JAFPro: Joint Appearance Fusion and Propagation for Human Video Motion Transfer from Multiple Reference Images

Xianggang Yu*†
FNii, CUHK-Shenzhen
xianggangyu@link.cuhk.edu.cn

Haolin Liu*†
FNii, CUHK-Shenzhen
haolinliu@link.cuhk.edu.cn

Xiaoguang Han‡†§
SRIBD and FNii, CUHK-Shenzhen
hanxiaoguang@cuhk.edu.cn

Zhen Li‡†
SRIBD and FNii, CUHK-Shenzhen
lizhen@cuhk.edu.cn

Zixiang Xiong
Texas A&M University
zx@ece.tamu.edu

Shuguang Cui‡†
SRIBD and FNii, CUHK-Shenzhen
shuguangcui@cuhk.edu.cn

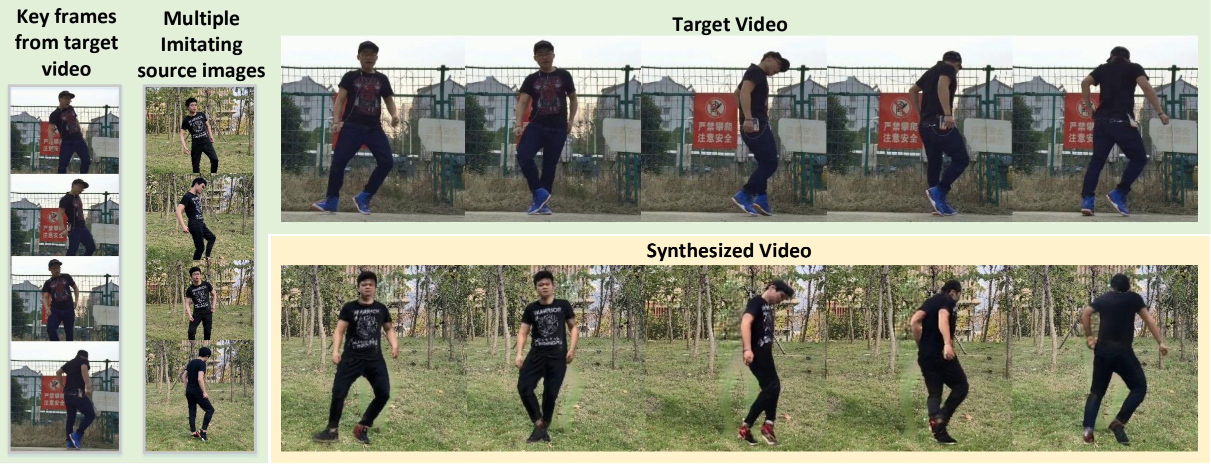
Key frames from target video
Multiple Imitating source images
Target Video
Synthesized Video

Figure 1: Illustration of human video motion transfer using multiple imitating source images. The goal is to generate a video with motion that is consistent with the target video but with appearance from the input source images. Our framework takes multiple source images as inputs and progressively improves the quality of the generated video as more imitating images are provided.

ABSTRACT

We present a novel framework for human video motion transfer. Deviating from recent studies that use only single source image, we propose to allow users to supply multiple source images by simply imitating some poses in the desired target video. To aggregate the appearance from multiple input images, we propose a JAFPro framework that incorporates two modules: an appearance fusion module that adaptively fuses the information in the supplied images and an appearance propagation module that propagates textures through flow-based warping to further improve the result. An attractive feature of JAFPro is that the quality of its results progressively improves as more imitating images are supplied. Furthermore, we build a new dataset containing a large variety of dancing videos in the wild. Extensive experiments conducted on this dataset demonstrate JAFPro outperforms state-of-the-art methods both qualitatively and quantitatively. We will release our code and dataset upon publication of this work.

KEYWORDS

Human video generation, Motion transfer, Multiple reference images, Generative adversarial network

ACM Reference Format:
Xianggang Yu, Haolin Liu, Xiaoguang Han, Zhen Li, Zixiang Xiong, and Shuguang Cui. 2020. JAFPro: Joint Appearance Fusion and Propagation for Human Video Motion Transfer from Multiple Reference Images. In *Proceedings*

*Both authors contributed equally to this research.
†The Future Network of Intelligence Institute, The Chinese University of Hong Kong, Shenzhen
‡Shenzhen Research Institute of Big Data.
§Corresponding author is Xiaoguang Han.

MM '20, October 12–16, 2020, Seattle, WA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3414001

of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3414001

1 INTRODUCTION

Human video motion transfer [1, 6, 10, 31, 38] is a human-centric topic that has recently attracted much attention in the computer vision community. Given a set of source images and a target video, its goal is to generate a video that preserves the appearance of the source images while following the motion in the target video. It has significant applications in (a) rendering synthetic training data of humans performing desired motion for video-related researches such as action recognition and video-based person re-identification [7, 30, 36], (b) the movie industry for transferring dangerous motion onto actors to avoid them doing it themselves, and (c) mass entertainment for allowing users to transfer the motion of an Internet footage (e.g. dance motions) to themselves by simply providing several personal images.

Previous studies on human video motion transfer generally follow two approaches. The first one requires a large number (e.g., more than 10,000) of source images [1, 6, 31] to synthesize a human video. It directly takes all source images as training samples and train a generative model with pose representations as conditional inputs. This approach can generate photo-realistic human video, but it requires every user to collect plenty of sample images and train their own model, which is time-consuming and user-unfriendly. The second approach takes as inputs only one source image of a person [4, 19, 20, 22, 25] and a target video to conduct motion transfer. The network is trained on a large video dataset, enabling the synthesized video to generalize to various identities. Although this approach is flexible, its main drawback is that a single source image (i.e. a single view of the person) obviously does not provide sufficient appearance information, thus the network can only hallucinate unseen textures, resulting in low-quality and unnatural videos.

Therefore, the motivation of this work is how to generate an appealing video of user while requiring as less user-efforts as possible. To this end, we propose to ask user to provide only a few source images by simply imitating some discrete poses from the target video. This is practically useful as it is arduous for a normal user to imitate the whole video sequence since the human motion in a video is dynamic and complicated, e.g. it typically takes 6~8 hours for beginners to learn a dance [2], even if the dance is with only simple arm motions. In contrast, roughly imitating some discrete poses is much easier, as the pose is static and easy to follow. We also use a simple algorithm (c.f. supplementary document) to select discrete poses from a given video for users to imitate. Based on this imitating setting, we propose a novel Joint Appearance Fusion and Propagation framework abbreviated as JAFPro that generates high-quality video by leveraging texture information in the input imitating images. It consists of two parts: an *Appearance Fusion Module* that aggregates information from multiple imitating images progressively and a *Flow-based Appearance Propagation Module* that propagates realistic textures from imitating images to nearby frames in the synthesized video.

JAFPro rectifies shortcomings of two existing approaches because *(1)* we take multiple source images as inputs to deal with the insufficient-information problem. Moreover, we devise an elaborate appearance fusion module that accepts arbitrary number of input images and integrates complementary information from them for progressively improved results as more images are supplied. Compared to simple appearance fusion methods that accept variable length of inputs (max fusion), our appearance fusion module can adaptively incorporate information from multiple inputs and thus preserve sharper details as validated in the experiment section 4.2; *(2)* Differ from the first approach [1, 6, 31], JAFPro is trained on a large video dataset consisting of various motion videos in the wild. This makes JAFPro more generalized in terms of synthesizing arbitrary human motion. Furthermore, providing input images in an imitating manner makes the input images posses similar appearance with their neighbors in the synthesized video, so that flow-based propagation can be utilized to convey textures from these imitating images to their neighboring frames so to improve results.

In summary, the key contributions of this work are:

- A novel human video motion transfer framework that progressively improves the quality of generated video as the number of input source images increases.
- A joint appearance fusion and propagation pipeline that fully exploits the information of multiple source images.
- A novel fusion scheme conducted on the texture atlas that effectively fuses textures from multiple source images.

2 RELATED WORK

Person Image Generation Given a reference image of a person and a target pose, person image generation aims to generate an image in which the person in reference image will act as the same pose as the target pose. The work of [20] is the pioneer for this task, in which they use 2D keypoints to represent a target pose, and synthesize an output image conditioned on the concatenation of source image and target pose in a coarse-to-fine manner. However, directly concatenate source image with target pose without modeling their misalignment will induce unpleasing artifacts. Several attempts have been made to deal with this unaligned problem. Ma *et al.* [21] propose a disentangled approach which separately encodes foreground, background and pose features, then tiles and decodes these features to an image to alleviate the unaligned effect. Zhu *et al.* [41] view the deformation as a transition on the person image manifold, and propose to progressively transfer reference image to target pose through a sequence of intermediate pose representations. On the other hand, a series of works [4, 25, 40] try to explicitly study the deformation. Specifically, they use a set of part-based affine transformations to approximate the non-rigid person deformation. Another branch of explicit deformation-modeling methods borrow the power of 3D human model, including [16, 19, 22, 39]. Our method also explicitly models the deformation between reference images and target video by leveraging 3D human models, but a major difference from previous methods is that we take as input multiple source images, therefore an adequate fusion method must be proposed upon aforementioned single source-based methods to aggregate multi-source information.

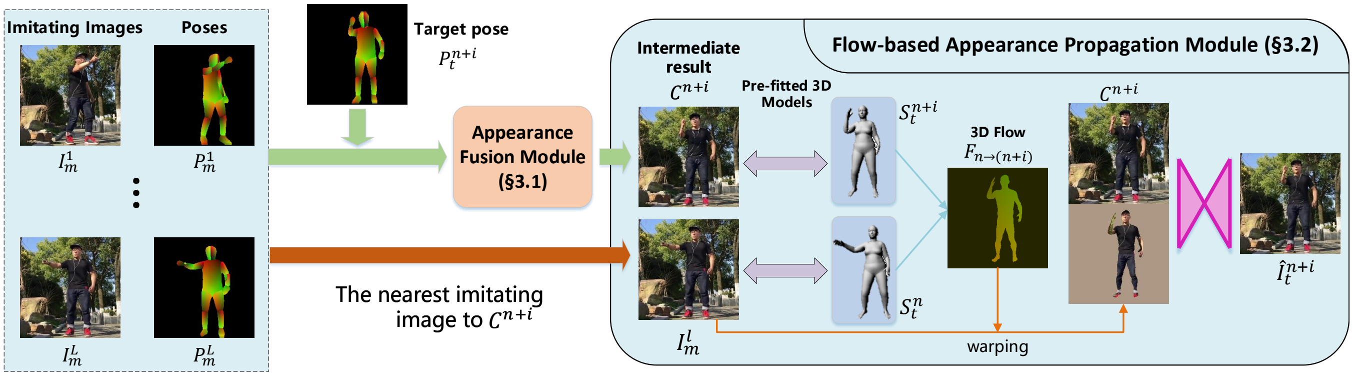

Figure 2: Overview of our JAFPro framework. It comprises an Appearance Fusion Module and a Flow-based Appearance Propagation Module, which are end-to-end trained to produce the final output.

Human Video generation In terms of the requirements of modeling human body dynamics and temporal coherence, human video generation is a more challenging task than its image counterpart [38]. There are two divisions in the literature of synthesizing human videos: 1) The first track [1, 6, 31] tries to train identity-specific model for generating photorealistic videos. More specifically, for each user, they first ask him/her to provide their own videos, with as diverse poses and as many frames as possible. Based on these videos, a pose-to-image model is trained using image-to-image translation networks [12, 32]. During the inference, a video exhibiting the user's appearance can be synthesized from any pose sequence using the trained model. Moreover, to generate production-quality videos, Liu *et al.* [18] introduce 3D model reconstruction combined with training a character-to-image model. Despite the high-quality results these methods present, their major difference from ours is that they require, for each identity, a time-consuming data collection and model training procedure. 2) The works in another division focus on generating a sequence of motion-consistent person images based on a sole starting frame [10, 24, 28, 38]. However, these methods are more similar to video prediction, in which the future dynamics must be inferred from the network [28, 38]. Compared with them, we concentrate on human motion transfer where the target dynamics is given.

Multi-reference Based Image Synthesis 3D human reconstruction from multi-view observations has been studied for several years [8, 9, 15, 17, 23, 29]. Recent studies also adapt this idea for multi-view based image synthesis [27, 37]. Sun *et al.* [27] propose a trainable network consisting of a flow prediction module and recurrent pixel generation module to exploit information from multiple viewpoints for synthesizing a novel view. However, without explicit 3D supervision, the flow prediction tends to fail when the object is non-rigid, e.g. the moving human body in our work. In [37], a network built upon 3D convolutions and attention map prediction module is proposed to aggregate appearances from multi-view sources. But the inputs of their methods are limited to 'photometric' images captured under specific condition, thus not feasible for our case.

3 ALGORITHM

Given a target motion video $[I_t^1, I_t^2, ..., I_t^n, ..., I_t^N]$ (where n is the frame index within the target video) and multiple source images $[I_m^1, I_m^2, ..., I_m^l, ..., I_m^L]$ (where l denotes the index within sources) imitating some poses of the target video, that is, the pose of I_t^n and I_m^l are roughly aligned, our goal is to generate a video that combines the appearance of $\{I_m^i\}_{i=1}^L$ and the motion of $\{I_t^i\}_{i=1}^N$. Moreover, we want our framework to be able to take as input arbitrary number of imitating images, i.e. L is not fixed, and progressively improve the quality of the synthesized video as L increases. To this end, as depicted in Figure 2, first, an Appearance Fusion Module (sec 3.1) absorbs multiple imitating images $\{I_m^i\}_{i=1}^L$ together with their Denseposes[3] $\{P_m^i\}_{i=1}^L$; then, it outputs an intermediate result C^{n+i} conditioned on the target Densepose P_t^{n+i}, where i is the frame interval between (I_t^n, I_t^{n+i}) and can be either negative (preceding frames), zero (current frame), or positive (subsequent frames). The appearance fusion module leverages a recurrent model to aggregate information from multiple imitating images so that it can progressively improve the quality of C^{n+i}. In the sequel, a Flow-based Appearance Propagation Module finds the nearest imitating frame I_m^l to C^{n+i} and propagates its textures to C^{n+i} via a dense flow field computed from 3D models (more details will be covered in sec 3.2). Finally, the output of the whole framework $\hat{I}_t^{n+i}$ is a combined version of the propagated textures and C^{n+i}.

3.1 Appearance Fusion on Texture Atlases

Given the user-provided source imitating images $\{I_m^i\}_{i=1}^L$, our goal is to fuse the appearance of them, which is a challenging task. Direct fusion (e.g. max) on feature space tends to result in blurry result because input images are not aligned. Therefore, as shown in Fig. 3, we first transfer the human appearance from RGB images onto texture atlases using their DensePoses [3] $\{P_m^i\}_{i=1}^L$, yielding a set of texture atlases $\{T_m^i\}_{i=1}^L$ (Fig. 3 only presents two consecutive atlases $[T_m^{l-1}, T_m^l]$ for simplicity); then, we conduct fusion on these texture atlases. The texture atlas stores the human texture in a structured way so that textures from different source images are approximately aligned. This make it more feasible to conduct fusion on texture atlases than on RGB images.

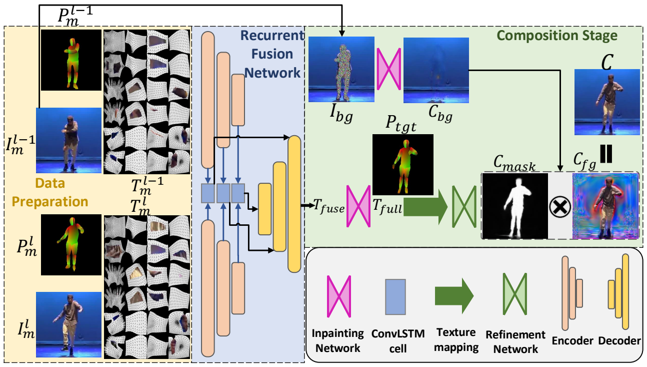

Figure 3: Illustration of the Appearance Fusion Module. We only present two consecutive imitating images (I_m^{l-1}, I_m^l) as an example for better illustrating the recurrent fusion mechanism.

On the other hand, designing an appropriate fusion method is essential, as textures extracted from different source images usually have both overlapping and exclusive areas. A proper fusion method is expected to retain all textures in exclusive areas while selectively absorbing information from different sources in the overlapping areas. To this end, inspired by [8, 27], we propose to use Convolutional Long-Short Term Memory [35] (ConvLSTM) for appearance fusion. This recurrent fusion scheme can not only adaptively fuse appearance from different sources, but also accept arbitrary number of input source images and improve the output video quality as more images are fed into the network. More details are given in the sequel.

Recurrent Fusion Network. The list of texture atlases $\{T_m^i\}_{i=1}^L$ are fed into a recurrent fusion network sequentially to aggregate their information. Specifically, for two consecutive atlases (T_m^{l-1}, T_m^l), we employ an encoder to encode their multi-scale features that flow into the ConvLSTM cells in the same order. The ConvLSTM cell will cast away unqualified features and only retain qualified information from the new incoming features. After features from all texture atlases are fused within the ConvLSTM cells, these fused multi-scale features are sent to the decoder in a skip-connection manner. In the end, a fused texture atlas T_{fuse} is decoded.

Composition Stage. Subsequently, we cut out the background from source images and inpaint the missing pixels to yield a new background C_{bg}. Although the fused texture atlas T_{fuse} aggregates information from all image sources, there are still some missing pixels. Hence another inpainting network is needed to hallucinate the missing parts before we obtain a full texture atlas T_{full}, which mapped back into the image using the target DensePose P_t later. The remapped image is then sent to a refinement network to obtain a foreground output C_{fg} and its corresponding mask C_{mask}. Finally, the background and foreground are combined by the following equation:

$$C = C_{fg} \circ C_{mask} + C_{bg} \circ (\mathbf{1} - C_{mask}), \tag{1}$$

where $\circ$ is element-wise multiplication. We treat C as an intermediate result as it will enter the Flow-based Appearance Propagation Module for further refinement as described in section 3.2.

Training Strategy. The training of AFM is very difficult because we do not have a complete texture atlas to supervise the fusion process. Therefore, to stabilize the training of AFM, we separately train the Recurrent Fusion Network (RFN) with a self-supervised setting. We refer readers to our supplementary document for more details.

3.2 Flow-based Appearance Propagation

Although the appearance fusion module introduced in section 3.1 can aggregate multi-source information as much as possible, there are still some missing high frequency textures in the fusion result. There are two reasons for this loss of details: 1) the dense correspondences between person images and texture atlases are sometimes inaccurate (i.e. the estimated DensePose is not accurate), thus a small portion of high frequency details will be lost during the process of texture mapping and remapping; 2) there are still a small part of texture information from multiple imitating source images carried onto the texture atlases are misaligned when they are fed into the recurrent fusion network. These fuzzy inputs cause the network to output a blurry image since the network is agnostic to which source of texture is the right-aligned one. Therefore, we propose a Flow-based Appearance Propagation Module (FAPM) to convey detailed textures from multiple source images to the blurry part caused by previous fusion module. The design of FAPM is based on the following observation: as imitating images are roughly aligned with some frames in the target motion video, we observe that the appearance of imitating images is similar to their neighboring frames in the synthesized video, as long as the motion between an imitating image and its neighboring frames is relatively small. Therefore, the high frequency details from an imitating image can be exploited to enhance the blurry parts of its neighboring frames through a flow-based propagation scheme.

3D Flow Calculation. We first estimate the dense motion flow between an imitating image and its neighboring frames. Instead of using optical flow for motion estimation, we calculate 3D flow [19] since the motion range in our video dataset is too large for optical flow to handle. Specifically, we fit two SMPL models [5, 13] for each pair of frames. This way a dense flow field can be computed from the matching vertices on two SMPL models.

Flow-based Propagation. As shown in Fig. 2, for an intermediate result C^{n+i} from AFM containing some blurry parts, we locate its nearest imitating image I_m^l (i.e. the imitating image that most similar to C^{n+i}). Let (S_t^n, S_t^{n+i}) denotes their pre-fitted SMPL models, then a 3D flow $F_{n\rightarrow(n+i)}$ can be calculated from two models; then, the imitating frame I_m^l can be warped by $F_{n\rightarrow(n+i)}$ via bilinear sampling (BS):

$$\widetilde{I}_m^l = \mathcal{BS}(I_m^l, F_{n\rightarrow(n+i)}), \tag{2}$$

where $\widetilde{I}_m^l$ is the warped imitating image that is aligned with C^{n+i}. Due to the temporal redundancy, most high frequency contents are preserved in $\widetilde{I}_m^l$ and ready to be added onto C^{n+i}. But there

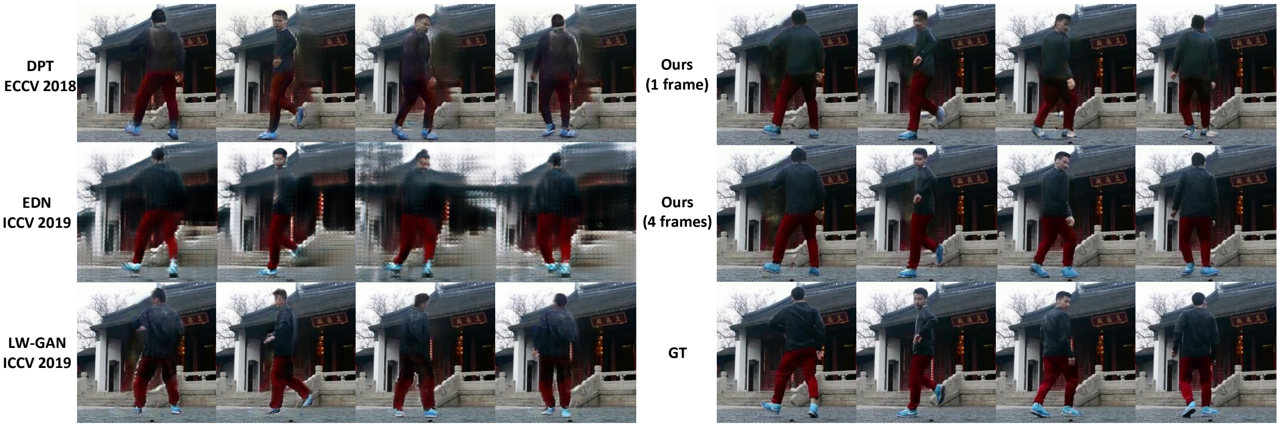

Figure 4: Qualitative comparison on the task of video reconstruction with state-of-the-art methods:DPT [22], EDN [6], and LW-GAN [19]. 5 frames are sampled from a generated video. We observe that our method reconstructs sharper and more detailed textures.

are still a small portion of misaligned and incorrect textures, so we further employ a encoder-decoder network to learn a confidence map to determine the correctness of each pixel. Let W denotes this confidence map, then the final result $\hat{I}_t^{n+i}$ is a weighted combination of $\widetilde{I}_m^l$ and C^{n+i}:

$$\hat{I}_t^{n+i} = C^{n+i} \circ W + \widetilde{I}_m^l \circ (\mathbf{1} - W), \tag{3}$$

where W is a fractional weighting factor.

3.3 Joint Training

Although AFM and FAPM can be trained separately, joint training can further improve the performance of our proposed algorithm because the propagated high-frequency textures from FAPM can be further regarded as a guidance for the fusion process, thus facilitate each other. Therefore, we jointly train the appearance fusion and propagation framework in an end-to-end fashion. During training, L source images are fed into the AFM to synthesize an intermediate result C. Then, one of the source image $I_m^l, 1 \leq l \leq L$ is randomly chosen and concatenated with C before being fed to the FAPM to yield $\hat{I}_t$. The final result is compared to the ground truth target image I_t.

Objective functions. There are four loss terms during training: an $L1$ loss $\mathcal{L}_1$, a perceptual loss $\mathcal{L}_{per}$, a global adversarial loss $\mathcal{L}_{GAN}$, and a facial adversarial loss $\mathcal{L}_{faceGAN}$. The total objective function is thus

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{per} + \lambda_3 \mathcal{L}_{GAN} + \lambda_4 \mathcal{L}_{faceGAN}, \tag{4}$$

where λ_1,λ_2,λ_3,λ_4 are the corresponding weighting factors.
The $\mathcal{L}_1$ loss is essential in stabilizing training, which is obtained by computing the L1 distance between the generated image $\hat{I}_t$ and the ground truth I_t:

$$\mathcal{L}_1 = \mathbb{E}[\| I_t - \hat{I}_t \|_1]. \tag{5}$$

The perceptual loss is defined as the L1 distance between the synthesized images and the ground truth in the feature space. The features are extracted from a VGG-19 [26] network. Five layers of features extracted from the VGG-19 are utilized to compute the perceptual loss as

$$\mathcal{L}_{per} = \mathbb{E}[\| \phi(I_t) - \phi(\hat{I}_t) \|_1]. \tag{6}$$

The adversarial loss is used to facilitate synthesis of details by fooling a discriminator D, which classifies the ground truth and synthesized frames as real or fake. The discriminator is also conditioned on the target DensePose P_t. The adversarial loss in our experiments is

$$\mathcal{L}_{GAN} = \mathbb{E}[\log D(\hat{I}_t, P_t) + \log(1 - D(I_t, P_t))]. \tag{7}$$

In addition, the facial adversarial loss is used to help make the resulting face image look realistic [6]. During training, the facial area in $\hat{I}_t$ and I_t are cropped and resized to a fix-size square image before going through the discriminator. Denote the cropped face images as $\hat{I}_{face}$ and I_{face}, then the adversarial loss is

$$\mathcal{L}_{faceGAN} = \mathbb{E}[\log D(\hat{I}_{face}) + \log(1 - D(I_{face}))]. \tag{8}$$

4 EXPERIMENT

Dataset. We build a DanceVideo dataset by collecting 1651 dance videos in the wild with static background from Internet. Each video has a length of two seconds with 15FPS (30 frames in total). The dance videos include two popular dance classes: Jazz and Popping. For each video, we use DensePose [3] to extract the poses of dancers. We then crop and resize all images and DensePose maps to 256 × 256 to construct the final dataset.

Implementation Details. We pretrain the Appearance Fusion Module for 100 epochs, then jointly train the whole framework for another 60 epochs. The batch size is set to 4 through the training. The whole training is conducted on four NVIDIA RTX 2080Ti, and costs 33 hours. For each training sample, the number of input source images randomly varies from 1 to 4 following an uniform distribution. The parameters in the objective function are set to $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 2, \lambda_4 = 2$. We use Adam [14] optimizer with learning rate of $1e - 4$ for the Appearance Fusion Module, $5e - 5$ for

Table 1: Quantitative comparison against the state-of-the-art methods. Values are calculated on entire image and foreground, respectively. ↑ means higher is better; ↓ means lower is better.

Methods	SSIM ↑	MS-SSIM↑	PSNR↑	L1 error↓	Flow error↓
EDN[6]	0.604/0.895	0.570/0.843	17.33/25.49	0.160/0.033	17.331/4.071
DPT[22]	0.840/0.919	0.873/0.916	23.93/26.97	0.067/0.026	2.937/2.094
LW-GAN[19]	0.828/0.905	0.815/0.853	22.13/24.55	0.069/0.034	3.788/2.393
Ours (1 frame)	0.873/0.932	0.898/0.926	24.90/27.94	0.052/0.023	2.718/2.019
Ours (4 frames)	**0.880/0.939**	**0.915/0.943**	**25.32/28.67**	**0.049/0.020**	**2.660/1.993**

(entire picture/foreground only)

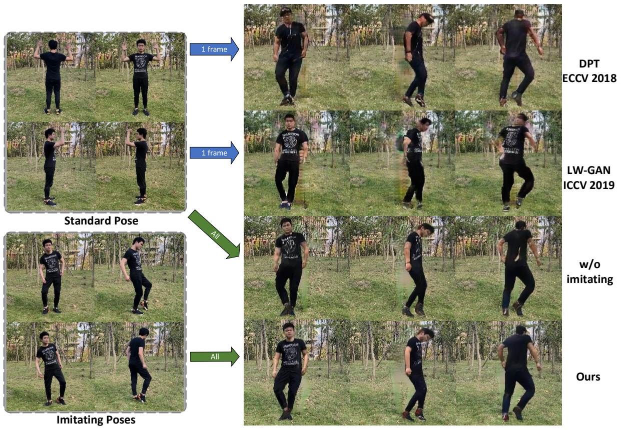

Figure 5: Qualitative comparison with EDN [6], DPT [22] and w/o imitating setting on the task of motion transfer. Images on the left are sources, while 3 frames are sampled from the transferred video are on the right (better viewed with zoom in).

the Flow-based Appearance Propagation Module, and $3e-6$ for the discriminator. For the GAN training, we use an alternating gradient descent scheme in which the discriminator is updated thrice per generator update.

4.1 Comparison against State-of-the-Arts

We compare our proposed JAFPro framework against three closely-related state-of-the-art methods: Everybody Dance Now[6] (abbr. EDN), DensePose Transfer[22] (abbr. DPT), and Liquid Warping GAN[19] (abbr. LW-GAN). EDN is a motion transfer framework focusing on dancing videos, which can be trained on our dataset and conduct comparison. DPT and LW-GAN are two single-reference based pose transfer framework, which leverages the Densepose and 3D flow for generation, respectively. We evaluate the performance on two tasks: 1) video reconstruction that takes one (in single-source setting) or multiple (in our multi-source setting) frames from a video as input and reconstructs the whole video; 2) motion transfer that transfers the motion of a video to a new appearance.

Evaluation Protocol. For video reconstruction task in which the ground truth video is available, we adopt structural similarity index (SSIM) [33] and its multi-scale variant MS-SSIM [34], Peak signal-to-noise ratio (PSNR), and $L1$ error to evaluate the quality of synthesized results. Moreover, we also calculate the mean difference between optical flows of synthesized video and ground truth video estimated by a pretrained flow estimator [11] to evaluate the temporal coherency. We evaluate the results of the entire picture and the foreground separately. For motion transfer task whose result has no ground truth, we only compare them qualitatively.

Video Reconstruction. For this task, we train all competing methods on our DanceVideo dataset and evaluate them on the validation set. Note that EDN [6] requires training a identity-specific model for each video in the validation set, so we use half of video (only 15 frames) to train a model and reconstruct the whole video utilizing the trained model. The quantitative comparisons are shown in Table. 1. Our method with multiple source images outperforms competitors in a large margin in terms of all evaluation metrics, the qualitative comparison in Fig. 4 also demonstrate this. Since DPT and LW-GAN take one source image as input, for a fair comparison, we also compare our method with 1 frame as input against them in Table. 1 and Fig. 4. Their results are still worse than ours, owing to the DensePose (DPT) or 3D flow (LW-GAN) they rely on are difficult to be accurately estimated in our scenarios as the videos contains more complex backgrounds. In contrast, this downside is somewhat alleviated in our method by the combination of AFM and FAPM with a joint training strategy. The ablation study in Sec 4.2 also confirms this.

Motion Transfer. In this task, we invite a volunteer who has never learned to dance, and transfer the motion of professional dance video to him (i.e. make him dance like a pro). The volunteer provides one source image with a front-view standard pose for DPT and LW-GAN, and supplies multiple source images to our method by imitating some discrete poses in the target motion video. Note that the volunteer is just asked to strike roughly similar poses to the references, so the imitating process is very quick. The synthesized videos are presented in Fig. 5. Our approach generates more realistic video than DPT and LW-GAN. In addition, we also ask volunteer to supply source images to our method without imitating, in this case four views covering most body areas are inputted. As shown in the third row of Fig. 5, despite offering more textures, the transferred video from standard poses preserves less details (especially for the

Table 2: Quantitative comparison of ablation studies. ↑ means higher is better; ↓ means lower is better.

Variants	SSIM ↑	MS-SSIM↑	PSNR↑	L1 error↓	Flow error↓
w/o fusion	0.827	0.841	22.623	0.0733	4.051
w/o propagation	0.869	0.900	24.795	0.0523	2.848
w/o joint	0.867	0.897	24.827	0.0546	2.854
full	**0.880**	**0.915**	**25.322**	**0.0492**	**2.660**
no shift (k=0)	0.880	0.915	25.322	0.0492	2.660
1 frame shift	0.876	0.911	25.147	0.0502	2.677
2 frames shift	0.874	0.908	25.044	0.0508	2.686
3 frames shift	0.873	0.906	24.980	0.0512	2.692
4 frames shift	0.872	0.905	24.962	0.0514	2.690
1 frame	0.873	0.898	24.901	0.0517	2.718
2 frames	0.878	0.911	25.218	0.0498	2.687
3 frames	0.879	0.914	25.284	0.0494	2.671
4 frames	**0.880**	**0.915**	**25.322**	**0.0492**	**2.660**
Direct fusion (3 frames)	0.867	0.895	24.659	0.0542	-
Max fusion (3 frames)	0.862	0.894	24.539	0.0538	-
ConvGRU (3 frames)	0.863	0.892	24.471	0.0543	-
ConvLSTM (3 frames)	**0.869**	**0.900**	**24.745**	**0.0524**	-

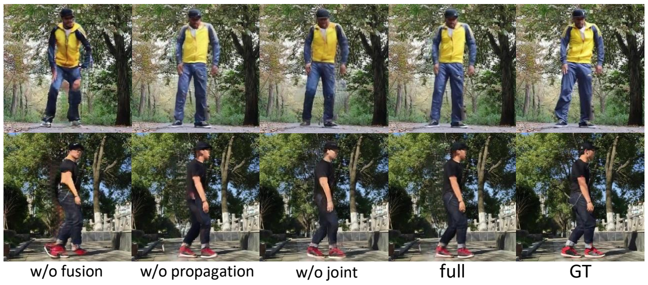

Figure 6: Zoom in for details. Qualitative comparison of different variants of our framework. By ablating different modules and joint-training setting, our full framework retains a large portion of sharper textures with less distortions.

wrinkles on pants, the textures on shoes and pattern on clothes), which confirms the importance of the imitating setting.

4.2 Ablation Study

To investigate the impacts of different modules and settings in our framework, we conduct ablation studies by removing/changing some components/settings in this section.

Efficacy of Key Contributions. To validate the effectiveness of our joint appearance fusion and propagation framework, we compare our full model with the following variants: *w/o appearance propagation*, *w/o appearance fusion*, and *w/o joint training*. The quantitative and qualitative results are shown in Table. 2 (first part) and Fig. 6, respectively. First, it can be seen that the appearance fusion module (AFM) contributes significantly to the output because the *w/o appearance propagation* variant achieves preferable numerical results. This demonstrates the elaborate AFM can aggregate information from multiple source images effectively. Second, without AFM providing the intermediate result, the propagation module itself cannot generate concordant results although preserves high frequency contents (c.f. results of *w/o appearance fusion* variant). This also indicates the advantage of our joint-framework design. Third, in the *w/o joint training* variant, we train two modules (AFM and FAPM) separately in a stage-by-stage manner. The inferior results demonstrate that the joint training of two modules (which corresponds to our full model) leads to the best results.

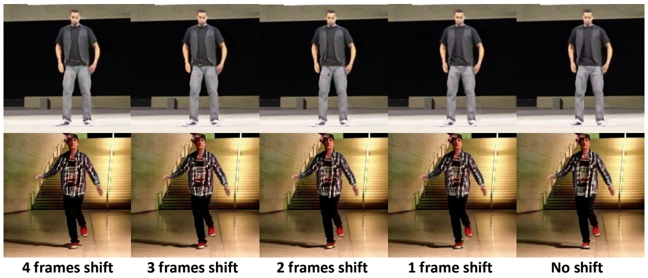

Figure 7: Zoom in for details. Visual comparison of the results from different number of frame shift. When the number of frame shift increases, the perceptual change between the synthesized video is very small, which demonstrate our robustness to imitating accuracy.

Robustness to imitating accuracy. As claimed before, our framework only requires user to roughly imitate some poses from target motion video, i.e. the imitating images can be slightly misaligned with references. To further investigate the effect of imitating accuracy, we conduct an experiment for the video reconstruction

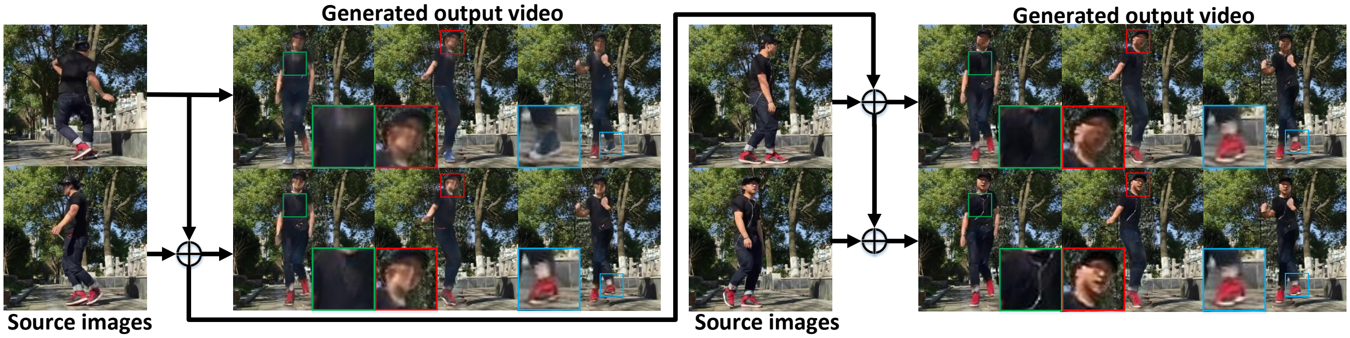

Figure 8: Qualitative illustration of using different number of source images. We gradually increase the number of source images from top to bottom until reaching 4 sources in total. Three frames are sampled from the synthesized video at each time. It is observed that increasing the number of source images greatly improves the results especially when the texture information are mutually complementary.

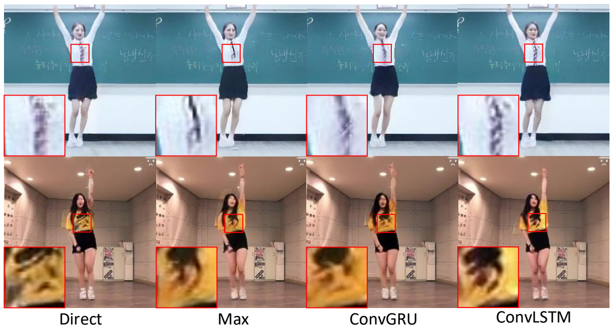

Figure 9: Qualitative comparison of using different fusion methods. Comparing to other methods, our ConvLSTM fusion method retains sharper textures as highlighted in the red boxes.

task by shifting the input source image to its reference pose by k frames. For example, given P_t^n as reference, input pair with $[I_t^n, P_t^n]$ is shifted to $[I_t^{n+k}, P_t^{n+k}]$, and textures from this shifted pair are propagated to the neighbors of P_t^n. This way we deliberately misplace the imitating image with its reference pose. Table. 2 (second part) and Fig. 7 show the results of varying k. When k increases (i.e. the imitating becomes more inaccurate), the decrease in numerical results are within a very small range, and the visual difference is also minor, which demonstrate our system is robust to slightly dissimilar imitating images.

Effect of Number of Source Images. As discussed in the introduction, existing methods based on only one source image will suffer from the insufficient-information problem. In contrast, our framework takes multiple source images as input and can progressively improve the synthesized results as more imitating images are supplied. To verify this, we conduct comparative experiments by ablating the number of source images. Specifically, we input $L, 1 \leq L \leq 4$ source images into our framework and compare the final synthesized video at each time. The quantitative results shown in Table. 2 (third part) indicate that increasing the number of source images improves the quality of synthesized video gradually. Moreover, from the qualitative result in Fig. 8, we can observe a significant promotion on the fine-grained textures such as clothing, shoes, and facial details as more source images are added.

Fusion Methods. In order to show the effectiveness of our proposed ConvLSTM fusion in AFM, we compare it with several other fusion methods: (1) *Direct* fusion which concatenates multiple texture atlases before feeding them to the network; (2) *Max* fusion which aggregates features of multiple texture atlases by a max pooling; (3) *ConvGRU* fusion that utilizes Convolutional Gated Recurrent Units in the RFN. We solely train the AFM with 3 frames as input for better ablating on the fusion method. The qualitative results in Fig. 9 show that our ConvLSTM fusion method preserves sharper textures while other methods tend to cast away important textures or blur the results. And the numerical comparison in Table. 2 (fourth part) also supports our superiority.

5 CONCLUSION

We propose a joint appearance fusion and propagation framework named JAFPro for human video motion transfer from multiple imitating images. By designing an elaborate fusion scheme, our framework is able to take arbitrary number of source images as input, and the information from multiple sources are effectively aggregated. The synthesized video are also progressively improved as more sources are supplied. With the imitating setting, our framework propagates realistic textures from imitating image to its nearby frames through flow based warping, which further improves the results. Compare to previous works, our framework gains superior results in terms of both visual perception and quantitative measures. Furthermore, a series of ablation studies also verify the efficacy of our key contributions.

ACKNOWLEDGMENTS

The work was supported in part by the Key Area R&D Program of Guangdong Province with grant No. 2018B030338001, by the National Key R&D Program of China with grant No. 2018YFB1800800, by Natural Science Foundation of China with grant NSFC-61629101, by Guangdong Zhujiang Project No. 2017ZT07X152, and by Shenzhen Key Lab Fund No. ZDSYS201707251409055.

REFERENCES

[1] Kfir Aberman, Mingyi Shi, Jing Liao, D Liscbinski, Baoquan Chen, and Daniel Cohen-Or. 2019. Deep Video-Based Performance Cloning. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 219–233.

[2] Adrian. 2017. How Long Does It Take For You To Dance Kpop? https://www.hellokpop.com/korea/how-long-does-it-take-to-dance-kpop/. Accessed: 2020-02-29.

[3] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7297–7306.

[4] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. 2018. Synthesizing Images of Humans in Unseen Poses. *arXiv preprint arXiv:1804.07739* (2018).

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision*. Springer, 561–578.

[6] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2018. Everybody dance now. *arXiv preprint arXiv:1808.07371* (2018).

[7] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. 2018. Deep association learning for unsupervised video person re-identification. *arXiv preprint arXiv:1808.07301* (2018).

[8] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*. Springer, 628–644.

[9] Peng Dai, Xue Wang, and Weihang Zhang. 2018. Coarse-to-fine multiview 3d face reconstruction using multiple geometrical features. *Multimedia Tools and Applications* 77, 1 (2018), 939–966.

[10] Qiyang Hu, Adrian Waelchli, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. 2018. Video synthesis from a single image and motion stroke. *arXiv preprint arXiv:1812.01874* (2018).

[11] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2462–2470.

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

[13] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7122–7131.

[14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* (2014).

[15] Po-Han Lee, Jui-Wen Huang, and Huei-Yung Lin. 2012. 3D model reconstruction based on multiple view image capture. In *2012 International Symposium on Intelligent Signal Processing and Communications Systems*. IEEE, 58–63.

[16] Yining Li, Chen Huang, and Chen Change Loy. 2019. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3693–3702.

[17] Junbang Liang and Ming C Lin. 2019. Shape-Aware Human Pose and Shape Reconstruction Using Multi-View Images. In *Proceedings of the IEEE International Conference on Computer Vision*. 4352–4362.

[18] Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. 2019. Neural rendering and reenactment of human actor videos. *ACM Transactions on Graphics (TOG)* 38, 5 (2019), 1–14.

[19] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. 2019. Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*. 5904–5913.

[20] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. 2017. Pose guided person image generation. In *Advances in Neural Information Processing Systems*. 406–416.

[21] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. 2018. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 99–108.

[22] Natalia Neverova, Rıza Alp Güler, and Iasonas Kokkinos. 2018. Dense Pose Transfer. *arXiv preprint arXiv:1809.01995* (2018).

[23] Hyewon Seo, Young In Yeo, and Kwangyun Wohn. 2006. 3D body reconstruction from photos based on range scan. In *International Conference on Technologies for E-Learning and Digital Entertainment*. Springer, 849–860.

[24] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2377–2386.

[25] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. 2018. Deformable gans for pose-based human image generation. In *CVPR 2018-Computer Vision and Pattern Recognition*.

[26] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

[27] Shao-Hua Sun, Minyoung Huh, Yuan-Hong Liao, Ning Zhang, and Joseph J Lim. 2018. Multi-view to novel view: Synthesizing novel views with self-learned confidence. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 155–171.

[28] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. 2018. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1526–1535.

[29] Tony Tung, Shohei Nobuhara, and Takashi Matsuyama. 2009. Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 1709–1716.

[30] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2018. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence* (2018).

[31] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. Video-to-Video Synthesis. *arXiv preprint arXiv:1808.06601* (2018).

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

[33] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[34] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Vol. 2. Ieee, 1398–1402.

[35] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*. 802–810.

[36] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. 2017. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE international conference on computer vision*. 4733–4742.

[37] Zexiang Xu, Sai Bi, Kalyan Sunkavalli, Sunil Hadap, Hao Su, and Ravi Ramamoorthi. 2019. Deep view synthesis from sparse photometric images. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 76.

[38] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. 2018. Pose guided human video generation. *arXiv preprint arXiv:1807.11152* (2018).

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

[40] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara Berg. 2019. Dance dance generation: Motion transfer for internet videos. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 0–0.

[41] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. 2019. Progressive Pose Attention Transfer for Person Image Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2347–2356.