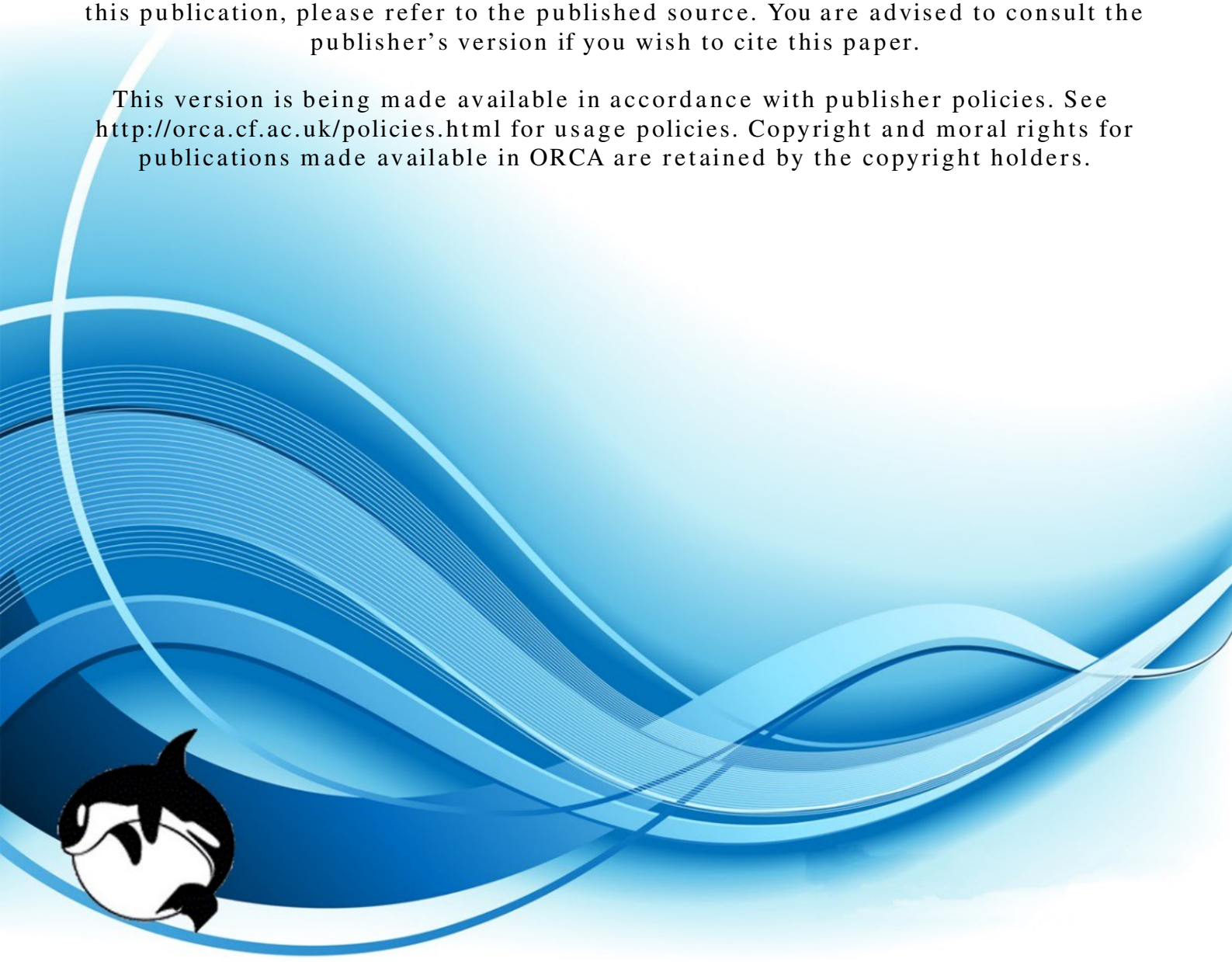


Human Motion Transfer with 3D Constraints and Detail Enhancement

Yang-Tian Sun, Qian-Cheng Fu, Yue-Ren Jiang, Zitao Liu, Yu-Kun Lai, Hongbo Fu, and Lin Gao*

Abstract—We propose a new method for realistic human motion transfer using a generative adversarial network (GAN), which generates a motion video of a target character imitating actions of a source character, while maintaining high authenticity of the generated results. We tackle the problem by decoupling and recombining the posture information and appearance information of both the source and target characters. The innovation of our approach lies in the use of the projection of a reconstructed 3D human model as the condition of GAN to better maintain the structural integrity of transfer results in different poses. We further introduce a detail enhancement net to enhance the details of transfer results by exploiting the details in real source frames. Extensive experiments show that our approach yields better results both qualitatively and quantitatively than the state-of-the-art methods.

Index Terms—Motion Transfer, Deep Learning, 3D Constraints, Detail Enhancement

1 Introduction

Video-based human motion transfer is an interesting but challenging research problem. Given two monocular video clips, one for a source subject and the other for a target subject, the goal of this problem is to synthesize a video by transferring the motion from the source person to the target, while maintaining the target person's appearance. Specifically, in the synthesized video, the subject should have the same motion as the source person, and the same appearance as the target person (including human clothes and background). Note this problem might also be seen as an appearance transfer problem if viewed from a single frame perspective. However, since from a holistic perspective, the goal is to transfer the motion from the source image domain to the target, we define this task as motion transfer. To achieve this, it is essential to produce high-quality image-to-image translation of individual frames, while ensuring temporal coherence.

The difficulty of this problem is how to effectively decouple and recombine the posture information and appearance information of the source and target characters. Based on generative adversarial networks (GANs), a powerful tool for high-quality image-to-image translation, Chan *et al.* [1] proposed to first learn a mapping from a 2D pose to a subject image from a target video, and then use the pose of a source subject as the input to the learned mapping for video synthesis. However, due to the difference between the source and target poses, this approach often results in noticeable artifacts, especially for the self-occlusion of body parts.

Observing that the self-occlusion issue is difficult to handle in the image domain, we propose to first reconstruct the 3D human model from the given monocular video for both the source and target subjects , and then adjust the pose of the target human model to match the source motion (while maintaining the target person's body shape). Intrinsic geometric description of the deformed target is then projected back to 2D to form an image that reflects the 3D structure and serves as condition for subsequent generation. We choose a few non-trivial eigenvectors of the Laplace matrix of the reconstructed mesh as the intrinsic geometry representation (dubbed "Laplacian feature" below), which is deformation-invariant and serves as a unique attribute of each vertex. Compared with other vertex-specific features, e.g., UV coordinates, the Laplacian feature is more suitable for the human generation task due to its spatially continuous embedding, implicitly encoded intrinsic geometry and higher feature dimensions. The boost of these characteristics to realistic human image generation is demonstrated in Section 4.2.2. The projection of Laplacian feature along with the 2D pose figure extracted from each source image is used as a constraint during GAN-based image-to-image translation, to effectively maintain the structural characteristics of human body under different poses.

In addition, previous methods [1], [2] only use the appearance of the target person in the training process of pose-to-image translation. When an input pose is very different from any poses seen during the training process, such solutions might lead to blurry results. Observing that the source video frame corresponding to the input pose might contain reusable rich details (especially for the body parts like hands where the source and target subjects share some similarity), we intend to selectively transfer geometry

* *Corresponding Author is Lin Gao (gaolin@ict.ac.cn).*

- *Y.-T. Sun, Y.-R. Jiang and L. Gao are with the Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, and also with the University of Chinese Academy of Sciences, Beijing, China. E-mail:{sunyangtian, gaolin}@ict.ac.cn, jiangyueren15@mails.ucas.ac.cn*
- *Q.-C. Fu is with the Department of Computer Science, Boston University E-mail:qcfu@bu.edu*
- *Z. Liu is with the TAL Education Group, Beijing, China and also with the Guangdong Institute of Smart Education, Jinan University, Guangzhou, China. E-mail:liuzitao@tal.com*
- *Y.-K. Lai is with the Visual Computing Group, School of Computer Science and Informatics, Cardiff University, Wales, UK. E-mail:LaiY4@cardiff.ac.uk*
- *H. Fu is with the School of Creative Media, City University of Hong Kong. E-mail: hongbofu@cityu.edu.hk*

Fig. 1. Given two monocular video clips, our method is able to transfer the motion of a source character (Top) to a target character (Middle), with realistic details (Bottom). We zoom in the enhanced details for better visualization.

details from real source frames to the synthesized video frames. This is achieved by our detail enhancement network. Figure 1 shows representative motion transfer results with rich details.

We summarize our contributions as follows: 1) We propose to use the Laplacian features of the reconstructed parametric human model as 3D constraints for the human motion transfer task. Such a novel representation serves as a more intrinsic representation compared with other 3D features, e.g., UV coordinates adopted in [3]. 2) We introduce the detail enhancement net (DE-Net), which utilizes the geometry information from real source frames to enhance details in generated results. Extensive experiments show that our method outperforms the state-of-the-art methods [1], [2], [3].

2 Related Work

Over the last decades, motion transfer has been extensively studied due to its ability for fast video content production. Some early solutions have mainly revolved around realigning existing video footage according to the similarity to the desired pose [4], [5]. However, it is not an easy task to find an accurate similarity measure for different actions of different subjects. Several other approaches have also attempted to address this problem in 3D, but they focus on the use of inverse kinematic solvers [6] and transfer motion between 3D *skeletons* [7], whereas we consider using a reconstructed 3D body mesh to guide motion transfer in the image domain in order to provide much richer constraints.

Recently, the rapid advances of deep learning, especially generative adversarial networks (GANs) and their variations (e.g., cGAN [8], CoGAN [9], CycleGAN [10], DiscoGAN [11]) have provided powerful tools for image-to-image translation, which has yielded impressive results across a wide spectrum of synthesis tasks and shown its ability to synthesize visually pleasing images from conditional labels. For example, pix2pix [12], based on a conditional GAN framework, is one of the pioneering works. CycleGAN [10] further presents the idea of cycle consistency loss for learning to translate between two domains in the absence of paired images, and Recycle-GAN [13] combines both spatial and temporal constraints for video retargeting tasks. Pix2pixHD [14] introduces a multi-scale conditional GAN to synthesize high-resolution images using both global and local generators, and vid2vid [2] designs specific spatial and temporal adversarial constraints for video synthesis.

Based on these variants of GANs, a lot of approaches [1], [15], [16], [17], [18] have been proposed for human motion transfer between two image domains. The key idea of these approaches is to decouple the pose information from an input image and use it as the input of a GAN network to generate a realistic image. For example, in [15], an input image is separated into two parts: the foreground (or different body parts) and the background, and the final realistic image is generated by separate processing and cross fusion of the two parts. Chan *et al.* [1] extract the pose information with an off-the-shelf human pose detector OpenPose [19] and use the pix2pixHD [14] framework together with a specialized face GAN to learn a mapping from a 2D pose figure to an image. Neverova *et al.* [20] adopt a similar idea but use the estimation of DensePose [21] to guide image generation. Wang *et al.* make a step further to adopt both OpenPose and DensePose in [2]. However, due to the lack of 3D semantic information , these approaches are highly sensitive to problems such as self-occlusions.

To solve the above problems, it is natural to add 3D information to the condition of generative networks. There are many robust 3D human mesh reconstruction methods (e.g., [22], [23], [24], [25]), which can reconstruct a 3D human model with its corresponding pose from a single image or a video clip. Benefiting from these accurate and reliable 3D body reconstruction techniques, we can study the issue of human motion transfer in a new perspective. Liu *et al.* [3]

present a novel warping strategy, which uses the projection of 3D models to tackle motion transfer, appearance transfer and novel view synthesis within a unified model. However, due to the generative characteristic of their network for multi-person, it does not perform particularly well for the specific person.

3 METHOD

We aim to generate a new video of a target person imitating the character movements in a source video, while keeping the structural integrity and detail features of the target subject as much as possible. To accomplish this, we use the mesh projection containing 3D information as the condition for the GAN, and introduce a detail enhancement mechanism to improve the details.

3.1 Overview

We tackle this problem with a coarse-to-fine strategy. At the first stage, we design a Motion Transfer Net (MT-Net) to get initial motion transfer results with the guidance of 3D information and a given appearance image. Note that adjacent pose labels are used for temporal consistence. Intuitively, the MT-Net can be trained within the target domain and the initial transfer results can be obtained by simply replacing the target subject's pose label with the source character's. However, such initial transfer results often exhibit blurring artifacts, especially in face and hands regions. To address this problem, we propose a Detail Enhancement Net (DE-Net) accompanied with a complicated but effective training pipeline based on the observation that although being different in clothes or genders, source and target characters usually have similar structure in faces and hands, where we often suffer from blurring artifacts. It is thus possible to use the information in the source frames to enhance the synthesized details. We first expound the main modules of our approach: pose label generation in Sec. 3.2, motion transfer net (MT-Net) in Sec. 3.3, and detail enhancement net (DE-Net) in Sec. 3.4, followed by the training strategy in Sec. 3.5 and the optimization objective in Sec. 3.6.

Notation. We denote $\mathcal{S} = \{S_i\}$ as a set of source video frames, and $\mathcal{T} = \{T_j\}$ as a set of target frames. For I_{sub}^{super}, the subscript denotes the attribute of the image, while the superscript denotes the domain it belongs to.

3.2 Pose Label Generation

Previous pose representations e.g. everybody [1] often take the form of keypoints or skeleton, and thus are difficult to deal with self-occlusion and fail to maintain the structure integrity. Therefore we propose to utilize the 3D geometry information of the underlying subject to produce label images as the GAN condition to regularize the generative network. The process of our 3D constraints generation is shown in Fig. 3.

3D Human Model Reconstruction. We first extract the 3D body shape β and pose θ information for both source and target videos using a state-of-the-art pre-trained 3D pose and shape estimator [25]. This leads to two 3D deformable mesh models (for the source and target videos), including the details of body, face, and fingers. When transferring between the two domains, the 3D human models allow easy generation of a 3D mesh with the pose from one domain and shape from the other. The extracted deformable mesh sequences might exhibit temporal incoherence artifacts due to inevitable reconstruction errors. We alleviate this issue by simply applying temporal smoothing to individual mesh vertices, since our mesh sequences have the same connectivity.

Human Model Projection. We project the reconstructed 3D human model onto 2D to obtain a label image, which will be used as the condition to guide the generator. The image should ideally contain intrinsic 3D information (invariant to pose changes) to guide the synthesis process such that such information can be color-encoded: a particular color corresponding to a specific location on the human body. To achieve this, we propose to extract the three non-trivial eigenvectors corresponding to the three smallest eigenvalues of the Laplace matrix of the reconstructed mesh and consider them as a 3-channel color assigned to each vertex [26]. The colored mesh is projected to 2D to form a 3D constraint image.

Note that although additional 3D information is available, 3D meshes extracted from 2D images may occasionally contain artifacts due to the inherent ambiguity. Therefore, we also adopt OpenPose [19] to extract a 2D pose figure as part of the condition, which is less informative but more robust in the 2D space. Our label image therefore is 6-channel after combining both 2D and 3D constraints actually. We will evaluate the impact of these two conditions in the ablation study in Sec 4.2.2.

3.3 Motion Transfer Net

We learn the mapping from temporally adjacent label images and an appearance image to a realistic image by training the MT-Net, consisting of an Appearance Encoder and a Pose GAN (Figure 4). Specifically, for $\mathcal{X} \in \{\mathcal{S}, \mathcal{T}\}$ let $I_{app}^{\mathcal{X}}$ and $I_{pose}^{\mathcal{X}}$ be the corresponding appearance image and the adjacent pose labels in the domain $\mathcal{X}$ (extracted from the current frame and the last frame), respectively, and we can obtain the reconstructed frame $I_{MT}^{\mathcal{X}}$. The Appearance Encoder encodes the appearance image $I_{app}^{\mathcal{X}}$ to a latent feature, and the Pose GAN produces an output image $I_{MT}^{\mathcal{X}}$ with the given appearance and pose/shape constraints. It is worth mentioning that in order to solve the problem of poor temporal continuity caused by single frame generation, similar to [1], we involve adjacent frames in $I_{pose}^{\mathcal{X}}$ to improve temporal coherence.

Appearance Encoder E_{app}**.** The Appearance Encoder is a fully convolutional network that extracts appearance features from an input image $I_{app}^{\mathcal{X}}$, and the extracted feature is used as a condition for the Pose GAN. It takes randomly selected frames as input and outputs appearance features corresponding to that domain.

Pose GAN Generator G_{pose}**.** The Pose GAN is the main part of MT-Net, which consists of three sub-modules: Downsampling, ResNet blocks, and Upsampling. It works on both label images and the appearance features extracted by the Appearance Encoder, and produces synthesized results with the corresponding pose and appearance. As shown in Fig. 4, the output of the Appearance Encoder is concatenated to

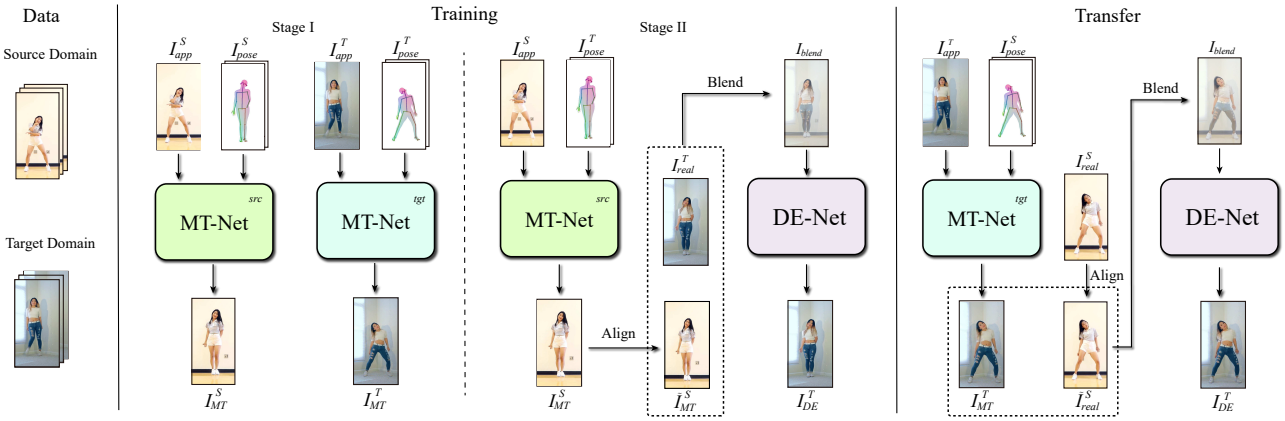


Fig. 2. The architecture of our pipeline. *Training:* At Stage I, we train MT-Net(src) and MT-Net(tgt) separately within the corresponding domains. The MT-Net transfers the motion from temporally adjacent pose labels to the image domain represented by the appearance image. At Stage II, we send a source appearance image $I^{\mathcal{S}}_{app}$ and adjacent target pose labels $I^{\mathcal{T}}_{pose}$ to MT-Net(src) to obtain a coarse source image $I^{\mathcal{S}}_{MT}$. After aligned to the target character's position and size, it is converted to the blending domain by blended with a corresponding real target frame. The DE-Net learns to translate the blended image to the target domain. *Transfer:* Our system obtains a coarse transfer result $I^{\mathcal{T}}_{MT}$ with target appearance and the source pose labels by MT-Net(tgt), and converts $I^{\mathcal{T}}_{MT}$ to the blending domain by blending it with the aligned real source frame. Finally the DE-Net translates the blended image I_{blend} to the target domain with details enhanced.

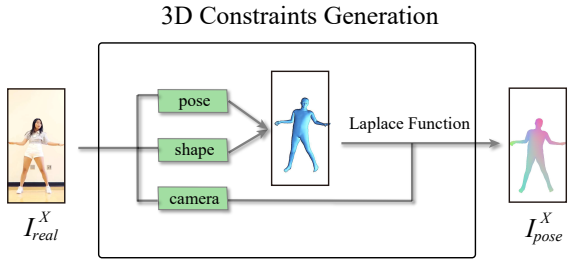


Fig. 3. Illustration of the 3D constraints generation process. We reconstruct a 3D mesh of a source or target subject image I^{X}_{real}, associate the three smallest eigenvalues of its Laplace matrix as intrinsic features (visualized in RGB color) with the mesh vertices, and project the colored mesh to form a 3D constraint image, denoted as I^{X}_{pose}.

the pose features and the concatenated features are used as input to the ResNet blocks.

Pose GAN Discriminator $D^{\mathcal{X}}_{MT}$. We use the multi-scale discriminator presented in pix2pixHD [14], since discriminators of different scales can give the discrimination of images at different levels. In our method, we use two discriminators $D^{\mathcal{S}}_{MT}$ and $D^{\mathcal{T}}_{MT}$ to predict the probability of generated images belonging to the corresponding domains, each with 3 scales.

The generator and discriminator networks drive each other: the generator learns to synthesize more realistic images conditioned on the input to fool the discriminator, while the discriminator in turn learns to discern the "real" images (ground truth) and "fake" (generated) images.

Note that MT-Net is not a generalized network that can be directly applied to any video character generation. Similar to pix2pixHD [14], each trained MT-Net is bound to a specific video character. We condition the network on the additional appearance image since the common appearance information between frames helps the network converge faster and become more stable, as we demonstrate in the supplementary material.

As previously mentioned, in order to meet the need of the subsequent detail enhancement, we train the MT-Net in both the source domain and the target domain, denoted as MT-Net^{src} and MT-Net^{tgt}, respectively. Note that these two networks share the same weights of the downsampling and ResNet parts, while having separate weights for the upsamplling part.

Temporal Smoothing. We use the time smoothing strategy in [1] to enhance the temporal continuity between adjacent generated frames. The generation of the current frame is not only related to the current label image I_{label}, but also related to the previous frame I'_{label}.

Therefore, our conditional GAN has the following objective:

$$\min_{\substack{E_{app},G_{pose},\\ D^{\mathcal{X}}_{MT}}} \mathcal{L}^{\mathcal{X}}_{\text{MT}}(I^{\mathcal{X}}_{pose}, I'^{\mathcal{X}}_{pose}, I^{\mathcal{X}}_{real}, I^{\mathcal{X}}_{MT}) = \mathbb{E}[\log D^{\mathcal{X}}_{MT}(I^{\mathcal{X}}_{pose}, I'^{\mathcal{X}}_{pose}, I^{\mathcal{X}}_{real})] + \mathbb{E}[\log(1 - D^{\mathcal{X}}_{MT}(I^{\mathcal{X}}_{pose}, I'^{\mathcal{X}}_{pose}, I^{\mathcal{X}}_{MT})]. \quad (1)$$

Here the discriminator $D^{\mathcal{X}}_{MT}$ takes the pose labels and images in the domain $\mathcal{X}$ as input, and classifies them to real images (from the training set), or fake images (I_{MT} generated by the Pose GAN).

3.4 Detail Enhancement Net

Through the first stage of training, we can obtain initial transfer results which, however, suffer from blurring artifacts. Here we design the DE-Net for the recovery of blurred details. Specifically, we aim to enhance details in the initial source-to-target transfer results $I^{\mathcal{T}}_{MT}$ with the help of the information from source frames. However, it is intractable for a neural network to extract desired details from the source domain and transfer them to the target domain. We address this problem based on the following observations: 1) blurred regions concentrate in the hands and face regions of initial transfer results for both source-to-target and target-to-source; 2) hands and faces between different characters always have similar patterns, hence the blending of blurred regions with real frames can introduce reasonable structure information 3) the blending of source

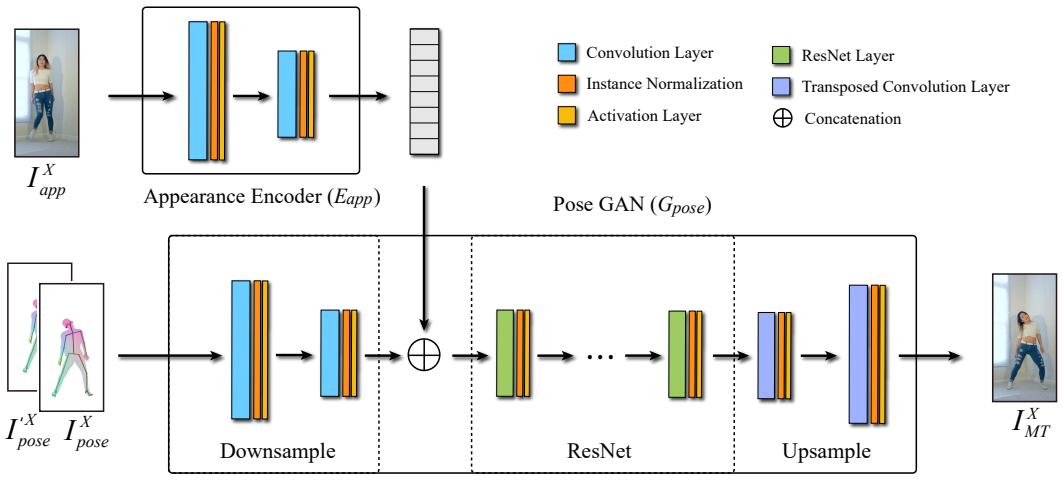


Fig. 4. Architecture of the generative network. Appearance image $I^{\mathcal{X}}_{app}$ is randomly selected and sent to the Appearance Encoder (denoted as E_{app}) to obtain appearance features. Adjacent pose label images including $I^{\mathcal{X}}_{pose}$ (for the current frame) and $I'^{\mathcal{X}}_{pose}$ (for the previous frame) are sent to Pose GAN (denoted as G_{pose}) together with the appearance features to generate a reconstructed image $I^{\mathcal{X}}_{MT}$.

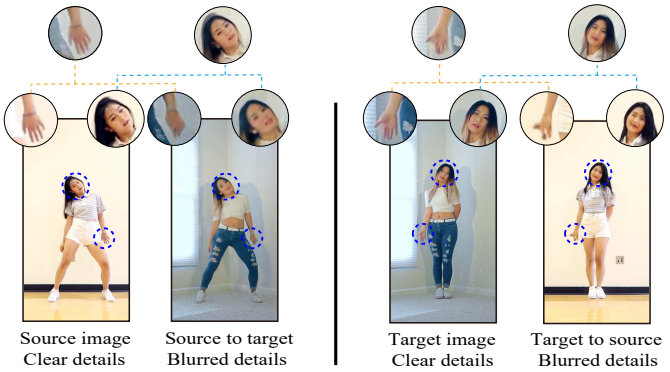


Fig. 5. Comparison of a source image $I^{\mathcal{S}}_{real}$ and a source-to-target transfer result $I^{\mathcal{T}}_{MT}$, as well as a target image $I^{\mathcal{T}}_{real}$ and a target-to-source transfer result $I^{\mathcal{S}}_{MT}$. The upper circles show the blending results in hands and face regions. The blending results are of a similar style for the two opposite blending modes.

frames with blurred source-to-target transfer results has a similar style to the blending of target frames with blurred target-to-source transfer results, as illustrated in Fig. 5. Note that there are real target frames with the same poses as the target-to-source transfer results. This motivates us to train the DE-Net to learn the transfer from the blending domain to the target domain with supervision. The learned mapping can be used to enhance details for the source-to-target transfer results in the transfer stage.

Alignment and Blending. Note that in different videos, subjects might have different builds or positions relative to the camera. In such cases, before blending we need to align the source frame ($I^{\mathcal{S}}_{MT}$ or $I^{\mathcal{S}}_{real}$) in accordance with the target by applying the transformation calculated from the reconstructed meshes with the source and target parameters (i.e., pose, body shape, and position). As we will show in Figure 12, this step is effective in improving the details and avoiding artifacts. More details are included in the supplementary materials. We blend the images from the source and target domains linearly with a learnable parameter θ. Note that θ is updated during the training process and fixed in the inference(transfer) stage.

Blending Domain to Target Domain. The purpose of our DE-Net is to transfer the image from the blending domain to the target domain, while preserving useful details and eliminating redundant information from source frames. It is based on the GAN framework, where the generator G_{DE} is a U-net for synthesizing images in the target domain with clear details, as illustrated in Figure 6. The discriminator D_{DE} discerns the "real" images (ground truth) and "fake" images (synthesized by G_{DE}). In the training stage, we use the blending of image pair $(I^{\mathcal{S}}_{MT}, I^{\mathcal{T}}_{real})$ as input and train DE-Net in a supervised manner with $I^{\mathcal{T}}_{real}$ as ground truth. The use of blended images instead of concatenation avoids the output overfitting to $I^{\mathcal{T}}_{real}$. We have verified the advantage of our blending manner in the supplementary. We optimize the DE-Net by minimizing the following objective:

$$\min_{G_{DE}, D_{DE}} \mathcal{L}_{\text{DE}}(I^{\mathcal{T}}_{pose}, I^{\mathcal{T}}_{real}, I^{\mathcal{T}}_{DE}) = \mathbb{E}[\log D_{DE}(I^{\mathcal{T}}_{pose}, I^{\mathcal{T}}_{real})] + \mathbb{E}[1 - \log D_{DE}(I^{\mathcal{T}}_{pose}, I^{\mathcal{T}}_{DE})], \quad (2)$$

where $I^{\mathcal{T}}_{pose}$ is the corresponding pose label.

In the transfer stage, we use the source-to-target transfer result $I^{\mathcal{T}}_{MT}$ and the corresponding source image $I^{\mathcal{S}}_{real}$ to obtain enhanced transfer results.

3.5 Data Flow in Different Phases

We re-interpret the overall data flow in this subsection, including the training and transfer phases, as illustrated in Fig. 2. The training part can be divided into two stages, i.e., the within-domain pre-training of MT-Net and the training of DE-Net.

Within-Domain Pre-Training of MT-Net. To stabilize the training process, we first pre-train MT-Net using within-domain samples for both the source and target domains separately, namely the training of MT-Net^{src} and MT-Net^{tgt}. We use the superscript to distinguish them since MT-Net contains different parameters for different characters. To be specific, in the $\mathcal{X}$ domain the MT-Net$^{\mathcal{X}}$ is trained to map the appearance image $I^{\mathcal{X}}_{app}$ and adjacent pose labels $I^{\mathcal{X}}_{pose}$ to the reconstructed source frame. Note that for each training $I^{\mathcal{X}}_{app}$

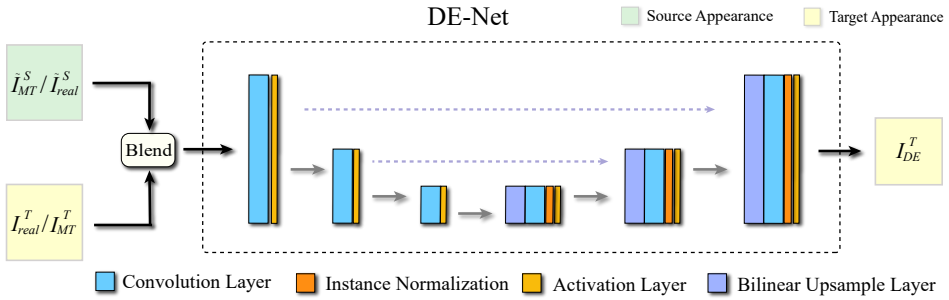


Fig. 6. Architecture of Detail Enhancement Net (DE-Net). The main part of our DE-Net is a U-net, which takes the linear blending of paired data $(\tilde{I}^{\mathcal{S}}_{MT}, I^{\mathcal{T}}_{real})$ or $(\tilde{I}^{\mathcal{S}}_{real}, I^{\mathcal{T}}_{MT})$ as input, and synthesizes a target image $I^{\mathcal{T}}_{DE}$ with details enhanced.

is randomly selected from video frames and fixed during the training process. We have demonstrated the effect of appearance image selection in the supplementary material.

Training of DE-Net. Essentially, DE-Net is trained to learn the mapping from the blending domain to the target domain for detail extraction and preservation with the real target character frames as supervision. As illustrated in Stage II of Fig. 2, in order to train DE-Net we generate an initially transferred image $I^{\mathcal{S}}_{MT}$ of *the source character driven by a given target pose* $I^{\mathcal{T}}_{pose}$. This process seems to running counter to the final objective (i.e., generating a target image from source character pose $I^{\mathcal{S}}_{pose}$), however, it sets the stage for the supervised training of DE-Net with $I^{\mathcal{T}}_{real}$ as the ground truth.

Transfer. Our *transfer (or inference) pipeline* is straightforward. By sending $I^{\mathcal{T}}_{app}$ and $I^{\mathcal{S}}_{pose}$ to MT-Net, we can obtain the initial transfer result $I^{\mathcal{T}}_{MT}$. DE-Net takes the blending of $I^{\mathcal{T}}_{MT}$ and its corresponding source frame as input, and outputs the final result with details enhanced.

3.6 Full Objective

The training of our network is divided into two stages. First we train the Motion Transfer Net. The full objective has the following form, containing adversarial loss $\mathcal{L}^{\mathcal{X}}_{\text{MT}}$, perceptual loss $\mathcal{L}_P$, and discriminator feature-matching loss $\mathcal{L}_{FM}$:

$$\begin{aligned}\min_{E_{app},G_{pose}} \sum_{\mathcal{X}\in\mathcal{S},\mathcal{T}} ((\max_{D^{\mathcal{X}}_{MT}} \mathcal{L}^{\mathcal{X}}_{\text{MT}}(I^{\mathcal{X}}_{pose}, I^{'\mathcal{X}}_{pose}, I^{\mathcal{X}}_{real}, I^{\mathcal{X}}_{MT})) \\ + \lambda_P \mathcal{L}_P(I^{\mathcal{X}}_{MT}, I^{\mathcal{X}}_{real}) \\ + \lambda_{FM}\mathcal{L}_{FM}(I^{\mathcal{X}}_{pose}, I^{'\mathcal{X}}_{pose}, I^{\mathcal{X}}_{real}, I^{\mathcal{X}}_{MT}). \end{aligned} \quad (3)$$

Here, $\mathcal{L}^{\mathcal{X}}_{\text{MT}}$ is defined in Eq. 1. The perceptual loss $\mathcal{L}_P$ regularizes a generated result $I^{\mathcal{X}}_{MT}$ to be close to the ground truth $I^{\mathcal{X}}_{real}$ in the VGG-19 [27] feature space, defined as

$$\mathcal{L}_P(I_1, I_2) = ||\text{VGG}(I_1) - \text{VGG}(I_2)||_1. \quad (4)$$

The discriminator feature-matching loss $\mathcal{L}_{FM}$ presented in pix2pixHD [14] similarly regularizes the output using intermediate results of the discriminator, and is calculated as

$$\begin{aligned}\mathcal{L}_{FM}(I^{\mathcal{X}}_{pose}, I^{'\mathcal{X}}_{pose}, I^{\mathcal{X}}_{real}, I^{\mathcal{X}}_{MT}) = \\ \mathbb{E}\sum_{i=1}^{T}\frac{1}{N_i}[||D^{(i)}_k(s,x) - D^{(i)}_k(s, G(s))||_1], \end{aligned} \quad (5)$$

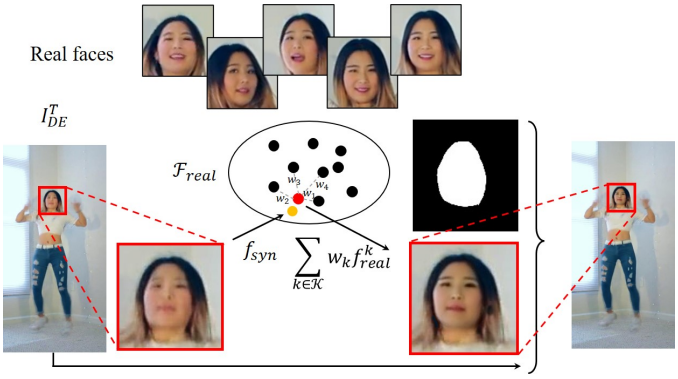


Fig. 7. The illustration of post-processing facial enhancement. The cropped facial part f_{syn} of the synthesized frame $I^{\mathcal{T}}_{DE}$ is embedded into the StyleGAN3 latent space together with real faces $\mathcal{F}_{real} = \{f^i_{real}\}$. Through a retrieval-and-interpolation strategy, the latent embedding is refined to get an enhanced face, which is finally fused to the synthesized frame with a predicted facial mask.

where D denotes the discriminator $D^{\mathcal{X}}_{MT}$, T is the number of layers, N_i is the number of elements in the i-th layer, and k is the index of discriminators in the multi-scale architecture. s is the condition of cGAN and x is the corresponding ground truth.

The DE-Net is optimized with the following objective

$$\begin{aligned}\min_{G_{DE}}((\max_{D_{DE}} \mathcal{L}_{\text{DE}}(I^{\mathcal{T}}_{pose}, I^{\mathcal{T}}_{real}, I^{\mathcal{T}}_{DE})) + \lambda_P \mathcal{L}_P(I^{\mathcal{T}}_{DE}, I^{\mathcal{T}}_{real}) + \\ \lambda_{FM}\mathcal{L}_{FM}(I^{\mathcal{X}}_{pose}, I^{'\mathcal{X}}_{pose}, I^{\mathcal{X}}_{real}, I^{\mathcal{X}}_{MT})) + \lambda_{reg}||\theta - 0.5||_2. \end{aligned} \quad (6)$$

Here $\mathcal{L}_{\text{DE}}$ is defined in Eq. 2. The perceptual and discriminator feature-matching losses are defined in Eqs. 4 and 5, respectively. The last term regularizes the blending coefficient to avoid the overfitting to target frames.

3.7 Facial Enhancement

To further improve the quality of synthesized human frames, we add a post-processing step to enhance facial details. We assume that the StyleGAN3 [28] latent codes of the same person's facial images constitute a low-dimensional, locally-linear manifold. Based on this assumption, the synthesized facial part can be enhanced with a retrieval-and-interpolation approach by projecting it into the StyleGAN3 latent space, as similarly done in DeepFaceDrawing [29]. Fig. 7 illustrates this idea. Specifically, the facial images with different expressions and angles are first cropped from real frames, and then encoded to the StyleGAN3 latent space with e4e [30]. These latent features are denoted as $\mathcal{F}_{real} = \{f^i_{real}\}$, where the superscript i indicates the ith

feature. Given a synthesized frame of the same character, the facial part can be extracted and embedded as f_{syn} with the same approach. We aim to project f_{syn} to the manifold constructed by $\mathcal{F}_{real}$. We first find the K-nearest neighbors of f_{syn} in $\mathcal{F}_{real}$ under the Euclidean space, whose indexes form a set $\mathcal{K}$. With the locally linear assumption, f_{syn} can be projected as a linear combination of these neighbors with coefficients w_k $(k \in \mathcal{K})$ by minimizing the reconstruction error. This can be formulated as the following optimization problem

$$\min ||f_{syn} - \sum_{k \in \mathcal{K}} w_k f^k_{real}|| \qquad s.t. \sum_{k \in \mathcal{K}} w_k = 1. \tag{7}$$

The coefficients can be determined by solving a constrained least-squares problem. Accordingly, f_{syn} can be refined as $\sum_{k \in \mathcal{K}} w_k f^k_{real}$, which is then fed into StyleGAN3 for more detailed facial part generation. Finally the enhanced facial image is fused to the synthesized frame ($I^{\mathcal{T}}_{DE}$) according to the detected facial mask. We show the effect of facial enhancement in Fig. 15.

4 Experiments

We compare our method with state-of-the-art methods and ablation variants, both quantitatively and qualitatively.

4.1 Setup

Datasets. To evaluate the performance of our method, we collected three types of data: the dataset published by [1], a few single-dancer videos from YouTube and Bilibili, and 5 videos filmed by ourselves (2 with ordinary background and 3 with green screen). Each subject wears different clothes and performs different types of action such as freestyle dancing and stretching exercises. We obtained these subjects' consent to use their video data for our experiments. We cut out the start and end parts that contain no action, crop and normalize each frame to 1024×512 resolution by simple scaling and translation. Note that each video in our dataset can serve as either a source or target video.

Implementation Details. We adopt a multi-stage training strategy in our method using Adam optimizer with the learning rate of 0.0001. In the first stage, we pre-train the MT-Net for 20 epochs. In the next stage, the parameters of MT-Net are fixed and the DE-Net is trained individually for 10 epochs. We set hyper-parameters $\lambda_{FM} = 10$ and $\lambda_P = 5$ for both stages and $\lambda_{reg} = 10$ for the second stage. More details about MT-Net and DE-Net training are given in the supplementary materials.

Existing Methods. We compare our method with state-of-the-art methods *vid2vid* [2], *Everybody Dance Now* [1] and *Liquid Warping GAN* [3], using their official implementation.

4.2 Quantitative Results

Evaluation Metrics. We use objective metrics for quantitative evaluation under two different conditions: 1) To directly measure the quality of generated images, we perform self-transfer, in which the source and target are the same subject, and then use SSIM [31] and learning-based perceptual similarity (LPIPS) [32] to assess the similarity between source and target images. We split frames of each subject into a training set and a validation set at the ratio of 8:2 for this evaluation. 2) We also evaluate the performance of cross-subject transfer, where the source and target are different subjects, using inception score [33] and Fréchet Inception Distance (FID) [34] as metrics. It should be noted that we compute the FID score between the original and generated target images since there exists no ground truth for comparison in this case. We also fine-tune the inception module [35] with the Market-1501 dataset [36] on the Person-ReID task, which serves as a more accurate feature extractor for synthesized human images due to the reduced domain gap, as demonstrated in [37]. We denote this metric as "**IS-ReID**". We exclude the green screen dataset in quantitative evaluation to focus on more challenging cases.

TABLE 1
Quantitative comparisons with the state-of-the-art methods on the dance dataset. It can be seen that our method outperforms SOTA in both self-transfer and corss-transfer condition.

Metric		Method			
		vid2vid	Chan *et al.*	LW-GAN	ours
Self-trans	SSIM ↑	0.781	0.836	0.790	**0.891**
	LPIPS ↓	0.096	0.067	0.106	**0.039**
Cross-trans	IS-ReID ↑	3.568	3.794	3.274	**4.015**
	FID ↓	59.98	57.02	81.20	**51.26**

The metrics mentioned above are all based on single frames, which cannot reflect the smoothness of generated image sequences. The effect of mesh filtering in time series can be observed in the video results and quantitatively measured by the user study in Sec 4.2.3.

4.2.1 Comparison with Existing Methods.

Comparison results with the state-of-the-art methods are reported in Table 1. It can be found that our method performs better than others, in both self-transfer and cross-subject transfer.

4.2.2 Ablation Study

We perform an ablation study to verify the impact of each key component of our model, including using 3D constraints ("**3D**") and DE-Net ("**DE**"). Our full pipeline is indicated as "Full". Note that "**Full (ave)**" means blending images from the source and target domain in an average way.

Table 2 shows the results of the ablation study. It is obvious that our full proposed framework performs better than its variants. Both 3D constraints and DE-Net are able to enhance the results. Although there is no explicit 3D loss, the Laplace projection of 3D meshes effectively defines the shape and geometry information and serves as a condition of Pose GAN. The 3D constraints plays an important role in generation, as shown in **MT (3D)** and **MT (2D)**. The scores of **MT(2D+3D)** show the complementarity of 2D and 3D conditions on this task. The comparison between **Full** and **MT(2D+3D)** (or between **MT(2D)+DE** and **MT(2D)**) proves the positive effect of DE-Net. The difference of the blending approaches is reflected in the "**Full (ave)**" and "**Full**", showing the advantage of the learned blending parameter.

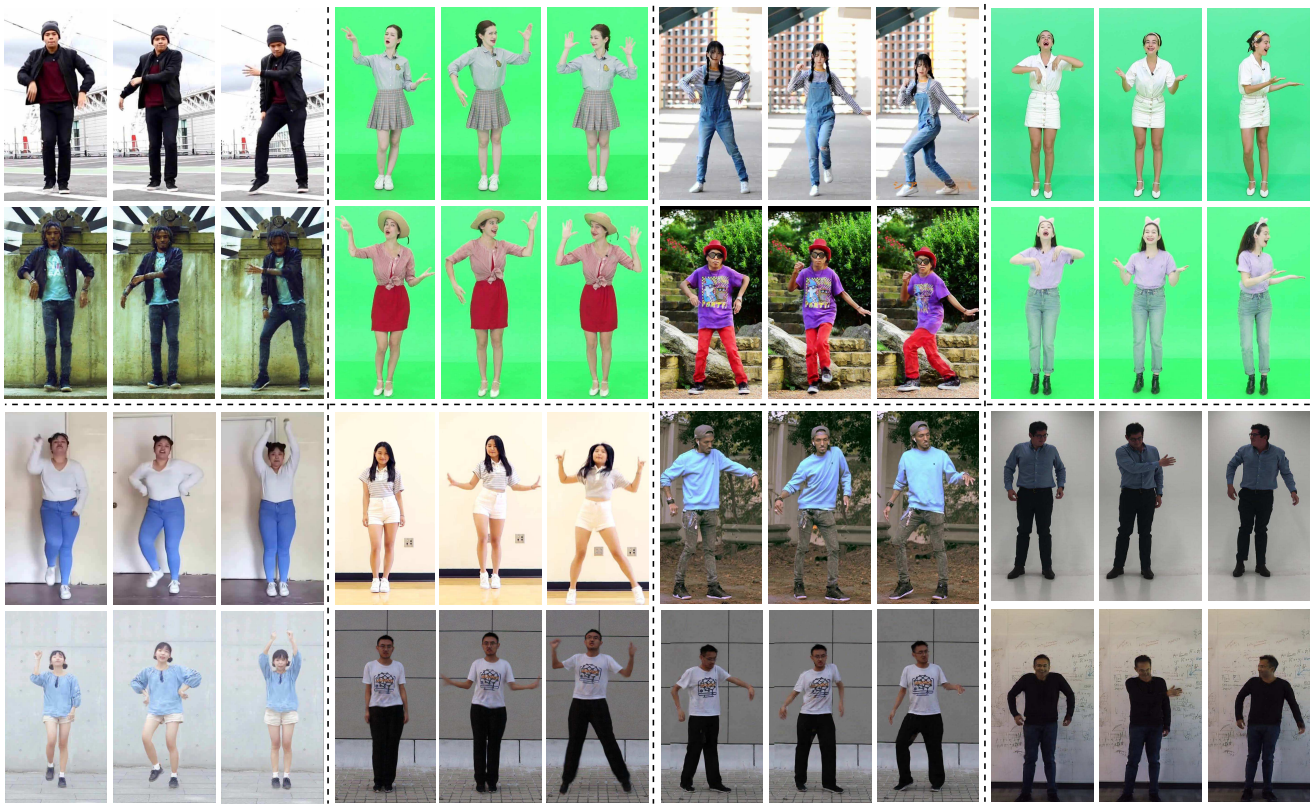


Fig. 8. Motion transfer results. We show the generated frames of several subjects with different genders, races, and builds. In each group, the top row shows the source subject and the bottom row shows the generated target subject. Please refer to the supplemental materials for synthesized videos.

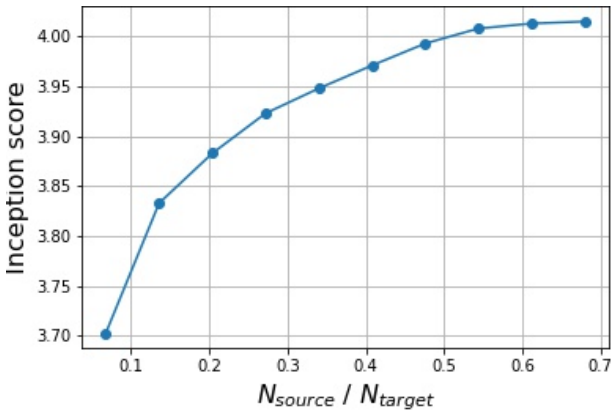


Fig. 9. The effect of the number of source frames on transfer results. We show the inception score of transfer results (y-axis) with respect to the ratio of source frame number to target (x-axis). For the inception score, the higher is the better.

Furthermore, we can observe that the scores of self-transfer between **MT(2D)** and **MT(2D+3D)** (or between **MT(2D)+DE** and **Full**) are similar. This is mainly because source and target subjects share the same body shape in self-transfer, thus somewhat limiting the effectiveness (and necessity) of 3D constraints. In contrast, the scores of cross-subject transfer indicate the important role of the 3D information on transfer between different subjects with different shapes.

Note that Liu *et al.* [3] also use the projection of a 3D model to guide the generation of GAN. However, compared with the correspondence map colored by UV coordinates, our Laplacian features provide a more intrinsic and continuous representation for 3D models, which is more suitable for the human image generation task. Some other approaches, e.g., SMPLpix [38], utilize the 3D geometry information by conditioning the depth map of human mesh vertices, which lacks structure information and is not suitable for the generative model. The results in Table 4 and Fig. 14 show that our approach leads to better scores and results.

To demonstrate the effect of post-processing facial enhancement, we illustrate a pair of results with and without facial enhancement, denoted as **w/ FE** and **w/o FE**, respectively. Moreover, we also evaluate a baseline method, denoted as **Closest FE**, which is based on direct retrieval (i.e., enhancing the facial image by searching for the closest facial embedding in the latent space), instead of our retrieval-and-interpolation approach. The comparisons are illustrated in Fig. 15.

4.2.3 User Evaluation

We also conducted a user study to measure the human perceptual quality for cross-subject transfer results. In our experiments, we compared videos generated by vid2vid, the method of Chan *et al.*, Liquid Warping GAN, and our method. Specifically, we showed to volunteers a series of videos by each of the methods at the resolution of 1024×512. We invited 50 participants for this study. Without any time limit, each of them was asked to select among results from the four approaches: 1) the clearest result with rich details; 2) the temporally most stable result; and 3) the overall best result. As shown in Table 3, our method leads

TABLE 2
Ablation study. Our approach achieves the best scores among the compared variants.

Metric		Method					
		MT(2D)	MT(3D)	MT(2D+3D)	MT(2D)+DE	Full (ave)	Full
Self-trans	SSIM ↑	0.828	0.831	0.856	0.877	0.887	**0.891**
	LPIPS ↓	0.064	0.063	0.058	0.043	0.041	**0.039**
Cross-trans	IS-ReID ↑	3.523	3.632	3.731	3.587	3.954	**4.015**
	FID ↓	58.62	56.68	55.77	57.56	53.76	**51.26**

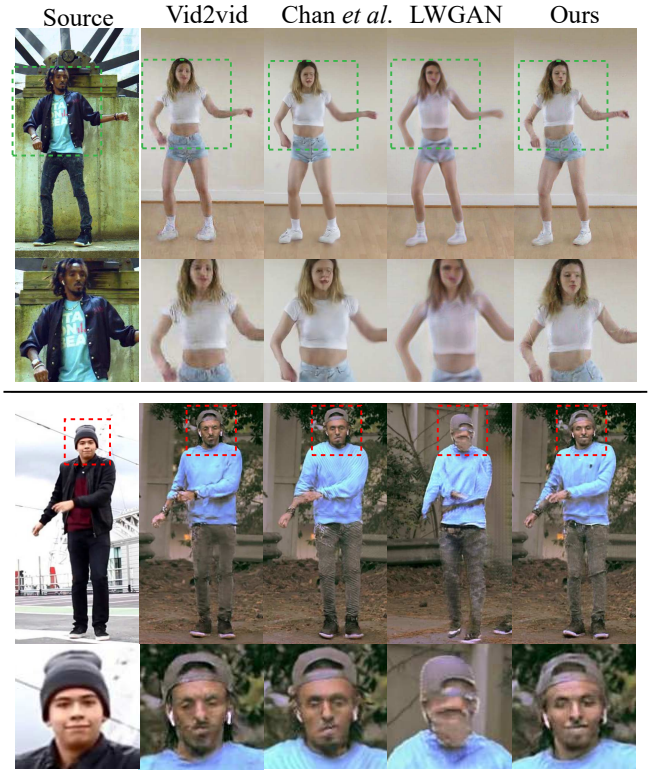


Fig. 10. Comparisons with the state-of-the-art methods. We show the generated results by vid2vid, Everybody Dance Now by Chan et al., Liquid Warping GAN (LW-GAN), and our method. Our method produces richer and more realistic details.

to more realistic results, with richer details and with better temporal stability in comparison with other methods.

TABLE 3
User study results. We report the percentage of the participants' choice as the best result among the results of 4 methods in three different aspects, respectively.

Quality	Method			
	vid2vid	chan *et al.*	LW-GAN	ours
Detail and clarity	22.2%	16.7%	7.07%	**54.0%**
Temporal stability	24.7%	15.1%	9.60%	**50.5%**
Overall feeling	26.3%	17.7%	8.08%	**48.0%**

TABLE 4
Ablation study for pose labels. Our 3D condition achieves higher IS and lower FID scores, indicating better results.

Metric	Pose labels in [3]	Ours
IS-ReID ↑	3.612	**4.015**
FID ↓	58.42	**51.26**

4.2.4 *Effect of Number of Source Frames on Results*

The previous motion transfer methods such as [1] and [2] only use target frames at the training stage, and the quality of their generative models is thus not directly related to the number of source frames. While in our method, source and target frames are both involved in training. Therefore, it is meaningful to explore the influence of source frame numbers on the quality of generated results. We carry out experiments on all subjects and record the average evaluation of generated image quality with respect to the ratio of source frame number to target when the number of target frames is fixed, as shown in Fig. 9. We choose the inception score as the metric. It can be seen that the loss has converged when the ratio is around 0.5, which serves as a guidance for the data preparation.

4.3 Qualitative Results

We visualize our generated results in Figure 8. It can be seen that our method successfully drives the motions of different targets with structural integrity and rich details, particularly in the face and hands. We also demonstrate that our method outperforms the existing methods in Figures 10 and 11.

As illustrated in the first row of Figure 10, our method can maintain the structural integrity of human parts, and avoid the mixture of arms and hands, which is often one of the main artifacts with the other methods. At the same time, our method can also characterize the details of the generated results more accurately, such as facial expression, as shown in the second row of Figure 10. Figure 12 demonstrates the effectiveness of alignment transformation for DE-Net: it effectively aligns the source to the target, and thus improves the clarity and avoids artifacts in generated results. Figure 13 shows the advantage of using 3D constraints and DE-Net.

4.4 Limitations and Ethical Discussions

Although our model is able to synthesize motion transferred images with high authenticity and details, there are still several limitations. First, both the MT-Net and DE-Net are person-specific generation models. Therefore, their generalization ability is limited and they have to be trained from scratch for any new source-target character pair. Second, the abnormal movement of source characters might cause large changes in the human body shape, e.g., perturbations in hair or clothing, thus making the DE-Net fail to eliminate these suddenly emerging human parts. We show some failure cases with visual artifacts in Figure 16. In the left example, our model fails to eliminate the long hair of the source character in the result, while in the right, some undesired part of clothes appears in the generated images because of the loosely dressed source subject.

The intentions of our work are undoubtedly towards a comprehensive and positive perspective. However, potential harms might still be brought about if our technique is maliciously used. To reduce the impact of misuse (specifically, deep-fakes), previous works [39] have proposed ways for detecting fake videos. We reasonably expect that these

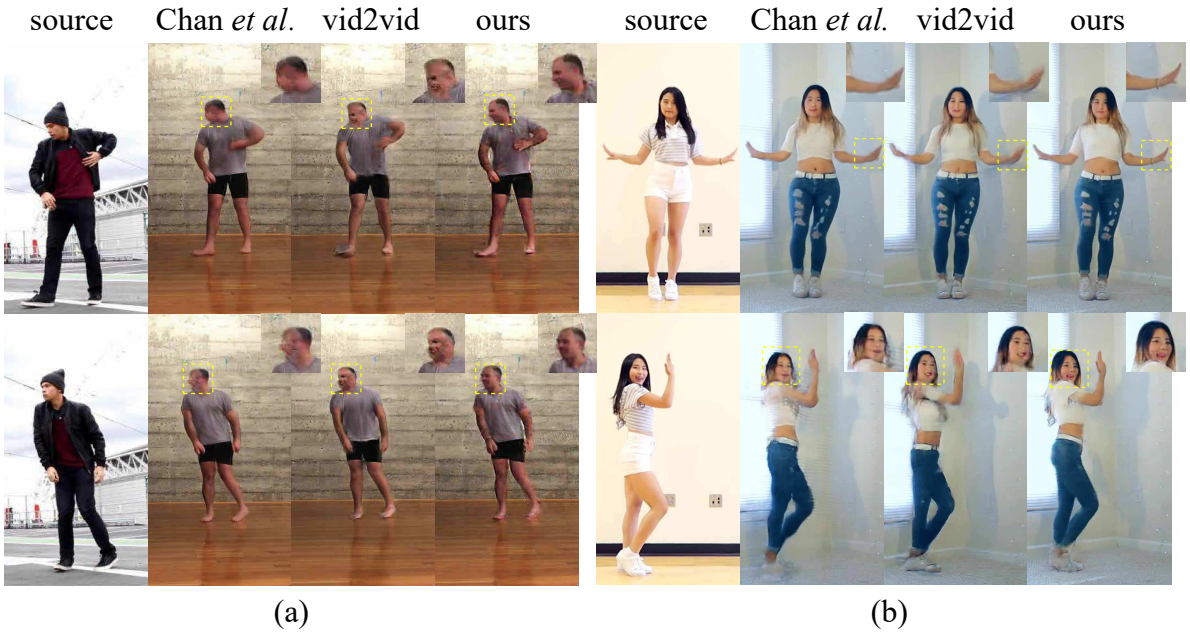


Fig. 11. We compare our approach with the method by Chan *et al.* [1] and vid2vid [2] on the data published by [1] (a) and the data used in [2] (b) for the sake of fairness. Our method produces more realistic results for occluded actions such as side faces and bending fingers.

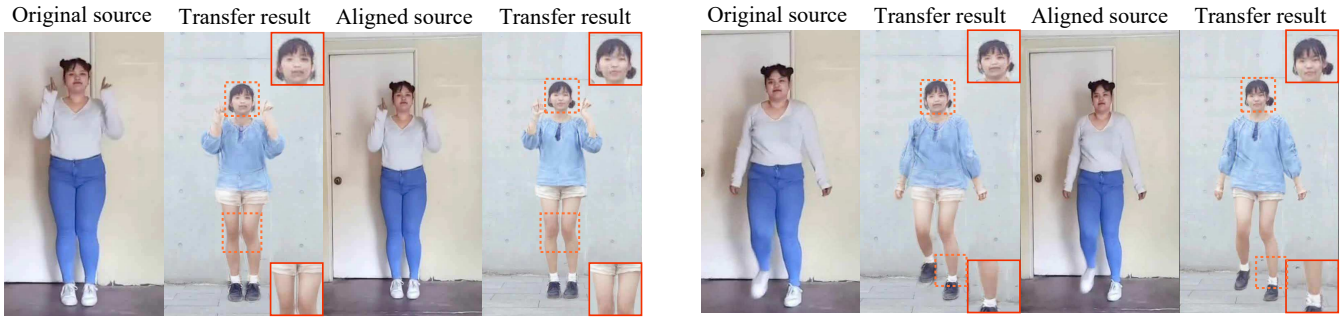


Fig. 12. Effects of aligning transformation for DE-Net. It can be seen that the absence of transformation will make DE-Net unable to match source and target characters accurately, resulting in unclear results with artifacts and body shape changes.

works could discriminate the generated videos by our method from real ones and address concerns about the visual misinformation.

5 Conclusion

We have proposed a new approach to human motion transfer. It employs the 3D body shape and pose constraints as a condition to regularize the generative adversarial learning framework, and the new condition is more expressive and complete than 2D. We also design an enhancement mechanism to enhance the detailed characteristics of synthesized results using detailed information from real source frames. Extensive experiments show that our method outperforms existing methods both qualitatively and quantitatively.

Acknowledgment

This work was supported by the Beijing Municipal Natural Science Foundation for Distinguished Young Scholars (No. JQ21013), the National Key R&D Program of China (No.2020AAA0104500), the National Natural Science Foundation of China (No. 62061136007 and No. 61872440), Royal Society Newton Advanced Fellowship (No. NAF\R2\192151), the Centre for Applied Computing and Interactive Media (ACIM) of School of Creative Media, City University of Hong Kong and the Youth Innovation Promotion Association CAS.

References

[1] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *IEEE International Conference on Computer Vision*, 2019, pp. 5932–5941.

[2] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *Advances in Neural Information Processing Systems*, 2018, p. 1152–1164.

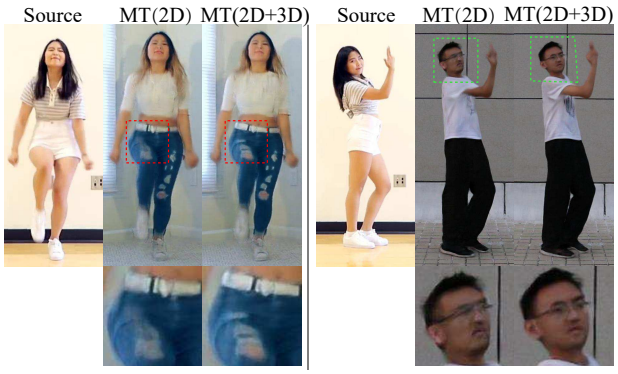
[3] W. Liu, Z. Piao, M. Jie, W. Luo, L. Ma, and S. Gao, "Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis," in *IEEE International Conference on Computer Vision*, 2019, pp. 5903–5912.

[4] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: driving visual speech with audio," in *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, 1997, p. 353–360.

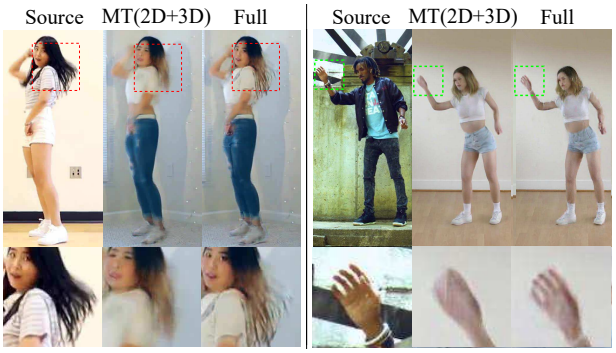
[5] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *IEEE International Conference on Computer Vision*, 2003, pp. 726–733.

[6] J. Lee and S. Y. Shin, "A hierarchical approach to interactive motion editing for human-like figures," in *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, 1999, pp. 39–48.

[7] C. Hecker, B. Raabe, R. W. Enslow, J. DeWeese, J. Maynard, and K. van Prooijen, "Real-time motion retargeting to highly varied user-created morphologies," in *SIGGRAPH*, 2008, pp. 1–11.



(a) Effect of 3D constraints. With 3D constraints, the results are improved in movements with occlusion such as bending legs (left) and structural integrity such as profile synthesis (right).



(b) Effect of the DE-Net. The DE-Net shows superior results in the generation of details in the face (left) and hand (right) areas.

Fig. 13. Visual comparison for the ablation study. We show the generated results of different conditions set in the ablation study.

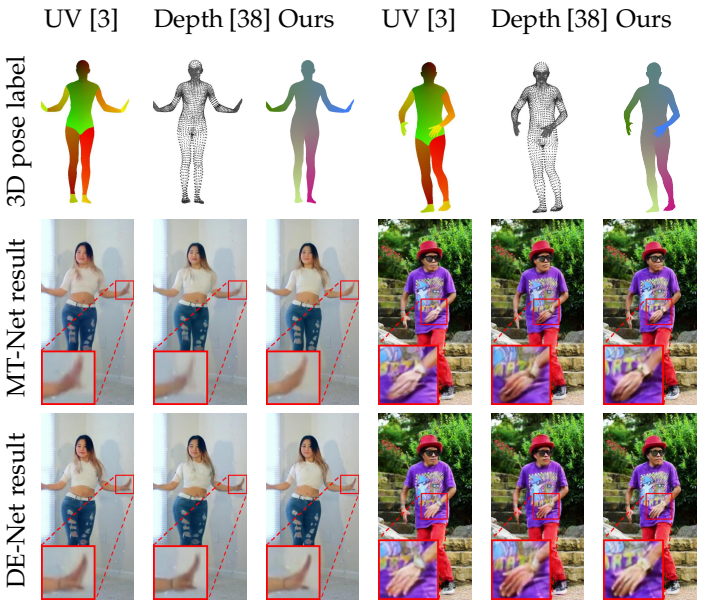


Fig. 14. Comparisons of using Laplacian feature with alternative 3D constraints. The first row visualizes different 3D constraints, the second row exhibits the direct transfer results of MT-Net, and the third row shows the detail-enhanced results of DE-Net. Note the inconsistent colors between wrists and hands caused by the discontinuous UV representation.

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

[9] M. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 469–477.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision*, 2017, pp. 2242–2251.

[11] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *International Conference on Machine Learning*, p. 1857–1865.

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[13] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, "Recycle-gan: Unsupervised video retargeting," in *European Conference on Computer Vision*, 2018, pp. 119–135.

[14] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8798–8807.

[15] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. V. Guttag, "Synthesizing images of humans in unseen poses," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8340–8348.

[16] P. Esser, E. Sutter, and B. Ommer, "A variational u-net for conditional appearance and shape generation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8857–8866.

[17] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Advances in Neural Information Processing Systems*, 2017, pp. 405–415.

[18] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 99–108.

[19] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, pp. 1302–1310.

[20] N. Neverova, R. Alp Guler, and I. Kokkinos, "Dense pose transfer," in *European Conference on Computer Vision*, 2018, pp. 123–138.

[21] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7297–7306.

[22] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3d deformation model for tracking faces, hands, and bodies," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8320–8329.

[23] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *IEEE Conference on Computer Vision and Pattern Regognition*, 2018, pp. 7122–7131.

[24] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 975–10 985.

[25] D. Xiang, H. Joo, and Y. Sheikh, "Monocular total capture: Posing face, body, and hands in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 957–10 966.

[26] M. Meyer, M. Desbrun, P. Schröder, and A. H. Barr, "Discrete differential-geometry operators for triangulated 2-manifolds," in *Visualization and mathematics III*, 2003, pp. 35–57.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[28] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," in *Advances in Neural Information Processing Systems*, 2021, pp. 852–863.

[29] S.-Y. Chen, W. Su, L. Gao, S. Xia, and H. Fu, "DeepFaceDrawing: Deep generation of face images from sketches," in *SIGGRAPH*, 2020, pp. 72:1–72:16.

[30] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for stylegan image manipulation," *ACM Transactions on Graphics*, pp. 1 – 14, 2021.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," in *IEEE Transactions on Image Processing*, 2004, pp. 600–612.

[32] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

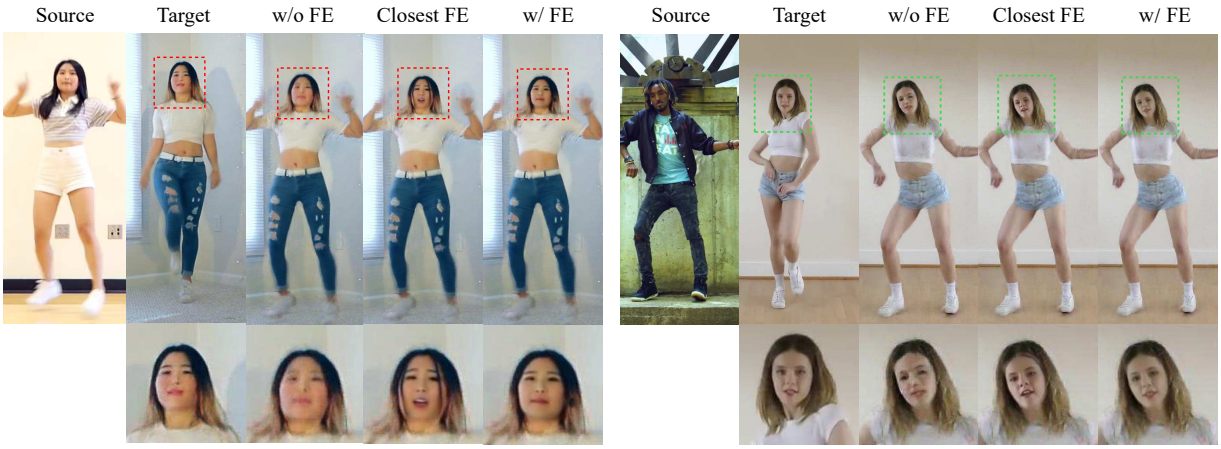
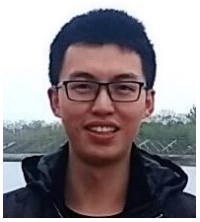


Fig. 15. Effect of the facial enhancement step. It can be seen that our approach can help generate more faithful character faces. In contrast the "Closest FE" baseline fails to keep the size or angle of faces, leading to fusion artifact (left) or incorrect poses (right).

Fig. 16. Failure cases. For each case, the source image is shown on the left and our transfer result on the right.

[33] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.
[34] 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 *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.
[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.
[36] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.
[37] Y. Li, C. Huang, and C. C. Loy, "Dense intrinsic appearance flow for human pose transfer," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3688–3697.
[38] S. Prokudin, M. J. Black, and J. Romero, "Smplpix: Neural avatars from 3d human models," in *IEEE Winter Conference on Applications of Computer Vision*, 2021, pp. 1809–1818.
[39] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A survey on deepfake video detection," *Iet Biometrics*, vol. 10, no. 6, pp. 607–624, 2021.



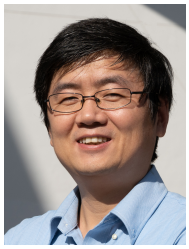
Yang-Tian Sun received his bachelor's degree in applied physics from the University of Science and Technology Beijing. He is currently a Master student in the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include computer graphics and computer vision.

Qian-Cheng Fu graduated from the North China University of Technology with a bachelor's degree in computer science. He is currently a PhD student studying computer science at Boston University.

Yue-Ren Jiang received his bachelor's degree in chemistry from the University of Chinese Academy of Sciences in 2019. He is currently a master student at the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include computer vision and computer graphics.

Zitao Liu received his Ph.D degree in Computer Science from University of Pittsburgh in 2016. He is currently the Head of Engineering, ThinkAcademy International at TAL Education Group. Before joining TAL, he was a senior research scientist at Pinterest. His research interests include AI in education, multimodal knowledge representation and user modeling. He serves as the Executive Committee of the International AI in Education Society.

Yu-Kun Lai received his bachelor's degree and PhD degree in computer science from Tsinghua University in 2003 and 2008, respectively. He is currently a Professor in the School of Computer Science & Informatics, Cardiff University. His research interests include computer graphics, geometry processing, image processing and computer vision. He is on the editorial boards of *Computer Graphics Forum* and *The Visual Computer*.



HongBo Fu received a BS degree in information sciences from Peking University, China, in 2002 and a PhD degree in computer science from the Hong Kong University of Science and Technology in 2007. He is a Full Professor at the School of Creative Media, City University of Hong Kong. His primary research interests fall in the fields of computer graphics and human computer interaction. He has served as an Associate Editor of The Visual Computer, Computers & Graphics, and Computer Graphics Forum.

Lin Gao received his PhD degree in computer science from Tsinghua University. He is currently an Associate Professor at the Institute of Computing Technology, Chinese Academy of Sciences. He has been awarded the Newton Advanced Fellowship from the Royal Society and the AG young researcher award. His research interests include computer graphics and geometric processing.