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, 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

The problem of 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 transfer 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). To achieve this, it is essential to produce high-quality image-to-image translation of 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 the target video, and then use the pose of the 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 a 3D human model from a 2D image of both the source and target subjects, and then adjust the pose of the target human body to match the source (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 3D structure.

This along with the 2D pose figure extracted from the 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, and does not fully utilize the appearance of the source. 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 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. Our problem may also be seen as an appearance transfer problem if viewed from a single frame perspective. However, from a holistic perspective, our goal is to transfer the motion from the source image domain to the target. Therefore, we define our task as motion transfer in this paper.

We summarize our contributions as follows: 1) We propose to reconstruct a 3D human body with its shape from a target frame and its pose from a source frame, and project it to 2D to serve as a GAN-based network condition. This contains rich 3D information including body shape, pose and occlusion to help maintain the structural characteristics of the human body in the generated images. 2) We introduce the detail enhancement net (DE-Net), which utilizes the information from the real source frames to enhance details

* *Corresponding Author.*

- *Y.T. Sun, Y.R. Jiang and L. Gao are with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. E-mail:sunyangtian@ict.ac.cn, jiangyueren15@mails.ucas.ac.cn, gaolin@ict.ac.cn*
- *Q.C. Fu is with the Department of Computer Science, Boston University E-mail:qcfu@bu.edu*
- *Z.T. Liu is with TAL AI Lab, TAL Education Group, Beijing, China E-mail:liuzitao@100tal.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.B. 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).

in the generated results. Extensive experiments show that our method outperforms the state-of-the-art methods, especially for challenging cases where the source and target have substantial differences.

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 [3], [4]. 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 [5] and transfer motion between 3D *skeletons* [6], whereas we consider using a reconstructed 3D body mesh to guide motion transfer in the image domain, which provides much richer constraints.

Recently, the rapid advances of deep learning, especially generative adversarial networks (GANs) and their variations (e.g., cGAN [7], CoGAN [8], CycleGAN [9], DiscoGAN [10]) have provided a powerful tool for image-to-image translation, which has yielded impressive results across a wide spectrum of synthesis tasks and shows its ability to synthesize visually pleasing images from conditional labels. Pix2pix [11], based on a conditional GAN framework, is one of the pioneering works. CycleGAN [9] further presents the idea of cycle consistency loss for learning to translate between two domains in the absence of paired images, and Recycle-GAN [12] combines both spatial and temporal constraints for video retargeting tasks. Pix2pixHD [13] 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], [14], [15], [16], [17] have been proposed for human motion transfer between two domains. The key idea of these approaches is to decouple the pose information from the input image and use it as the input of a GAN network to generate a realistic output image. For example, in [14], the input image is separated into two parts: the foreground (or different body parts) and background, and the final realistic image is generated by separately processing and cross fusion of the two parts. Chan *et al.* [1] extract pose information with an off-the-shelf human pose detector OpenPose [18], [19], [20], [21] and use the pix2pixHD [13] framework together with a specialized Face GAN to learn a mapping from a 2D pose figure to an image. Neverova *et al.* [22] adopt a similar idea but use the estimation of DensePose [23] 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 such as [24], [25], [26], [27], which can reconstruct a 3D model with 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.* [28] present a novel warping strategy, which also uses the projection of 3D models to tackle motion transfer, appearance transfer and novel view synthesis within a unified model. However, due to the diversity of their network functionalities, it does not perform particularly well in the aspect of motion transfer.

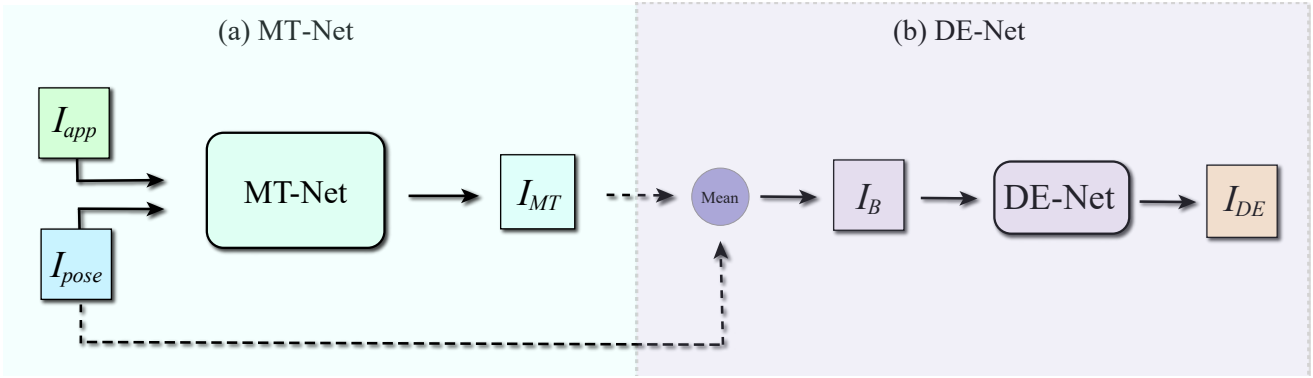


Fig. 2. The architecture of our Motion Transfer Net (MT-Net) and Detail Enhancement Net (DE-Net). (a) MT-Net takes two images I_{app} and I_{pose} as two inputs, and outputs a new image I_{MT}, which has the appearance of I_{app} and the pose of I_{pose}. (b) DE-Net takes image I_B as input, which is the blending of raw transfer result I_{MT} possibly with blurry artifacts and corresponding real frame I_{pose} with rich details, and the aim is to generate an image I_{DE} in the target domain with the details enhanced.

3 METHOD

We aim to generate a new video of the target person imitating the character movements in the 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 denote $\mathcal{S} = \{S_i\}$ as a set of source video frames, and $\mathcal{T} = \{T_j\}$ as a set of target frames. Our architecture can be divided into two modules, as shown in Figure 2: the Motion Transfer Net (MT-Net) on account of motion transfer across two domains and the Detail Enhancement Net (DE-Net) used for the enhancement of details. More specifically, MT-Net takes **two real video frames** I_{app} and I_{pose} as input, and generates an output image I_{MT} that has the same appearance as I_{app} and the pose as I_{pose}. DE-Net takes image I_B as input, which is the blending of raw transfer result I_{MT} with blurred details and corresponding real frame I_{pose} with rich details, and aims to generate an image I_{DE} in the **target domain** with the details enhanced. Our *training pipeline* is as follows:

Within-Domain Pre-Training of MT-Net. To stabilize the training process, we first pre-train MT-Net using within-domain samples. For the domain $\mathcal{S}$, let $I_{app}, I_{pose} \in \mathcal{S}$, and we can obtain I_{MT}, which is the reconstructed source frame. For the domain $\mathcal{T}$, let $I_{app}, I_{pose} \in \mathcal{T}$, and we should obtain the reconstructed target frame. This process initializes the MT-Net. Note that for each I_{pose}, I_{app} is randomly selected from the corresponding domain and fixed during the training process.

Training of DE-Net. After the pre-training of MT-Net, let $I_{app} \in \mathcal{S}$ and $I_{pose} \in \mathcal{T}$, we generate initial transferred image I_{MT} by the MT-Net, which is often blurred. We then calculate a blended image I_B which is an average of I_{MT} and corresponding real frame I_{pose} that contains clear details. We then train the DE-Net to discern and generate details from the blended result selectively to produce output image I_{DE} with details enhanced that matches the target domain.

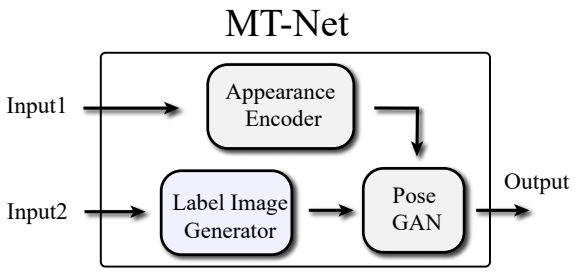


Fig. 3. Architecture of MT-Net, which synthesizes a motion transferred image with the appearance of Input1 and pose of Input2.

Our *transfer pipeline* is as follows: Let $I_{app} \in \mathcal{T}$ and $I_{pose} \in \mathcal{S}$, we can get the initial transfer result I_{MT} (with the source pose and target appearance) using MT-Net. And then we can obtain the final result I_{DE} with details enhanced by the DE-Net.

Note that the domains of I_{app} and I_{pose} for training DE-Net and transfer are swapped, because in the training setting, I_{MT} has the appearance as the source and pose as the target, and DE-Net aims to produce an image with the appearance and pose both in the target domain, so the ground truth of I_{DE} is available (which is exactly the corresponding $I_{pose} \in \mathcal{T}$). This provides supervision for training I_{DE} to enhance details in the target domain (i.e. with the appearance of the target subject). Such supervision is not available if the transfer setting is used. Although both our method and Ma *et al.* [16] use two-stage pipelines, our method is essentially different: our novelty lies in the regularization of 3D constraints in the MT-Net and innovation of utilizing source subject details in the DE-Net; see details in the remaining subsections.

3.2 Motion Transfer Net

As illustrated in Figure 3, the Motion Transfer Net consists of 3 parts: Label Image Generator that produces label images that encode human pose and shape information, Appearance Encoder that encodes the appearance of the

input image, and Pose GAN that produces an output image with given appearance and pose/shape constraints.

3.2.1 Label Image Generation via 3D Human Models

To maintain the structural integrity of the generated results and produce realistic images for actions involving self-occlusions, we utilize the 3D geometry information of the underlying subject to produce label images as the GAN condition to regularize the generative network. The architecture of our label image generator is shown in Figure 4.

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 [27]. This leads to a 3D deformable mesh model including the details of body, face and fingers. When transferring between two domains, the 3D human models also allow the 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. This can be alleviated by simply applying temporal smoothing to 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 a particular color corresponds to a specific location on the human body. To achieve this, we propose to extract the three non-trivial eigenvectors corresponding to three smallest eigenvalues and consider them as a 3-channel image assigned to each vertex [29], which 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 [18], [19], [20], [21] 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. And we will discuss the roles of these two conditions and their combination play separately in the ablation study.

3.2.2 Pose to Image Translation

We learn the mapping from label image sequences to realistic image sequences by training a conditional GAN, consisting of Appearance Encoder and Pose GAN. The design of Pose GAN is similar to pix2pixHD [13]: It is composed of a generator G_{pose}, and two multi-scale frame discriminators with the same architecture for images in the source and target domains, respectively. The two 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. The difference of our network from pix2pixHD is that the data we use to learn the mapping includes not only target video, but also source video. We have done

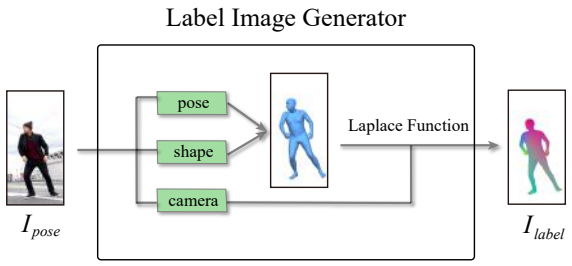


Fig. 4. Architecture of our label image generator. We reconstruct a 3D mesh of the transferred human body, assign eigenvectors corresponding to the three smallest eigenvalues of its Laplace matrix as intrinsic features (visualized in RGB color), and project it to form a 3D constraint image, denoted as I_{label}.

this by conditioning G_{pose} on both label images and appearance features, extracted by the Label Image Generator and Appearance Encoder respectively. See Figure 5 for the generative network architecture. It is worth mentioning that in order to solve the problem of poor continuity caused by single frame generation, similar to [1], adjacent frames are involved in training to improve temporal coherence.

Appearance Encoder E_{app} and Pose GAN Generator G_{pose}. As said above, to make full use of the given data and meet the need of the subsequent detail enhancement, we train the generative network using data from both source video and target video. However, training two separate conditional GANs has a high overhead for computing resources and time. In order to simplify this process, we introduce an Appearance Encoder, and use label images containing 2D/3D constraints and appearance features together to guide Pose GAN to produce the reconstructed image (for within-domain input) or initial transfer result I_{MT} (for cross-domain input). Note that for a new source subject video, our framework only needs to fine-tune the upsampling part of Pose GAN for the generation of the new subject.

Appearance Encoder is a fully convolutional network that extracts appearance features of the input image I_{app}, which 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 is the main part of MT-Net, which consists of three submodules: Downsampling, ResNet blocks and Upsampling. It works on both label images and the appearance features extracted by the Appearance Encoder, and synthesize results with the corresponding pose and appearance. As shown in Figure 5, the output of Appearance Encoder is added to the intermediate ResNet blocks in the generator.

Pose GAN Discriminator. We use the multi-scale discriminator presented in pix2pixHD [13]. Discriminators of different scales can give the discrimination of images at different levels. In our method, we use two discriminators $D^{\mathcal{S}}_{pose}$ and $D^{\mathcal{T}}_{pose}$ to discriminate the probability of generated images belonging to the corresponding domain, each with 3 scales.

Temporal Smoothing. We use the time smoothing strategy in [1] to enhance the continuity between adjacent gener-

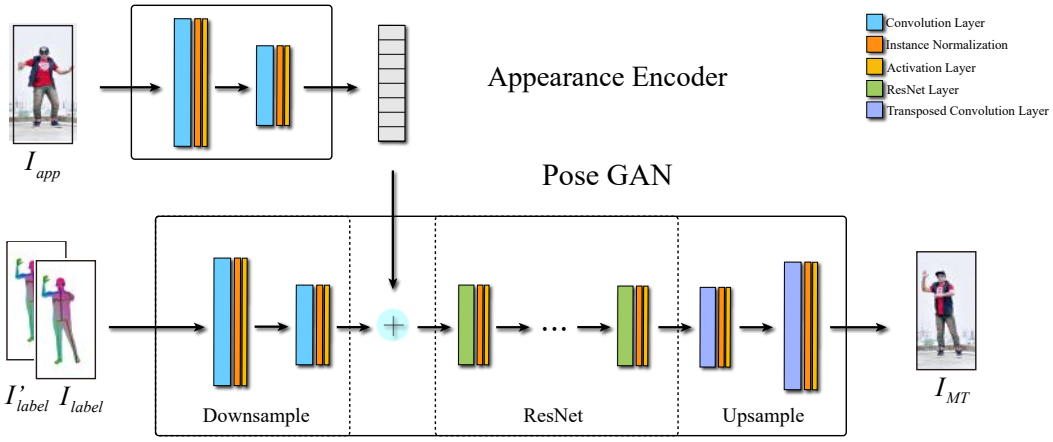


Fig. 5. Architecture of the generative network. Appearance image I_{app} is randomly selected and sent to Appearance Encoder (denoted as E_{app}) to obtain appearance features. Adjacent label images I_{label} (current frame) and I'_{label} (previous frame) are sent to Pose GAN (denoted as G_{pose}) together with appearance features to generate the reconstructed image or initial transfer result I_{MT}.

ated 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, let $d \in \{\mathcal{S}, \mathcal{T}\}$ denote the domain in which the training images are selected, our conditional GAN has the following objective:

$$\begin{aligned}\mathcal{L}^d_{\text{MT}}(E_{app}, G_{pose}, D^d_{pose}) = \\ \mathbb{E}[\log D^d_{pose}(I_{label}, I'_{label}, I_{pose}, I'_{pose})] + \\ \mathbb{E}[\log(1 - D^d_{pose}(I_{label}, I'_{label}, I_{MT}, I'_{MT})].\end{aligned} \quad (1)$$

Here the discriminator D^d_{pose} takes a pair of adjacent images in the domain d, and classifies them to real images (the current frame I_{pose}, and previous frame I'_{pose} from the training set), or fake images (I_{MT} and the previous frame output I'_{MT} generated by the Pose GAN).

3.3 Detail Enhancement Net

Through the first stage of training, we can obtain initial transfer results with blurred details: source to target transfer result I_{S2T} as well as target to source transfer result I_{T2S}. We can also construct paired data (I_S, I_{S2T}) and (I_T, I_{T2S}), as shown in Figure 6, where images in the same pair have the same pose but different appearances, and different clarity of details, which motivates us to use a DE-Net to enhance the details from the blended image. Note that in different videos, subjects might have different builds or positions relative to camera. In such cases, before sending the paired data to DE-Net we need to align the source frame in accordance with target by applying the transformation calculated from reconstructed mesh with source parameters and target respectively. We have proved the effectiveness of this step in Fig 11 and more details will be included in supplementary material.

The purpose of our DE-Net is to generate clear details of target images from the blended image pair. It is a GAN where the generator G_{DE} is a U-net which synthesizes images in the target domain with clear details, as illustrated

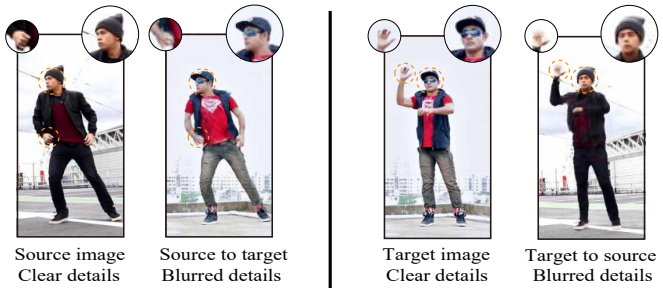


Fig. 6. Comparison of the source image I_S and source to target transfer result I_{S2T}, as well as the target image I_T and target to source transfer result I_{T2S}.

in Figure 7. The discriminator D_{DE} discerns the "real" images (ground truth) and "fake" images (synthesized by G_{DE}).

In the training stage, we use the blended image pair (I_{T2S}, I_T) as input and supervisely train DE-Net with I_T as ground truth. The use of mean blended image instead of concatenation avoids the output overfitting to I_T. We optimize the DE-Net by the following objective:

$$\begin{aligned}\mathcal{L}_{\text{DE}}(G_{DE}, D_{DE}) = \mathbb{E}[\log D_{DE}(I_{label}, I'_{label}, I_T)] \\ + \mathbb{E}[1 - \log D_{DE}(I_{label}, I'_{label}, I_{DE})],\end{aligned} \quad (2)$$

where I_{label} and I'_{label} are the label images used to generate I_{T2S}.

In the transfer stage, we use the source to target transfer result I_{S2T} and the corresponding source image I_S to obtain enhanced transfer result.

3.4 Full Objective

The training of our network is divided into two stages. First we train the Motion Transfer Net, which consists of Pose GAN and Appearance Encoder. The full objective contains

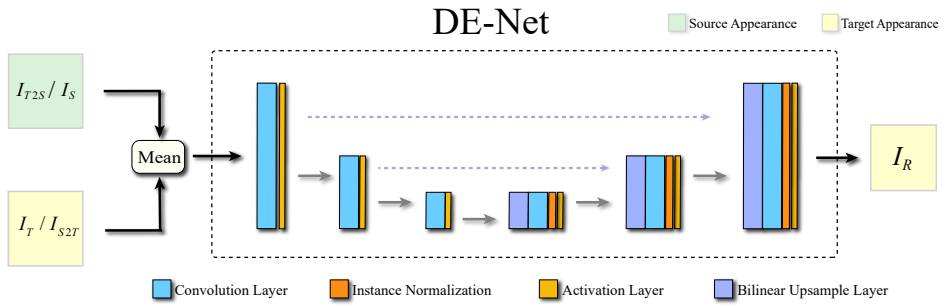


Fig. 7. Architecture of Detail Enhancement Net (DE-Net). The main part of our DE-Net is a U-net, which takes the mean of paired data (I_{T2S}, I_T) or (I_S, I_{S2T}) as input, and synthesizes a target image I_{DE} with details enhanced.

adversarial loss, perceptual loss and discriminator feature-matching loss, which has the following form:

$$\begin{aligned}\min_{E_{app},G_{pose}} ((&\max_{D^{\mathcal{S}}_{pose}} \mathcal{L}^{\mathcal{S}}_{\text{MT}}(E_{app}, G_{pose}, D^{\mathcal{S}}_{pose})) \\ &+ (\max_{D^{\mathcal{T}}_{pose}} \mathcal{L}^{\mathcal{T}}_{\text{MT}}(E_{app}, G_{pose}, D^{\mathcal{T}}_{pose})) \\ &+ \lambda_P \mathcal{L}_P(I_{MT}, I_{pose}) \\ &+ \lambda_{FM} \mathcal{L}_{FM}((E_{app}, G_{pose}), D^{\mathcal{S}}_{pose}) \\ &+ \lambda_{FM} \mathcal{L}_{FM}((E_{app}, G_{pose}), D^{\mathcal{T}}_{pose})). \end{aligned} \quad (3)$$

Here, $\mathcal{L}^{\mathcal{S}}_{\text{MT}}$ and $\mathcal{L}^{\mathcal{T}}_{\text{MT}}$ are defined in Eq. 1. The perceptual loss $\mathcal{L}_P$ regularizes the generated result I_{MT} to be closer to the ground truth I_{pose} in the VGG-19 [30] 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 $\mathcal{L}_{FM}$ is presented in pix2pixHD [13] and similarly regularizes the output using intermediate result of the discriminator, calculated as

$$\mathcal{L}_{FM}(G, D_k) = \mathbb{E} \sum_{i=1}^{T} \frac{1}{N_i} [||D_k^{(i)}(s, x) - D_k^{(i)}(s, G(s))||_1], \quad (5)$$

where T is the number of layers, N_i is the number of elements in the ith 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

$$\begin{aligned}\min_{G_{DE}}((&\max_{D_{DE}} \mathcal{L}_{\text{DE}}(G_{DE}, D_{DE})) + \lambda_P \mathcal{L}_P(I_{DE}, I_{pose}) \\ &+ \lambda_{FM} \mathcal{L}_{FM}(G_{DE}, D_{DE})). \end{aligned} \quad (6)$$

Here $\mathcal{L}_{\text{DE}}$ is defined in Eq. 2. The perceptual and discriminator feature-matching losses are defined in Eqs. 4 and 5.

4 Experiment

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

4.1 Setup

Dataset. To verify the performance of our method, we collected three types of data: the dataset published by [1], 10 in-the-wild single-dancer videos from YouTube (including the data used by [2]) and 5 videos filmed by ourselves, out of which 2 with ordinary background and 3 with green screen. All videos are at 1920×1080 resolution or 1024×512 resolution. Each subject wears different clothes and performs different types of action such as freestyle dancing and stretching exercises. To prepare for training and testing, We cut the start and end parts that contain no action, and crop and normalize each frame to the same size by simple scaling and translation.

Implementation details. We adopt a multi-stage training strategy in our method using Adam optimizer with learning rate 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 DE-Net is trained individually for 10 epoches. We set hyperparameters $\lambda_{FM} = 10$ and $\lambda_P = 5$ for both stages. More details about MT-Net and DE-Net are given in supplementary material.

Existing methods. We compare our performance with existing state-of-the-art methods vid2vid [2], Everybody Dance Now [1] and Liquid Warping GAN [28], using 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 the 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 training and test 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 Frchet Inception Distance [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 exclude the green screen dataset in quantitative evaluation to focus on more challenging cases.

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.

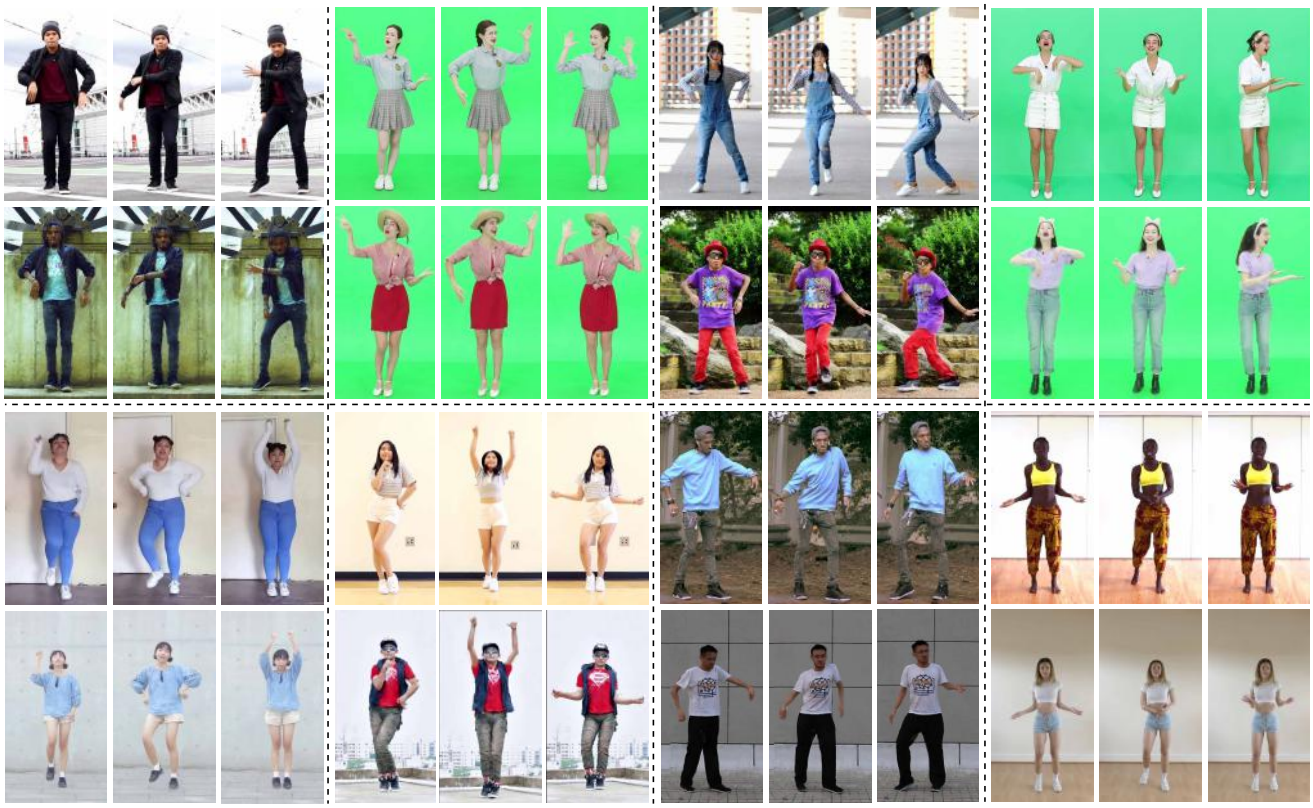


Fig. 8. Transfer results. We show the generated frames of several subjects with different genders, races or builds
. In each group, the top row shows the source subject and the bottom row shows the generated target subject.

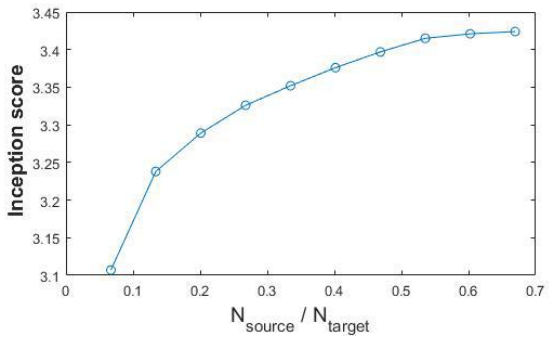


Fig. 9. The effect of number of source frames on transfer results. We show the inception score of transfer results (y-axis) w.r.t the ratio of source frames number to target (x-axis). For inception score, higher is better.

Metric		Method			
		vid2vid	Chan *et al.*	LW-GAN	ours
Self-trans	SSIM	0.781	0.836	0.790	**0.887**
	LPIPS	0.096	0.067	0.106	**0.041**
Cross-trans	IS	3.186	3.238	3.064	**3.424**
	FID	59.98	57.02	81.20	**53.76**

TABLE 1
Quantitative comparison with state of the arts on the dance dataset. Metrics are averaged over all subjects. For SSIM and IS, higher is better. For LPIPS and FID, lower is better.

4.2.1 *Comparison with existing methods.*

Comparison results with state of the arts are reported in Table 1.

It can be found that our method performs better than others.

4.2.2 *Ablation study*

In this part, we perform an ablation study to verify the impact of each component of our model, including using 3D constraints ("3D") and DE-Net ("DE"). Our full pipeline is indicated "Full". When 3D is disabled, we use 2D pose figures as default label images.

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 literally defines the shape and geometry information and serves as a condition of Pose GAN, which plays an important role in generation, as is shown in **MT(3D only)** and **MT** . And the score of **MT+3D** shows the complementarity of 2D and 3D condition on this task. The comparison of **Full** and **MT+3D** (or **MT+DE** and **MT**) proves the performance of DE-Net.

Furthermore, we can observe that scores of self-transfer between **MT** and **MT+3D** (or **MT+DE** and **Full**) are similar. This is because source and target subjects share the same body shape in self-transfer, which somewhat limits the effectiveness of 3D constraints, where the scores of cross-subject

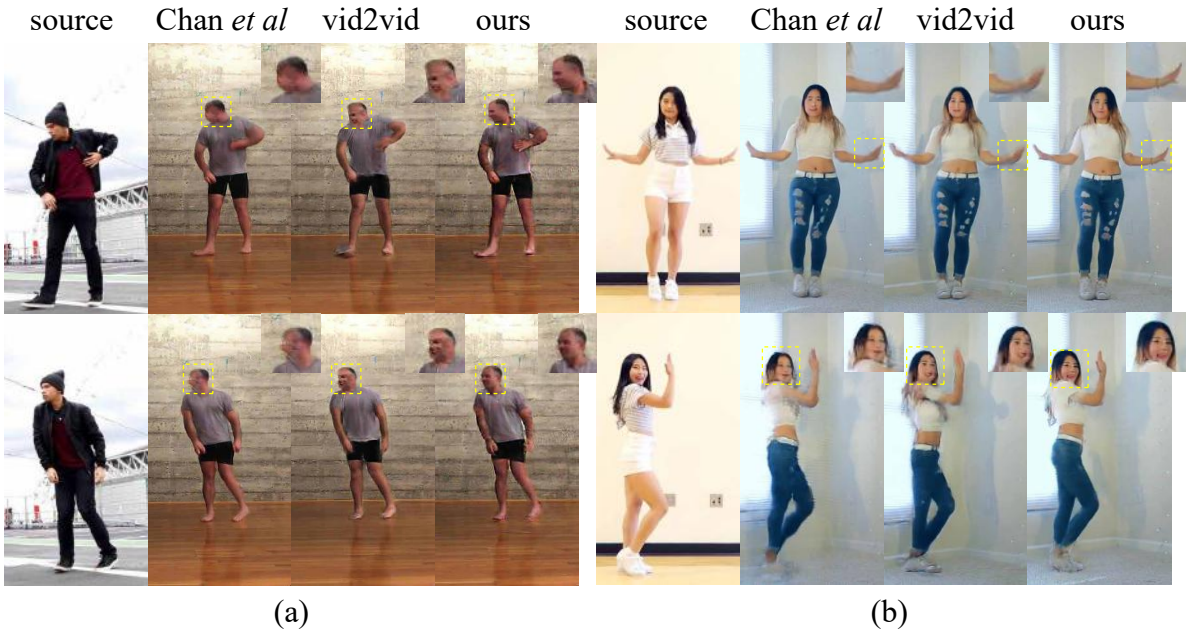


Fig. 10. We compare with Chan *et al.* [1] and vid2vid [2] on the data published by [1] (a) and the data used in [2] (b) for the sake of fairness. Our method obtains more accurate results in details for occlusion actions such as side faces and bending fingers.

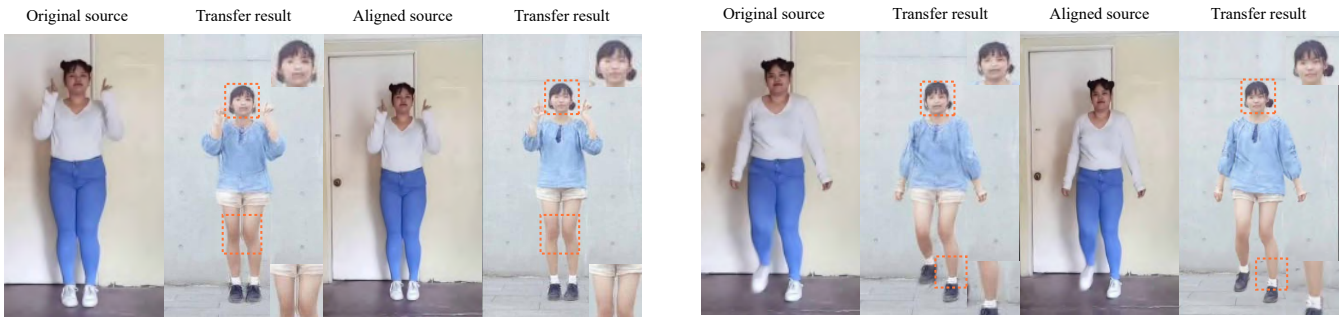


Fig. 11. Effects of aligning transformation. It can be seen that the absence of transformation will make DE-Net unable to match source & target characters accurately, resulting in vague results and body shape changes.

transfer demonstrate the important role 3D information plays on transfer between different subjects with different shapes.

Metric		Method				
		MT	MT(3D only)	MT+3D	MT+DE	Full
Self-trans	SSIM	0.828	0.831	0.856	0.877	**0.887**
	LPIPS	0.064	0.063	0.058	0.043	**0.041**
Cross-trans	IS	3.178	3.265	3.325	3.224	**3.424**
	FID	58.62	56.68	55.77	57.56	**53.76**

TABLE 2
Ablation study. For SSIM and IS, higher is better. For LPIPS and FID, lower is better.

4.2.3 User evaluation

We also conduct a user study to measure the human perceptual quality for cross-subject transfer results. In our experiments, we compare videos generated by vid2vid, Chan *et al.*, Liquid Warping GAN and our method. Specifically, we show to volunteers a series of videos by each of the methods at the resolution of 1024×512, and the volunteers are given unlimited time to make responses. 50 distinct participants are involved and each of them is asked to select: 1) the clearest result with rich details; 2) the most temporally stable result; and 3) the overall best result. As shown in Table 3, our method is more realistic, with richer details and with better temporal stability in comparison with other methods.

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 3
User study. We report the percentage of participants' choice in three different aspects respectively.

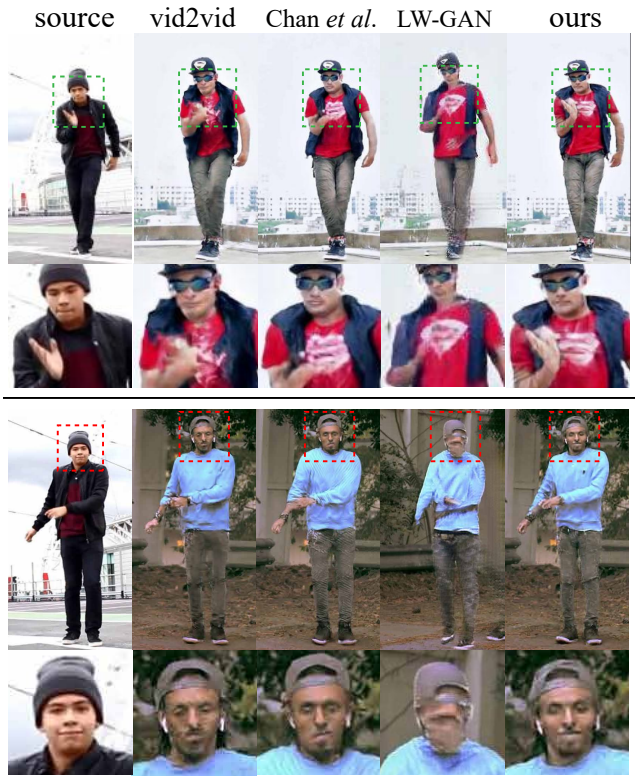


Fig. 12. Comparison with state-of-the-art methods. We show the generated results by vid2vid, Everybody Dance Now, Liquid Warping GAN and our method. Only our method has reconstructed the wriggling hand and smile face.

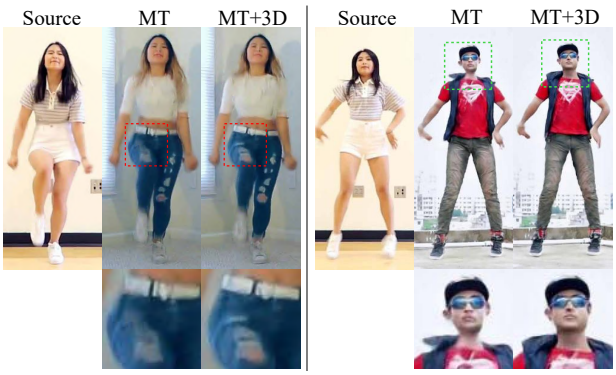
4.2.4 *Effect of number of source frames on results*

The previous motion transfer methods such as [1] and [2] only use target frames at training stage, and the quality of their generative model is not directly related to the number of source frames. While in our method, source frames and target are both involved in training. Therefore, it's meaningful to explore the influence of source frames number on the generated results. We carry out experiments on all subjects and record the averaged evaluation of generated iamge quality with respect to the ration of source frames number to target when the number of target frames is fixed, as is shown in Fig 9. Note that we choose inception score as metric. It can be seen that the loss has converged when the ratio is around 0.5.

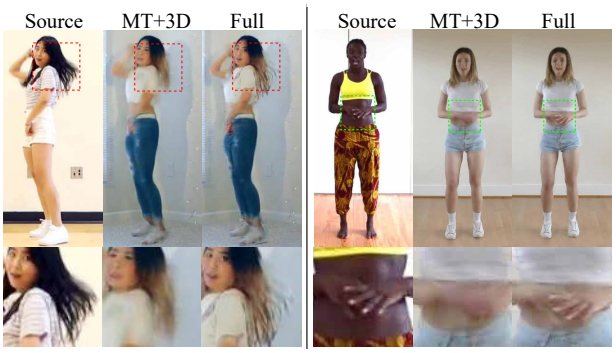
4.3 Qualitative Results

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 existing methods in Figure 12 and Figure 10.

As illustrated in the first row of Figure 12, our method can enhance the structural integrity of arms and legs, and avoid the missing hands in the case that other methods fail to generate. At the same time, our method can also characterize details of the generated results more accurately, such as facial expression shown in the second row of Figure 12.



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



(b) Effect of the DE-Net. The DE-Net shows superior results in the generation of details in 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. Failure cases. For each case, the source image is shown on the left and our transfer result on the right.

Figure 13 shows the advantage of using the 3D constraints and DE-Net.

4.4 Limitations and Discussions

Although our model is able to synthesize motion transferred images with high authenticity and details, there are still several limitations. We show some failure cases with visual artifacts shown in Figure 14. In the left example, our model fails to eliminate the long hair of the source character in result, while in the right, some undesired part of clothes appears in the generated image because of the loose source clothes. These failure cases are mainly attributed to the abnormal movement of source character, which causes large changes in human body shape (e.g., perturbations in hair or clothing) and makes the DE-Net fail to eliminate extra

details. Our future work will focus on improving the ability of DE-Net to avoid the appearance of undesired details in the transfer results.

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, which is more expressive and complete than 2D. We also design a enhancement mechanism to reinforce the detail characteristics of synthesized results using detailed information from real source frames. Extensive experiments show that our method outperforms existing methods both visually and quantitatively.

References

[1] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody Dance Now," 2018.

[2] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," 2018.

[3] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: driving visual speech with audio," in *SIGGRAPH*, 1997.

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

[5] J. Lee and S. Y. Shin, "A hierarchical approach to interactive motion editing for human-like figures," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1999, Los Angeles, CA, USA, August 8-13, 1999*, W. N. Waggenspack, Ed. ACM, 1999, pp. 39–48. [Online]. Available: https://doi.org/10.1145/311535.311539

[6] 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*, 2008.

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

[8] M. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 469–477. [Online]. Available: http://papers.nips.cc/paper/6544-coupled-generative-adversarial-networks

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[10] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," *CoRR*, vol. abs/1703.05192, 2017. [Online]. Available: http://arxiv.org/abs/1703.05192

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

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

[13] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. (2017) High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs.

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

[15] P. Esser, E. Sutter, and B. Ommer, "A variational u-net for conditional appearance and shape generation," 2018.

[16] 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.

[17] 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.

[18] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," in *arXiv preprint arXiv:1812.08008*, 2018.

[19] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[20] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *CVPR*, 2017.

[21] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.

[22] N. Neverova, R. Alp Guler, and I. Kokkinos, "Dense pose transfer," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 123–138.

[23] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," 2018.

[24] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3d deformation model for tracking faces, hands, and bodies," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[25] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Computer Vision and Pattern Regognition (CVPR)*, 2018.

[26] 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 *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[27] D. Xiang, H. Joo, and Y. Sheikh, "Monocular total capture: Posing face, body, and hands in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[28] W. Liu, W. L. L. M. Zhixin Piao, Min Jie, and S. Gao, "Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis," in *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

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

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

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[32] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[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.