_t is the original image.

D. CONTEXTUAL LOSS

We use the proposed method [72] to measure the similarity between generated image and ground truth in the human pose transfer task. This method regards an image as a set of features, measuring the feature similarity between images rather than the spatial locations of features. We match all the corresponding features of the original images and the generated images in the global image context to measure the image similarity. Then calculating the similarity between the images according to the similarity between the matching features. Different from previous loss functions, contextual loss is applied to the patch blocks at the corresponding position of the feature layers instead of the entire features. The contextual loss can be formulated as [64]:

$$\mathcal{L}_{context} = -log(CX(\mathcal{F}^l(I_g), \mathcal{F}^l(I_t))) \tag{13}$$

where $\mathcal{F}^l(I_g)$ and $\mathcal{F}^l(I_t)$ are the feature maps extracted from layer l by pretrained VGG19 model for images I_g and I_t severally. CX calculates the similarity between features.

Algorithm 1 End-to-End Training for Our Network

Input: Training images $\left\{I_s^i, P_s^i, I_t^i, P_t^i\right\}_{i=1}^N$, $\left\{\hat{I}_f^i, P_f^i\right\}_{i=1}^N$.

1: Initialize the network parameters.
//Train Human Pose Transfer Network
2: With $\left\{I_s^i, P_s^i, I_t^i, P_t^i\right\}_{i=1}^N$, train $\{\mathcal{G}, \mathcal{D}\}$ to optimize $\mathcal{L}_{\mathcal{D}}^{\ell_1}$, $\mathcal{L}_{rec}$, $\mathcal{L}_{per}^{\ell_1}$, and $\mathcal{L}_{context}$.
//Train Coherent Interframe Consistency Network
3: With $\left\{P_t^i, I_t^i, \hat{I}_t^i\right\}_{i=1}^N$, train $\{G_1, D_1\}$ to optimize $\mathcal{L}_{ts}$.
//Train Face Refinement Network
4: With $\left\{\hat{I}_f^i, P_f^i\right\}_{i=1}^N$, train $\{G_2, D_2\}$ to optimize the loss $\mathcal{L}_{face}$.

Output: $\{\mathcal{G}, \mathcal{D}\}$, $\{G_1, D_1\}$, and $\{G_2, D_2\}$.

E. THE OVERALL LOSS

We adopt an adversarial loss $\mathcal{L}_{GAN}$ with a discriminator $\mathcal{D}$ which differentiates synthesized images from real consecutive frames to help the generator $\mathcal{G}$ learn the distribution of the real data.

$$\mathcal{L}_{GAN} = \mathbb{E}_{(P,I)}[\log \mathcal{D}(P, I)] + \mathbb{E}_I[\log(1 - \mathcal{D}(I, \mathcal{G}(I)))] \quad (14)$$

The total loss function for our human motion transfer network is a linearly weighted sum of above terms:

$$\mathcal{L}_{full} = \min_{\mathcal{G}} \max_{\mathcal{D}} \psi_1 \mathcal{L}_{\mathcal{D}}^{\ell_1} + \psi_2 \mathcal{L}_{rec} + \psi_3 \mathcal{L}_{per}^{\ell_1} + \psi_4 \mathcal{L}_{context} + \psi_5 \mathcal{L}_{ts} + \psi_6 \mathcal{L}_{face} \quad (15)$$

where ψ_i denotes the weight of corresponding loss respectively. The whole training is implemented by solving the objective min-max optimization problem.

Our training process can be summarized in Algorithm 1.

V. EXPERIMENTS

In this section, we conduct extensive experiments on human motion transfer task to evaluate the performance of our proposed framework. The tests include a comparison of several state-of-the-art methods and illustrate its superiority over the baseline models. Furthermore, both the quantitative analysis and ablation study to verify the effectiveness of the network in this paper.

A. DATASETS AND SETTINGS

We use several types of short videos downloaded from YouTube to create a dataset and perform validation. The dataset consists of ballet, jazz and street dance videos, etc. Besides, we also collect a high-quality indoor person video dataset, which we filmed ourselves from 3 to 5 minutes of videos with 1920 × 1080 resolution. We use a stabilizer to hold the phone at a fixed place to avoid jittering in the recorded videos and ensure a static background in all frames. After that, we use the cellphone camera to film the target person of real time footage at 25 frames per second. During this process, we require the subjects to move slowly and act randomly, or try to imitate the actions of source person in a downloaded video. Each video contains the dance of one person, and we train our model separately for each video. Fig. 7 shows our model performs well in dance transfer.

B. EVALUATION METRICS

Similar to previous research, we use several evaluation indicators which are commonly used, such as Structure Similarity (SSIM) [73], Inception Score (IS) [74], Peak Signal-to-Noise Ratio (PSNR), Sharpness Difference (SD), Learned Perceptual Image Patch Similarity (LPIPS) [75] and Fréchet Inception Score (FID) [76] to assess the quality of generated images in the human motion transfer task. These indicators have their own strengths and weaknesses, in order to make the experimental results more convincing, the results of each metric are taken into account. In the following, we will describe them in detail.

1) SSIM

From the perspective of image composition, the structure similarity models distortions as a combination of three factors: luminance, contrast, structure. According to the theory of structural similarity, the image signals are highly structured, and there is a correlation between the closest pixels in the spatial domain. However, SSIM is sensitive to the non-structural distortions of the images (e.g., translation, rotation, scale, etc.), and cannot evaluate images with local or cross distortions well.

2) IS

Inception Score uses Inception V3-Network pre-trained on ImageNet as the classifier, which inputs the generated images and the output values (image categories) of the network are statistically analyzed. Nevertheless, inception score is biased towards the internal weights of the network and the score will be unsatisfied if the images have a different distribution than ImageNet. In particular, IS cannot reflect whether the generate model is over-fitting.

3) FID

Different from IS only considering the generated images, FID uses Gaussian functions to model the features extracted from the Inception network, then calculates the score by computing Fréchet distance between two Gaussians fitted to feature representations. FID has better robustness to noise compared with IS, and the training set is more extensive.

4) LPIPS

LPIPS metric computes distance in AlexNet feature space (conv1-5, pretrained on ImageNet) with linear weights to match human perceptual judgments well. It learns the perceptual similarity and train the network by cross-entropy loss instead of calculating the similarity of the output features via a pre-trained classifier.

5) PSNR

PSNR provides an objective criterion to measure image distortion or noise level, and it is always used in image compression and other fields. However, it is based on error sensitivity between the corresponding pixels, ignoring the

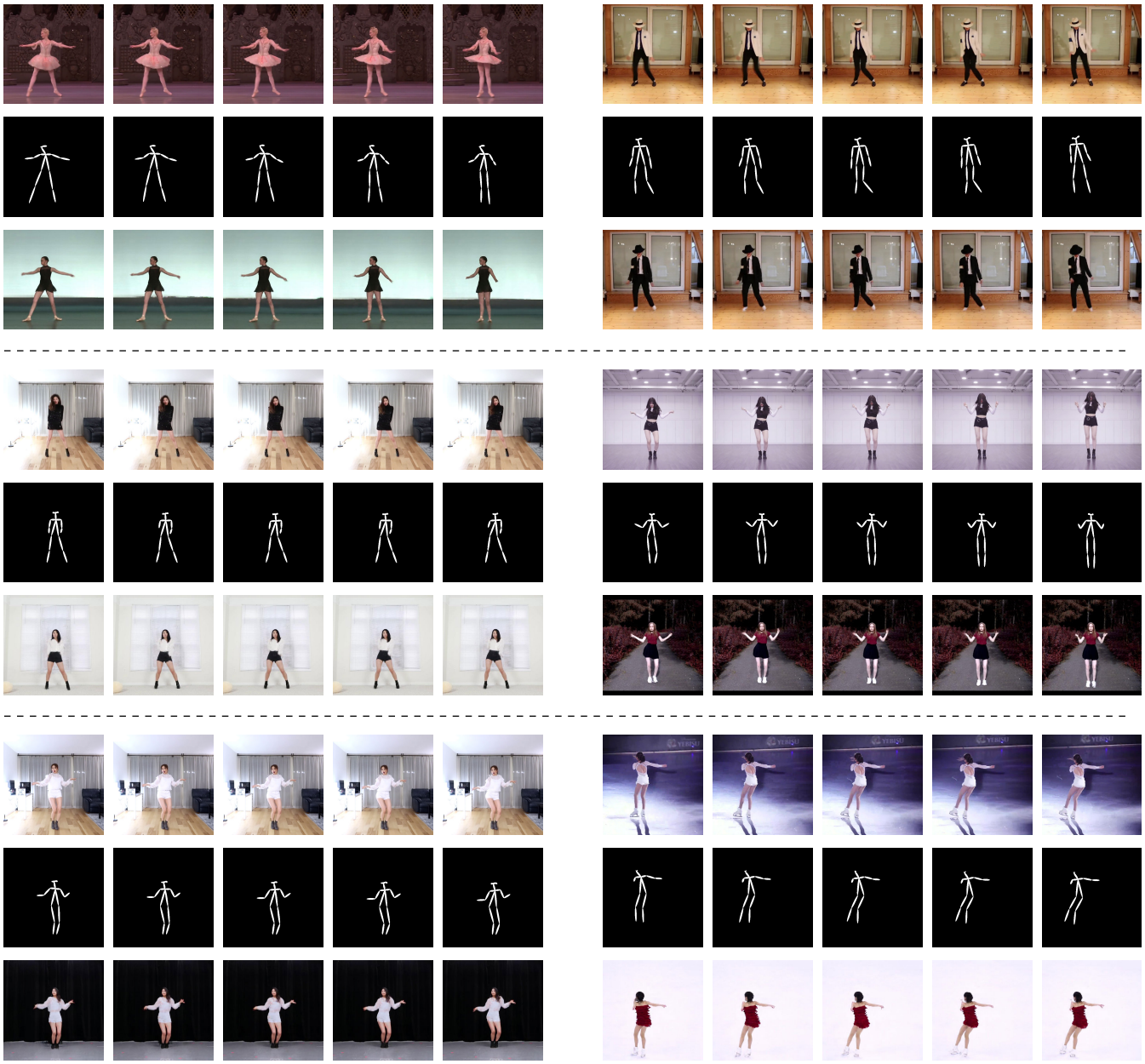

FIGURE 7. The results of human motion transfer. These examples consist of six short videos, we select five consecutive frames from each video. The first row shows the source person, the second row shows the normalized pose skeleton images, and the third row is the generated frames of the target person by our network. The bottom right video sequence shows that our model can not only generate the back of person, but also perform well in some sports pose transfer tasks (Note that other five videos are the results of dance pose transfer).

characteristics of the human visual system (HVS), so the results are often inconsistent with subjective perception.

6) SD

Sharpness difference is an important factor to measure image quality which determines the amount of details an image can represent. It is defined by the boundaries between zones of different tones or colors. A higher SD value indicates a smaller visual difference between the two images on the edge features.

C. ABLATION STUDY

In this section, we conduct extensive ablation study to validate the performance of each component in our network. Fig. 8 shows that temporal smoothing and face enhancement GANs are very important modules to generate realistic images. We also observe that adding each of loss functions can improve the quality of output (see Fig. 9). To be specific, we gradually remove components of the full framework and describe the evaluation of each variant in detail. All the various variants are trained by using the same implementation details to ensure the validity of experiments.

w/o. $\mathcal{L}_{ts}$. This model adopts CoCosNet generator with a face refinement GAN. The reconstructive face images are added to the experimental results.

w/o. $\mathcal{L}_{face}$. The purpose of this experiment is to verify the effectiveness of the face refinement GAN. This variant consists of CoCosNet baseline and temporal smoothing module.

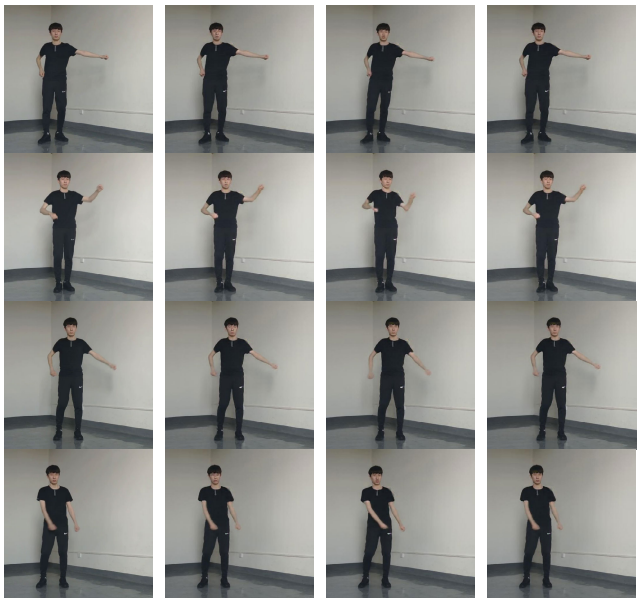

FIGURE 8. The visualization results of ablation study. The left column is a sequence of frames cropped from the target video. It is obvious that the face enhancement component can improve the facial details. Compared with the ground truth, the actions are incontinuous without $\mathcal{L}_{ts}$.

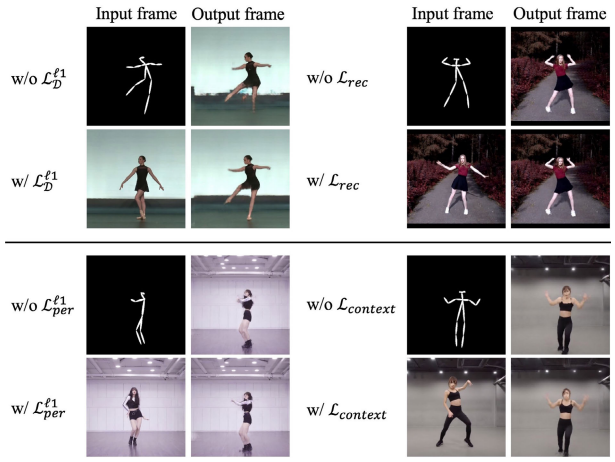

FIGURE 9. The ablation study of different loss functions in human motion transfer module. As we can see, the $\mathcal{L}_{\mathcal{D}}^{\ell 1}$ is effective at synthesizing details. The $\mathcal{L}_{rec}$ is helpful to reduce distortions. Without $\mathcal{L}_{per}^{\ell 1}$, the information of human body is not well preserved. Besides, the $\mathcal{L}_{context}$ ensures the generator to retain the texture in output.

w/o. $\mathcal{L}_{\mathcal{D}}^{\ell 1}$. In this experiment we remove the domain alignment loss to verify this loss can preserve the pixels efficiently. **w/o.** $\mathcal{L}_{rec}$. In order to test the reconstruction term we proposed, this model shows the performance of generating frames without reconstruction loss. **w/o.** $\mathcal{L}_{per}^{\ell 1}$. We analyze the rationality of perceptual loss [77], whether it can output more realistic frames.

w/o. $\mathcal{L}_{context}$. In this test, we evaluate the contextual loss and verify its effect.

Full. We use our proposed full framework in this experiment.

TABLE 1. Ablation study of our proposed model.

Model	SSIM↑	IS↑	FID↓	LPIPS↓
w/.o $\mathcal{L}_{ts}$	0.827	3.326	30.914	0.038
w/.o $\mathcal{L}_{face}$	0.814	3.371	32.257	0.042
w/.o $\mathcal{L}_{\mathcal{D}}^{\ell 1}$	0.796	3.143	34.104	0.053
w/.o $\mathcal{L}_{rec}$	0.810	**3.407**	29.836	0.044
w/.o $\mathcal{L}_{per}^{\ell 1}$	0.803	3.125	33.471	0.051
w/.o $\mathcal{L}_{context}$	0.787	3.168	31.535	0.047
Full	**0.835**	3.364	**27.682**	**0.035**

TABLE 2. Quantitative comparison with state-of-the-art methods on our dataset. (∗) denotes the results tested on our test set.

Method	Metrics					
	SSIM↑	IS↑	FID↓	LPIPS↓	PSNR↑	SD↑
pix2pix*	0.748	3.148	44.801	0.073	19.643	18.476
BicycleGAN*	0.779	3.175	38.279	0.061	20.377	19.027
pix2pixHD*	**0.841**	3.206	32.734	0.044	22.725	20.650
PoseWarp*	0.810	2.975	35.496	0.053	21.870	20.269
vid2vid*	0.839	3.282	28.317	0.038	23.169	20.935
Ours	0.835	**3.364**	**27.682**	**0.035**	**23.472**	**21.183**
Real Data	1.000	4.172	/	0.000	/	/

The assessment results of the ablation study are exhibited in Table 1. Compared with all variants, we can observe that the temporal smoothing significantly improves the capability of keeping frame to frame coherence and always generates photo-realistic human video sequence in consecutive frames. Furthermore, without the guidance of face enhancement network, the facial appearances of the outputs are blurred. The results show that it is unable to refine facial details clearly in the generated frames after removing the face GAN, in addition, temporal information produces positive effectiveness in frame-by-frame generation. We also study the role of each term in the objective loss functions, the results suggest our proposed losses not only substantially help generate natural frames but also significantly improve the visual plausibility of output images. In general, the full model outperforms other variants in terms of generating correct human pose, vivid faces and consistent background, which yields better SSIM, IS, FID and LPIPS scores.

D. COMPARISON WITH STATE-OF-THE-ARTS

1) QUANTITATIVE COMPARISON

In Table 2, we show the quantitative comparison measured by SSIM, IS, FID, LPIPS, PSNR, and SD. But it should be noted that the dataset we used in this work is not same as others, thus we train their models on our dataset instead of using the well-trained models released by their previous works [1], [8], [13], [79], [80]. The results manifest that our method achieves the best performance in terms of most metrics. It indicates that the network proposed in this work can not only produce more realistic details with higher IS value, but more natural textures. Due to the blurry images always achieve higher SSIM score, which has been proved in other works [58], [59], [67], [77], [78], our SSIM value is slightly lower than baselines, but it cannot deny the effectiveness of our model.

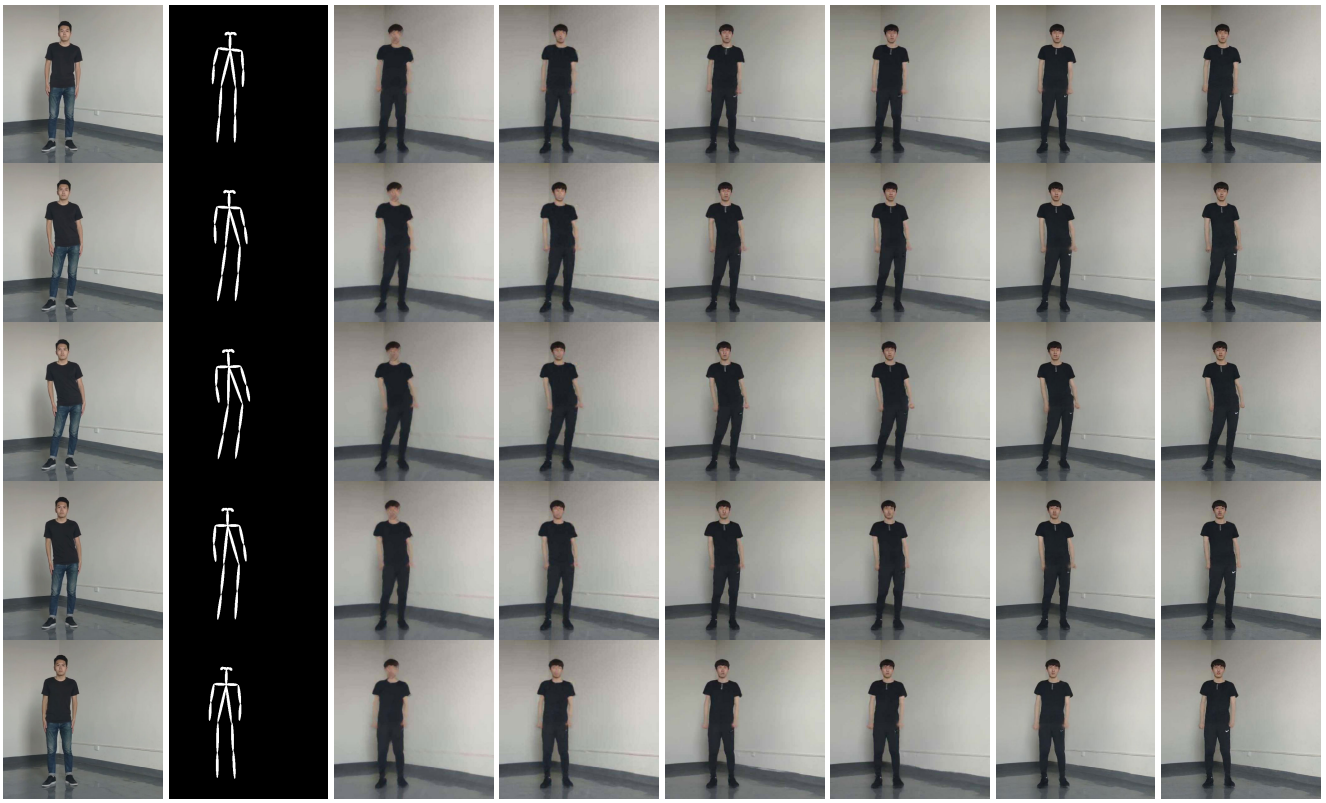

FIGURE 10. The qualitative comparison with several state-of-the-art methods, such as pix2pix, BicycleGAN, PoseWarp, pix2pixHD, and vid2vid. The results show our model preserve the facial details and clothes attributes (e.g., color, texture, logo) better.

Overall, our proposed approach outperforms state-of-the-art methods in human motion transfer task.

2) QUALITATIVE COMPARISON

We compare our proposed method with state-of-the-art methods, i.e., pix2pix [13], BicycleGAN [79], pix2pixHD [8], PoseWarp [80], and vid2vid [1]. The quantitative results are shown in Fig. 10. All the results of these baselines are available by using the open source codes released by authors. As we can see, our model generates more realistic images. Moreover, our method is especially superior in reducing the image artifacts and preserving the global structures and detailed textures, including background, clothes, color, style and shadow.

E. FAILURE CASES ANALYSIS

Overall our proposed method is available to generate convincing results in human motion transfer task. Although our model performs well in most cases, sometimes a few results are not satisfactory. We approximately split them into four categories and expound on each type of errors. Fig. 11 illustrates the examples of failure cases produced by our model. The 1st row shows that our method fails to synthesize frames with rare pose transformation, e.g., crossing limbs flexibly or overlapping legs correctly during an intricate dance of the output video. The 2nd row demonstrates our model is difficult to deduce the frontal face appearance accurately from the back or profile of the target person. From the examples in the 3rd row, we can see that the pose normalization module is unable to produce the pose maps on the scale of the source person, which mainly causes the body pixel-level alignment error and misaligned background pixels. As shown in the 4th row, our model outputs some unsatisfactory results due to the complexity of image background and foreground. In order to solve this challenging issue, [59], [68], [80], [81] introduce the multistage framework which includes three networks: 1) a foreground transformer network synthesizes the human

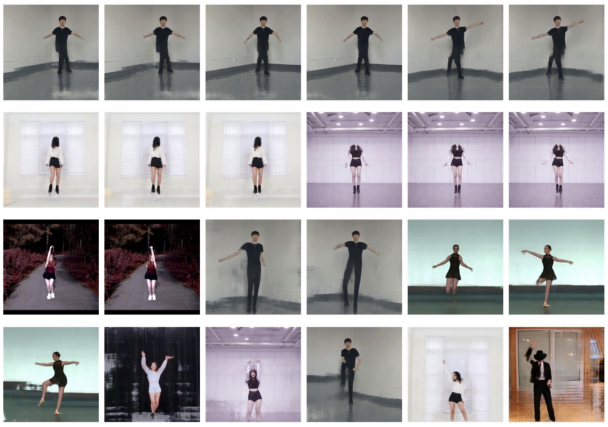

FIGURE 11. Example failure cases caused by pose attributes or modules in our model. We illustrate four different types of typical errors.

body motion. 2) a background transformer network for generating the realistic texture and style of the target images. 3) a fusion module is utilized to blend the generated person images with corresponding background images, then outputs the final results.

VI. CONCLUSION

In this paper, we propose a novel framework for weakly supervised pose-to-video generation, which outperforms in human motion transfer task. Our method uses pose labels as an intermediate representation for video-to-video synthesis. The model consists of a three-stage pipeline. Stage-I represents the input videos as two sequences of frames respectively. Stage-II extracts the pose stick labels, normalizes the pose skeleton figures, and transfers the normalized pose from the source person to the target person. Stage-III combines the frames into video according to the temporal sequences. Extensive experimental results show that our proposed method obtains superior performance in both subjective visual perception and objective metric scores. Additionally, the rationality of each component has been experimentally verified through the ablation study.

However, our model outputs some unsatisfactory results in dealing with large pose changes (e.g., turn around, leap, cross legs or arms). For each video, we should train our model separately, i.e., a well-trained model can only generate videos of the person in the training set, which limits the domain generalization capability. In the future, our work could focus on improving the consistency of generated contents across the whole video, including appearance and background. Moreover, the model can be extended to other persons who are not in the training dataset.

REFERENCES

[1] T. C. Wang, M. Y. Liu, J. Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 1144–1156.

[2] C. Chan, S. Ginosar, T. Zhou, and A. Efros, "Everybody dance now," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Korea South, Oct. 2019, pp. 5932–5941, doi: 10.1109/ICCV.2019.00603.

[3] T. C. Wang, M. Y. Liu, A. Tao, G. Liu, J. Kautz, and B. Catanzaro, "Few-shot video-to-video synthesis," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 1–14.

[4] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, "Dance dance generation: Motion transfer for Internet videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Seoul, South Korea, Oct. 2019, pp. 1208–1216, doi: 10.1109/ICCVW.2019.00153.

[5] J. Ren, M. Chai, S. Tulyakov, C. Fang, X. Shen, and J. Yang, "Human motion transfer from poses in the wild," 2020, *arXiv:2004.03142*. [Online]. Available: http://arxiv.org/abs/2004.03142

[6] Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei, and Y. A. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2242–2251, doi: 10.1109/ICCV.2017.244.

[8] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8798–8807, doi: 10.1109/CVPR.2018.00917.

[9] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[10] L. Yang, P. Wang, C. Liu, Z. Gao, P. Ren, X. Zhang, S. Wang, S. Ma, X. Hua, and W. Gao, "Towards fine-grained human pose transfer with detail replenishing network," 2020, *arXiv:2005.12494*. [Online]. Available: http://arxiv.org/abs/2005.12494

[11] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen, "Cross-domain correspondence learning for exemplar-based image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 5142–5152, doi: 10.1109/CVPR42600.2020.00519.

[12] J. Lin, Y. Xia, T. Qin, Z. Chen, and T.-Y. Liu, "Conditional image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5524–5532, doi: 10.1109/CVPR.2018.00579.

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 5967–5976, doi: 10.1109/CVPR.2017.632.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[15] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1857–1865.

[16] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2868–2876, doi: 10.1109/ICCV.2017.310.

[17] S. Mo, M. Cho, and J. Shin, "InstaGAN: Instance-aware image-to-image translation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–26.

[18] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Long Beach, CA, USA, Jun. 2019, pp. 1954–1963, doi: 10.1109/CVPRW.2019.00247.

[19] J. Pan, J. Dong, Y. Liu, J. Zhang, J. Ren, J. Tang, Y. W. Tai, and M.-H. Yang, "Physics-based generative adversarial models for image restoration and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 24, 2020, doi: 10.1109/TPAMI.2020.2969348.

[20] L. Feng, X. Zhang, S. Wang, Y. Wang, and S. Ma, "Coding prior based high efficiency restoration for compressed video, in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Taipei, Taiwan, Sep. 2019, pp. 769–773, doi: 10.1109/ICIP.2019.8803398.

[21] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Recurrent temporal aggregation framework for deep video inpainting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1038–1052, May 2020, doi: 10.1109/TPAMI.2019.2958083.

[22] S. Lee, S. W. Oh, D. Won, and S. J. Kim, "Copy-and-Paste networks for deep video inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 4412–4420, doi: 10.1109/ICCV.2019.00451.

[23] H. Zhang, L. Mai, H. Jin, Z. Wang, N. Xu, and J. Collomosse, "An internal learning approach to video inpainting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 2720–2729, doi: 10.1109/ICCV.2019.00281.

[24] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Trans. Comput. Imag.*, vol. 2, no. 2, pp. 109–122, Jun. 2016, doi: 10.1109/TCI.2016.2532323.

[25] L. Xiao, S. Nouri, M. Chapman, A. Fix, D. Lanman, and A. Kaplanyan, "Neural supersampling for real-time rendering," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 1–12, Jul. 2020, doi: 10.1145/3386569.3392376.

[26] D. Li, Y. Liu, and Z. Wang, "Video super-resolution using non-simultaneous fully recurrent convolutional network," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1342–1355, Mar. 2019, doi: 10.1109/TIP.2018.2877334.

[27] E. Faramarzi, D. Rajan, F. C. A. Fernandes, and M. P. Christensen, "Blind super resolution of real-life video sequences," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1544–1555, Apr. 2016, doi: 10.1109/TIP.2016.2523344.

[28] Y. Huang, W. Wang, and L. Wang, "Video super-resolution via bidirectional recurrent convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 1015–1028, Apr. 2018, doi: 10.1109/TPAMI.2017.2701380.

[29] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 4482–4490, doi: 10.1109/ICCV.2017.479.

[30] P. Yi, Z. Wang, K. Jiang, Z. Shao, and J. Ma, "Multi-temporal ultra dense memory network for video super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2503–2516, Aug. 2020, doi: 10.1109/TCSVT.2019.2925844.

[31] S. Li, F. He, B. Du, L. Zhang, Y. Xu, and D. Tao, "Fast spatio-temporal residual network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 10514–10523, doi: 10.1109/CVPR.2019.01077.

[32] T. H. Kim, S. Nah, and K. M. Lee, "Dynamic video deblurring using a locally adaptive blur model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2374–2387, Oct. 2018, doi: 10.1109/TPAMI.2017.2761348.

[33] L. Zhang, L. Zhou, and H. Huang, "Bundled kernels for nonuniform blind video deblurring," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 9, pp. 1882–1894, Sep. 2017, doi: 10.1109/TCSVT.2016.2565880.

[34] L. Pan, Y. Dai, M. Liu, F. Porikli, and Q. Pan, "Joint stereo video deblurring, scene flow estimation and moving object segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 1748–1761, 2020, doi: 10.1109/TIP.2019.2945867.

[35] W. Ren, J. Pan, X. Cao, and M.-H. Yang, "Video deblurring via semantic segmentation and pixel-wise non-linear kernel," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1086–1094, doi: 10.1109/ICCV.2017.123.

[36] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "DeblurGAN-v2: Deblurring (Orders-of-Magnitude) faster and better," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 8877–8886, doi: 10.1109/ICCV.2019.00897.

[37] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang, "Deep video deblurring for hand-held cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 237–246, doi: 10.1109/CVPR.2017.33.

[38] W. Ren, J. Zhang, X. Xu, L. Ma, X. Cao, G. Meng, and W. Liu, "Deep video dehazing with semantic segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1895–1908, Apr. 2019, doi: 10.1109/TIP.2018.2876178.

[39] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "AOD-net: All-in-one dehazing network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 4780–4788, doi: 10.1109/ICCV.2017.511.

[40] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "End-to-end united video dehazing and detection," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7016–7023.

[41] C. Li, C. Guo, J. Guo, P. Han, H. Fu, and R. Cong, "PDR-net: Perception-inspired single image dehazing network with refinement," *IEEE Trans. Multimedia*, vol. 22, no. 3, pp. 704–716, Mar. 2020, doi: 10.1109/TMM.2019.2933334.

[42] J. Zhang and D. Tao, "FAMED-net: A fast and accurate multi-scale end-to-end dehazing network," *IEEE Trans. Image Process.*, vol. 29, pp. 72–84, 2020, doi: 10.1109/TIP.2019.2922837.

[43] Y. Li, Q. Miao, W. Ouyang, Z. Ma, H. Fang, C. Dong, and Y. Quan, "LAP-net: Level-aware progressive network for image dehazing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 3275–3284, doi: 10.1109/ICCV.2019.00337.

[44] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 3194–3203, doi: 10.1109/CVPR.2018.00337.

[45] Z. Zhu, J. Lu, M. Wang, S. Zhang, R. R. Martin, H. Liu, and S.-M. Hu, "A comparative study of algorithms for realtime panoramic video blending," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2952–2965, Jun. 2018, doi: 10.1109/TIP.2018.2808766.

[46] Z. Wang, X. Chen, and D. Zou, "Copy and paste: Temporally consistent stereoscopic video blending," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3053–3065, Oct. 2018, doi: 10.1109/TCSVT.2017.2706197.

[47] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, and Y. Zhuang, "Self-supervised spatiotemporal learning via video clip order prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 10326–10335, doi: 10.1109/CVPR.2019.01058.

[48] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 220–237, Jan. 2021, doi: 10.1109/TPAMI.2019.2924417.

[49] J. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and L. Fei-Fei, "Peeking into the future: Predicting future person activities and locations in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 5718–5727, doi: 10.1109/CVPR.2019.00587.

[50] Y.-H. Kwon and M.-G. Park, "Predicting future frames using retrospective cycle GAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 1811–1820, doi: 10.1109/CVPR.2019.00191.

[51] A. Furnari and G. Farinella, "Rolling-unrolling LSTMs for action anticipation from first-person video," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, May 6, 2020, doi: 10.1109/TPAMI.2020.2992889.

[52] X. Liang, L. Lee, W. Dai, and E. P. Xing, "Dual motion GAN for future-flow embedded video prediction," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1744–1752.

[53] W. Lotter, G. Kreiman, and D. Cox, "Deep predictive coding networks for video prediction and unsupervised learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–18.

[54] D. Chen, J. Liao, L. Yuan, N. Yu, and G. Hua, "Coherent online video style transfer," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1114–1123, doi: 10.1109/ICCV.2017.126.

[55] H. Huang, H. Wang, W. Luo, L. Ma, W. Jiang, X. Zhu, Z. Li, and W. Liu, "Real-time neural style transfer for videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 783–791.

[56] M. Ruder, A. Dosovitskiy, and T. Brox, "Artistic style transfer for videos," in *Proc. German Conf. Pattern Recognit.*, 2016, pp. 26–36.

[57] A. Gupta, J. Johnson, A. Alahi, and L. Fei-Fei, "Characterizing and improving stability in neural style transfer," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 4087–4096, doi: 10.1109/ICCV.2017.438.

[58] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. V. Gool, "Pose guided person image generation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 406–416.

[59] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 99–108, doi: 10.1109/CVPR.2018.00018.

[60] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, "Soft-gated Warping-GAN for pose-guided person image synthesis," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Dec. 2018, pp. 472–482.

[61] R. A. Guler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7297–7306, doi: 10.1109/CVPR.2018.00762.

[62] Y. Li, C. Huang, and C. C. Loy, "Dense intrinsic appearance flow for human pose transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 3688–3697, doi: 10.1109/CVPR.2019.00381.

[63] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, USA, Jun. 2019, pp. 2342–2351, doi: 10.1109/CVPR.2019.00245.

[64] Y. Men, Y. Mao, Y. Jiang, W.-Y. Ma, and Z. Lian, "Controllable person image synthesis with attribute-decomposed GAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 5083–5092, doi: 10.1109/CVPR42600.2020.00513.

[65] Y. Ren, X. Yu, J. Chen, T. H. Li, and G. Li, "Deep image spatial transformation for person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 7687–7696, doi: 10.1109/CVPR42600.2020.00771.

[66] A. Pumarola, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer, "Unsupervised person image synthesis in arbitrary poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8620–8628, doi: 10.1109/CVPR.2018.00899.

[67] S. Song, W. Zhang, J. Liu, and T. Mei, "Unsupervised person image generation with semantic parsing transformation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 2352–2361, doi: 10.1109/CVPR.2019.00246.

[68] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Korea (South, Oct. 2019, pp. 5903–5912, doi: 10.1109/ICCV.2019.00600.
[69] G. H. Martinez, Y. Raaj, H. Idrees, D. Xiang, H. Joo, T. Simon, and Y. Sheikh, "Single-network whole-body pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 6981–6990, doi: 10.1109/ICCV.2019.00708.
[70] Z. Liu, H. Hu, Z. Wang, K. Wang, J. Bai, and S. Lian, "Video synthesis of human upper body with realistic face," in *Proc. IEEE Int. Symp. Mixed Augmented Reality Adjunct (ISMAR-Adjunct)*, Beijing, China, Oct. 2019, pp. 200–202, doi: 10.1109/ISMAR-Adjunct.2019.00-47.
[71] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[72] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 768–783.
[73] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[74] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 2234–2242.
[75] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[76] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 6626–6637.
[77] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[78] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 1874–1883, doi: 10.1109/CVPR.2016.207.
[79] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 465–476.
[80] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, "Synthesizing images of humans in unseen poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 8340–8348, doi: 10.1109/CVPR.2018.00870.
[81] C. Si, W. Wang, L. Wang, and T. Tan, "Multistage adversarial losses for pose-based human image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 118–126, doi: 10.1109/CVPR.2018.00020.
[82] T. Park, J.-Y. Zhu, O. Wang, J. Lu, E. Shechtman, A. Alexei Efros, and R. Zhang, "Swapping autoencoder for deep image manipulation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 1–23.
[83] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 2332–2341, doi: 10.1109/CVPR.2019.00244.

HONGYU WANG (Graduate Student Member, IEEE) received the B.S. degree in computer science and technology from Zhengzhou University. He is currently pursuing the M.S. degree with the College of Computer and Cyber Security, Hainan University, Haikou, China. His research interests include computer vision, pattern recognition, deep learning, machine learning, SLAM, and formation control of multi-robot.

MENGXING HUANG (Member, IEEE) received the Ph.D. degree from Northwestern Polytechnical University, in 2007. He joined as a Staff with the Research Institute of Information Technology, Tsinghua University, as a Postdoctoral Researcher. He joined Hainan University, in 2009. He is currently a Professor and a Ph.D. Supervisor of information and communication engineering, and the Dean of the School of Information Science and Technology. He is also the Executive Vice-President of the State Key Laboratory of Marine Resource Utilization in the South China Sea, and the Director of the Hainan Key Laboratory of Big Data and Smart Services. He has published more than 60 academic articles as the first or corresponding author. He has reported 12 patents of invention, owns three software copyright, and published two monographs and two translations. His current research interests include big data and intelligent information processing, ocean information perception and fusion AI and smart service, and so on. He is a Senior Member of the Chinese Computer Federation (CCF) and a member of the Information System Committee in CCF. He has been awarded two second class and one third class prizes of the Hainan Provincial Scientific and Technological Progress.

DI WU was born in 1991. He received the M.S. degree in information and communication engineering from Hainan University, Haikou, China. He is currently a Research Assistant with the School of Information and Communication Engineering, Hainan University. He is the author of over ten articles published in related journals and international conference proceedings. His major research interests include artificial intelligence and information processing.

YUCHUN LI received the B.S. degree from the University of Jinan, Jinan, China, in 2016, and the M.S. degree from the Nanjing University of Science and Technology, Nanjing, China, in 2019. She is currently pursuing the Ph.D. degree with Hainan University, China. Her current research interests include computer-aided diagnostic medical image processing and artificial intelligence.

WEICHAO ZHANG received the B.S. degree from Yangzhou University, Yangzhou, China, in 2019. He is currently pursuing the M.S. degree with Hainan University, Haikou, China. His current research interests include anomaly detection in videos and image processing.