Neural Rendering and Reenactment of Human Actor Videos

LINGJIE LIU, University of Hong Kong, Max Planck Institute for Informatics
WEIPENG XU, Max Planck Institute for Informatics
MICHAEL ZOLLHÖFER, Stanford University, Max Planck Institute for Informatics
HYEONGWOO KIM, FLORIAN BERNARD, and MARC HABERMANN, Max Planck Institute for Informatics
WENPING WANG, University of Hong Kong
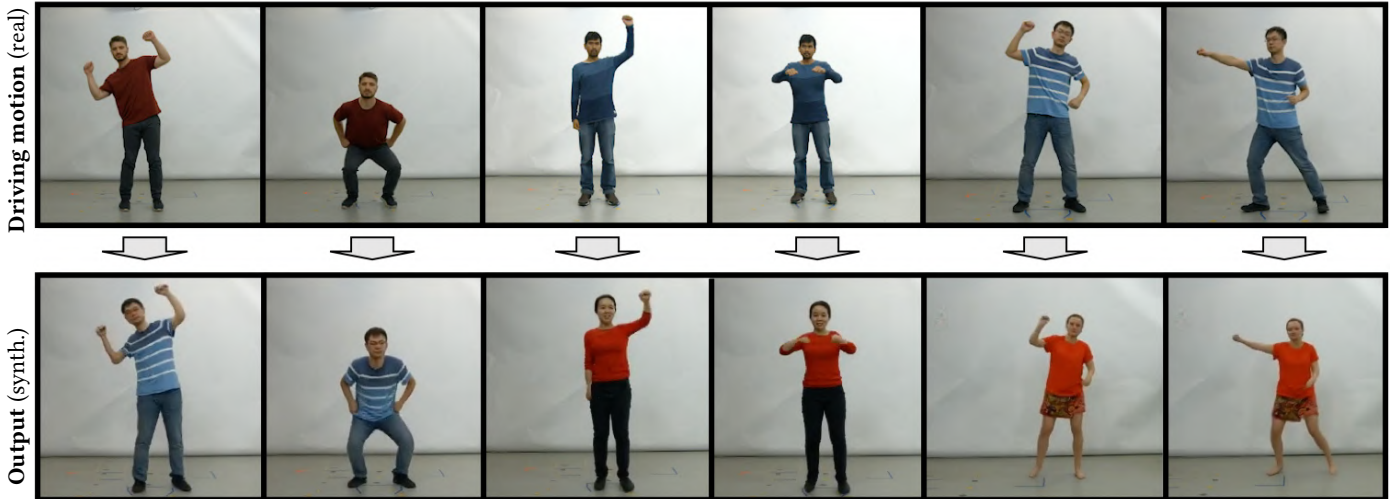CHRISTIAN THEOBALT, Max Planck Institute for Informatics

Fig. 1. We propose a novel learning-based approach for the animation and reenactment of human actor videos. The top row shows some frames of the video from which the *source motion* is extracted, and the bottom row shows the corresponding synthesized target person imagery reenacting the source motion.

We propose a method for generating video-realistic animations of real humans under user control. In contrast to conventional human character rendering, we do not require the availability of a production-quality photo-realistic 3D model of the human, but instead rely on a video sequence in conjunction with a (medium-quality) controllable 3D template model of the person. With that, our approach significantly reduces production cost compared to conventional rendering approaches based on production-quality 3D models, and can also be used to realistically edit existing videos. Technically, this is achieved by training a neural network that translates simple synthetic images of a human character into realistic imagery. For training our networks, we first track the 3D motion of the person in the video using the template model, and subsequently generate a synthetically rendered version of the video. These images are then used to train a conditional generative adversarial network that translates synthetic images of the 3D model into realistic imagery of the human. We evaluate our method for the reenactment of another person that is tracked in order to obtain the motion data, and show video results generated from artist-designed skeleton motion. Our results outperform the state-of-the-art in learning-based human image synthesis.

Authors' addresses: Lingjie Liu, liulingjie0206@gmail.com, University of Hong Kong, Max Planck Institute for Informatics; Weipeng Xu, wxu@mpi-inf.mpg.de, Max Planck Institute for Informatics; Michael Zollhöfer, zollhoefer@cs.stanford.edu, Stanford University, Max Planck Institute for Informatics; Hyeongwoo Kim, hyeongwoo.kim@mpi-inf.mpg.de; Florian Bernard, fbernard@mpi-inf.mpg.de; Marc Habermann, mhaberma@mpi-inf.mpg.de, Max Planck Institute for Informatics; Wenping Wang, wenping@cs.hku.hk, University of Hong Kong; Christian Theobalt, theobalt@mpi-inf.mpg.de, Max Planck Institute for Informatics.

1 INTRODUCTION

The creation of realistically rendered and controllable animations of human characters is a crucial task in many computer graphics applications. Virtual actors play a key role in games and visual effects, in telepresence, or in virtual and augmented reality. Today, the plausible rendition of video-realistic characters—i.e., animations indistinguishable from a video of a human—under user control is also

important in other domains, such as in simulation environments that render training data of the human world for camera-based perception algorithms of autonomous systems and vehicles [Dosovitskiy et al. 2017]. There, simulated characters can enact large corpora of annotated real world scenes and actions, which would be hard to actually capture in the real world. Also, training data for dangerous situations, like a child running unexpectedly onto a street, cannot be captured in reality, but such image data are crucial for training of autonomous systems.

With established computer graphics modeling and rendering tools, creating a photo-real virtual clone of a real human that is indistinguishable from a video of the person is still a highly complex and time consuming process. To achieve this goal, high quality human shape and appearance models need to be hand-designed or captured from real individuals with sophisticated scanners. Real world motion and performance data needs to be artist-designed or captured with dense camera arrays, and sophisticated global illumination rendering methods are required to display animations photo-realistically. In consequence, creation and rendering of a video-realistic virtual human is highly time consuming.

We therefore propose a new efficient and lightweight approach to capture and render video-realistic animations of real humans under user control. At runtime it requires only a monocular color video of a person as input (or any other motion source) that is then used to control the animation of a reenacted video of a different actor. In order to achieve this, we employ a learning-based approach that renders realistic human images merely based on synthetic human animations.

At the training stage, our method takes two short monocular videos of a person as input, one in static posture, and one in general motion. From the *static* posture video, a fully textured 3D surface model of the actor with a rigged skeleton is reconstructed. This character model is then used to capture the skeletal motion seen in the *motion* video using (a modified version of) the monocular human performance capture method of [Mehta et al. 2017b]. While this captures the 3D pose and surface motion of the actor, there is still a significant gap to the expected photo-realistic appearance of the virtual character. Hence, we train a generative neural network in an attempt to fill this gap. Specifically, based on the 3D character model and the tracked motion data, we first render out different image modalities of the animated character (color and depth images, body part segmentations), which correspond to the image frames in the motion video. Then, based on the so-created training data, we train a conditional GAN to reconstruct photo-realistic imagery of the motion video frames using our rendered images as conditioning input.

During testing, we animate the virtual 3D character of the target subject with a user-defined motion sequence, which can stem from an arbitrary source (e.g. motion capture (MoCap) data, artist-designed animations, or videos of an actor), and then render the color image, depth, and semantic masks for each frame of the output video of the virtual character. Then, we pass the rendered conditioning images to the network and thus obtain photo-realistic video of the same person performing the desired motion.

We emphasize that, compared to previous work that mapped face model renderings to realistic face video (*Deep Video Portraits* [Kim

et al. 2018]), translating complete articulated character renderings to video is a much more difficult problem due to more severe pose and appearance changes and the so-resulting non-linearities and discontinuities. Another difficulty is the inevitable imperfection in human body tracking, which directly results in a misalignment between the conditioning input and the ground truth image. Hence, established image-to-image translation approaches like *pix2pix* [Isola et al. 2017] and *Deep Video Portraits* are not directly applicable to full human body performances. To alleviate these problems, we propose a novel GAN architecture that is based on two main contributions:

(i) a part-based dense rendering of the model in RGB and depth channel as conditioning images to better constrain the pose-to-image translation problem and disambiguate highly articulated motions, and

(ii) an attentive discriminator network tailored to the character translation task that enforces the network to pay more attention to regions where the image quality is still low.

The proposed method allows us to reenact humans in video by using driving motion data from arbitrary sources and to synthesize video-realistic target videos. In our experiments, we show high quality video reenactment and animation results on several challenging sequences, and show clear improvements over most related previous work.

2 RELATED WORK

We focus our discussion on the most related performance capture, video-based rendering, and generative modeling approaches.

Video-based Characters and Free-viewpoint Video. Video-based synthesis tries to close the gap between photo-realistic videos and rendered controllable characters. First approaches were based on reordering existing video clips [Schödl and Essa 2002; Schödl et al. 2000]. Recent techniques enable video-based characters [Casas et al. 2014; Li et al. 2017b; Volino et al. 2014; Xu et al. 2011] and free-viewpoint video [Carranza et al. 2003; Collet et al. 2015a; Li et al. 2014; Zitnick et al. 2004] based on 3D proxies. Approaches for video-based characters either use dynamically textured meshes and/or image-based rendering techniques. Dynamic textures [Casas et al. 2014; Volino et al. 2014] can be computed by temporal registration of multi-view footage in texture space. The result is fully controllable, but unfortunately the silhouettes of the person match the coarse geometric proxy. The approach of Xu et al. [2011] synthesizes plausible videos of a human actors with new body motions and viewpoints. Synthesis is performed by finding a coherent sequence of frames that matches the specified target motion and warping these to the novel viewpoint. Li et al. [2017b] proposed an approach for sparse photo-realistic animation based on a single RGBD sensor. The approach uses model-guided texture synthesis based on weighted low-rank matrix completion. Most approaches for video-based characters require complex and controlled setups, have a high runtime, or do not generalize to challenging motions. In contrast, our approach is based on a single consumer-grade sensor, is efficient and generalizes well.

Learned Image-to-image Translation. Many problems in computer vision and graphics can be phrased as image-to-image mappings.

Recently, many approaches employ convolutional neural networks (CNNs) to learn the best mapping based on large training corpora. Techniques can be categorized into (variational) auto-encoders (VAEs) [Hinton and Salakhutdinov 2006; Kingma and Welling 2013], autoregressive models (AMs) [Oord et al. 2016], and conditional generative adversarial networks (cGANs) [Goodfellow et al. 2014; Isola et al. 2017; Mirza and Osindero 2014; Radford et al. 2016]. Encoder-decoder architectures [Hinton and Salakhutdinov 2006] are often used in combination with skip connections to enable feature propagation on different scales. This is similar to U-Net [Ronneberger et al. 2015]. CGANs [Isola et al. 2017; Mirza and Osindero 2014] have obtained impressive results on a wide range of tasks. Recently, high-resolution image generation has been demonstrated based on cascaded refinement networks [Chen and Koltun 2017] and progressively trained (conditional) GANs [Karras et al. 2018; Wang et al. 2018]. These approaches are trained in a supervised fashion based on paired ground truth training corpora. One of the main challenges is that paired corpora are often not available. Recent work on unpaired training of conditional GANs [Choi et al. 2018; Liu et al. 2017; Yi et al. 2017; Zhu et al. 2017] removes this requirement.

Generative Models for Humans. The graphics community has invested significant effort into realistically modeling humans. Parametric models for individual body parts, such as faces [Blanz and Vetter 1999; Li et al. 2017a], eyes [Bérard et al. 2014; Wood et al. 2016], teeth [Wu et al. 2016], hands [Romero et al. 2017], as well as for the entire body [Anguelov et al. 2005; Loper et al. 2015] have been proposed. Creating a complete photo-real clone of a human based on such models is currently infeasible. Recently, generative deep neural networks have been proposed to synthesize 2D imagery of humans. Approaches that convert synthetic images into photo-realistic imagery have been proposed for eyes [Shrivastava et al. 2017], hands [Mueller et al. 2017], and faces [Kim et al. 2018]. The approach of Ganin et al. [2016] performs gaze manipulation based on learned image-warping. In the context of entire human bodies, Zhu et al. [2018] generate novel imagery of humans from different view-points, but cannot control the pose. In addition, generative models for novel pose synthesis [Balakrishnan et al. 2018; Ma et al. 2017; Siarohin et al. 2018] have been proposed. The input to such networks are a source image, 2D joint detections, or a stick figure skeleton, and the target pose. Balakrishnan et al. [2018] train their network end-to-end on pairs of images that have been sampled from action video footage. Similar to our method, they also use a target body part-based representation as conditional input. However, their body part images are computed by a spatial transformation in 2D, while ours are obtained by rendering the 3D model animated by the target pose. Therefore, our approach yields more plausible body part images as the underlying geometry is correct. Ma et al. [2017] combine pose-guided image formation with a refinement network that is trained in an adversarial manner to obtain higher quality. Siarohin et al. [2018] introduce deformable skip connections to better deal with misalignments caused by pose differences. Lassner et al. [2017] proposed a 2D generative model for clothing that is conditioned on body shape and pose. Recently, a VAE for appearance and pose generation [Esser et al. 2018] has been proposed that enables training without requiring images of the same identity

with varying pose/appearance. In contrast to previous methods for pose synthesis, our approach employs dense synthetic imagery for conditioning and a novel adversarial loss that dynamically shifts its attention to the regions with the highest photometric residual error. This leads to higher quality results.

Human Performance Capture. The foundation for realistic video-based characters is reliable and high-quality human performance capture. The techniques used in professional production are based on expensive photometric stereo [Vlasic et al. 2009] or multi-view [Matusik et al. 2000; Starck and Hilton 2007; Waschbüsch et al. 2005] reconstruction setups. At the heart of most approaches is a skeletal motion prior [Gall et al. 2009; Liu et al. 2011; Vlasic et al. 2008], since it allows to reduce the number of unknown pose parameters to a minimum. A key for reliable tracking and automatic initialization of performance capture methods is the incorporation of 2D [Pishchulin et al. 2016; Wei et al. 2016] and 3D [Mehta et al. 2017a; Pavlakos et al. 2017; Zhou et al. 2016] joint detections of a pose estimation network into the alignment objective. Hybrid methods [Elhayek et al. 2015; Mehta et al. 2017b; Rosales and Sclaroff 2006] that rely on 2D as well as 3D detections combine both of these constraints for higher quality. High-quality reconstruction of human performances from two or more cameras is enabled by model-based approaches [Cagniart et al. 2010; De Aguiar et al. 2008; Wu et al. 2013]. Currently, the approaches that obtain the highest accuracy are multi-view depth-based systems, e.g., [Collet et al. 2015b; Dou et al. 2017, 2016; Wang et al. 2016]. Driven by the demand of VR and AR applications, the development of lightweight [Bogo et al. 2015, 2016; Helten et al. 2013; Yu et al. 2017; Zhang et al. 2014] solutions is an active area of research and recently even monocular human performance capture has been demonstrated [Xu et al. 2018].

3 METHOD

In this section we describe the technical details of our approach, which is outlined in Fig. 2. The main idea is to train a neural network that converts simple synthetic images of human body parts to an image of a human character that exhibits natural image characteristics such that the person appears (close to) photo-realistic. The main motivation for our approach is as follows: Photo-realistic rendering based on the conventional graphics pipeline requires expensive high-quality 3D human character modeling and the simulation of complex global light transport. In contrast, it is relatively simple and cheap to construct medium-quality synthetic human imagery based on commodity 3D reconstructions and direct illumination models. In order to compensate for the lack of photo-realism in such medium-quality images, we propose to train a generative deep neural network to bridge this gap, such that the person looks more realistic. In the following, we first describe the acquisition of suitable training data, followed by an in-depth explanation of the architecture of our Character-to-Image translation network.

3.1 Acquisition of the Training Corpus

In this section we describe how we acquire our training corpus. Our training corpus consists of pairs of rendered conditioning input images and the original image frames of a monocular training video (cf. Fig. 2). The conditioning images that are used as input
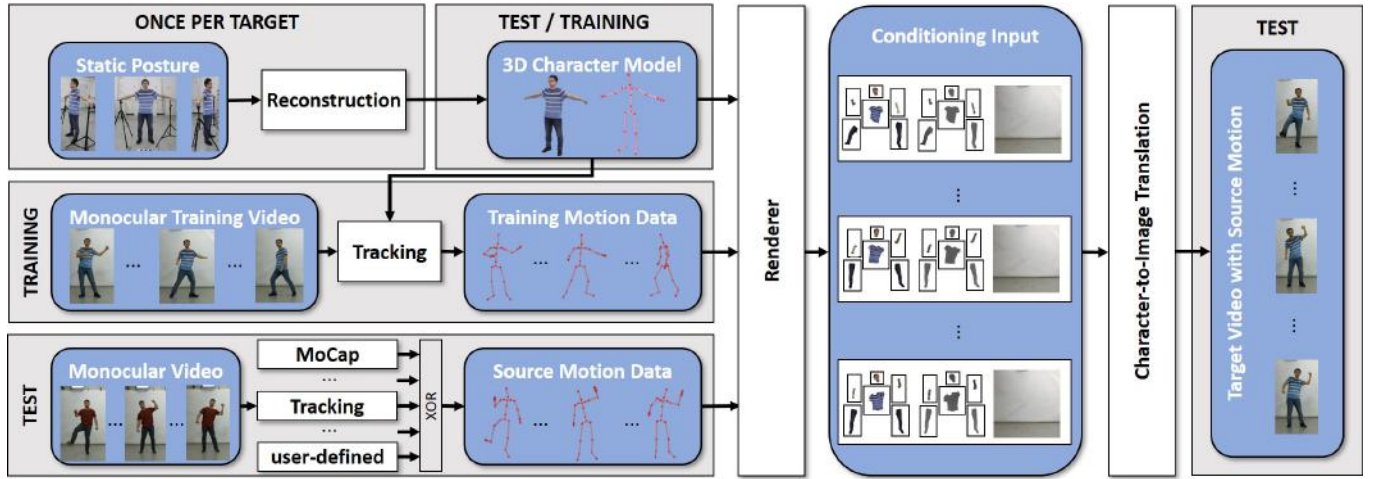
Fig. 2. Overview of our proposed approach for neural reenactment of human performances. Using a 3D character model of the target person, which is reconstructed from static posture images, we obtain the training motion data from the monocular training video based on an extension of the method of [Mehta et al. 2017b]. The motion data is then used to steer the 3D character model, from which we render color and depth body part images, which are fed into our Character-to-Image translation network in order to produce video-realistic output frames. At test time, the source motion data is used to render the conditional input (with the target 3D character model) in order to reenact the target human character. Note that the source motion can either come from a tracked monocular video, or from any other motion source (e.g., user-defined motion, or MoCap data).

for the network comprise individual depth and color renderings of six human body parts, i.e., *head, torso, left arm, right arm, left leg* and *right leg*, and an empty background image. In the following we describe how these images are obtained.

Data Acquisition. We capture our raw data using a Blackmagic video camera. For each actor, we record a motion sequence of approximately 8 minutes (about 12k frames). Similar to most learning based methods, our training data should resemble the distribution of real-world observations. Therefore, for each subject, we collect the training video such that it covers a typical range of general motions.

3D Character Model. Our method relies on a textured 3D template mesh of the target subject, which we obtain by capturing (around) 100 images of the target person in a static pose from different viewpoints, and then reconstructing a textured mesh using a state-of-the-art photogrammetry software[1]. The images need to be captured in such a way that a complete 3D reconstruction can be obtained, which is achieved by using a hand-held camera and walking around the subject. Afterwards, the template is rigged with a parameterized human skeleton model. More details on the template reconstruction can be found in [Xu et al. 2018].

Conditioning Input Images. In order to obtain the conditioning input images, we track the skeleton motion of the person in the training video using the skeletal pose tracking method of [Mehta et al. 2017b]. We extended this approach by a dense silhouette alignment constraint and a sparse feature alignment term based on a set of detected facial landmarks [Saragih et al. 2011]. Both additional terms lead to a better overlap of the model to the real-world data, thus simplifying the image-to-image translation task and leading to higher quality results. The output of the method is a sequence of

deformed meshes, all of them sharing the same topology (cf. *Training/Source Motion Data* in Fig. 2). We apply temporal smoothing to the trajectories of all vertices based on a Gaussian filter (with a standard deviation of 1 frame).

We then generate three different types of conditioning images by rendering synthetic imagery of the mesh sequence using the 3D character model, Specifically, we render (i) the textured mesh to obtain the color image $\mathcal{I}$, (ii) the depth image $\mathcal{D}$, and (iii) the binary semantic body part masks $\{\mathcal{M}_p \mid p \in \{1, ..., 6\}\}$. To render the binary semantic body part masks, we manually labeled the 6 body parts *head, torso, left arm, right arm, left leg, right leg* on the template mesh, and then generated a binary mask for each individual body part. Based on these masks, we extract the part-based color images $\{\mathcal{I}_p = \mathcal{I} \odot \mathcal{M}_p \mid p \in \{1, ..., 6\}\}$ and the part-based depth images $\{\mathcal{D}_p = \mathcal{D} \odot \mathcal{M}_p \mid p \in \{1, ..., 6\}\}$, where $\odot$ denotes the Hadamard product. Finally, we generate the conditioning input images by concatenating the 6 part-wise color images $\{\mathcal{I}_p\}$ and depth images $\{\mathcal{D}_p\}$, as well as the empty background image $\mathcal{B}$, along the channel axis (cf. *Conditioning Input* in Fig. 2), resulting in the input **X**. More details are described in the supplementary material.

3.2 Character-to-Image Translation

In this section we describe our *Character-to-Image* translation network (Fig. 3) in detail.

Motion transfer. In order to synthesize a video of the target person mimicking the motion of the source person, we transfer the per-frame skeletal pose parameters $\mathcal{P}_{\text{src}}$ (comprising the global location, global orientation and individual joint angles) from the source $\mathcal{S}_{\text{src}}$ to the target $\mathcal{S}_{\text{trgt}}$, where $\mathcal{S}_*$ denotes the skeleton model (comprising the skeleton topology and the bone lengths) [Xu et al. 2018]. Afterwards, we animate the virtual character of the target person

---

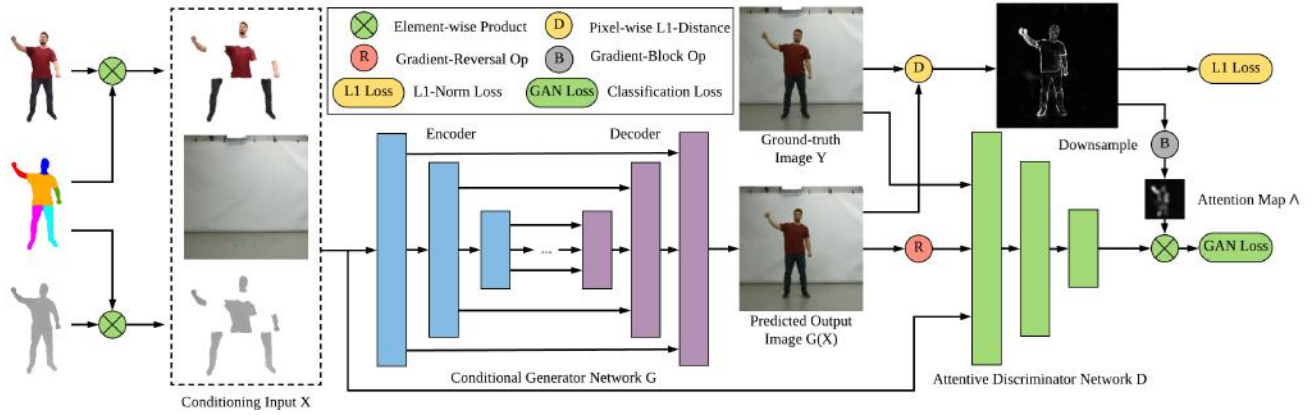[1]Agisoft Photoscan, http://www.agisoft.com/

Fig. 3. An illustration of the architecture of our Character-to-Image translation network. Both the encoder and the decoder of the generator have 8 layers with skip-connections between the same resolution levels. We use the gradient-reversal operation to reverse the sign of the gradients flowing into the generator and the gradient-block operation to stop backpropagation from the downsampled error map to avoid a cheating path.

with the transferred poses and finally render the target character to generate the conditioning images. We point out that we do not directly apply the source's skeletal pose parameters $\mathcal{P}_{\text{src}}$ to the target's skeleton $\mathcal{S}_{\text{trgt}}$ due to two issues: on the one hand, this would require that both skeletons have exactly the same structure, which may be overly restrictive in practical applications. On the other hand, and more importantly, differences in the rigging of the skeleton would lead to incorrect poses. To address these issues, we estimate the optimal pose $\mathcal{P}^*$ of the target person for each frame by solving the following inverse kinematics (IK) problem, which encourages that (corresponding) keypoints on both skeletons, including the joints and facial landmarks, match in 3D: s

$$\mathcal{P}^* = \underset{\mathcal{P}}{\text{argmin}} \sum_{k \in K} \left\| \mathbf{J}_k\left(\mathcal{S}_{\text{src}}, \mathcal{P}_{\text{src}}\right) - \mathbf{J}_k\left(\Phi(\mathcal{S}_{\text{trgt}}, \mathcal{S}_{\text{src}}), \mathcal{P}\right) \right\|_2^2 \quad (1)$$

Here K denotes the number of keypoints on the skeleton, $\mathbf{J}_k(\mathcal{S}_*, \mathcal{P}_*)$ is a function that computes the 3D position of the k-th keypoint given a skeleton $\mathcal{S}_*$ and a pose $\mathcal{P}_*$, and the function $\Phi(\mathcal{S}_{\text{trgt}}, \mathcal{S}_{\text{src}})$ returns the skeleton $\mathcal{S}_{\text{trgt}}$ after each individual bone length of $\mathcal{S}_{\text{trgt}}$ has been rescaled to match $\mathcal{S}_{\text{src}}$. To ensure that $\mathcal{S}_{\text{trgt}}$ is globally at a similar position as in the training corpus, we further translate $\mathcal{S}_{\text{trgt}}$ by a constant offset calculated with the root position of the skeleton in the test sequence and training sequence. Note that this IK step enables motion transfer between skeletons with different structures, and thus allows us to use motion data from arbitrary sources, such as artist designed motions or MoCap data, to drive our target character.

Network Architecture. Our Character-to-Image translation network (Fig. 3) consists of two competing networks, a conditional generator network $\mathbf{G}$ and an attentive discriminator network $\mathbf{D}$ based on the attention map $\boldsymbol{\Lambda}$.

The purpose of the generator network $\mathbf{G}$ (cf. *Generator Network* in Fig. 3) is to translate the input $\mathbf{X}$, which comprises synthetic color and depth renderings of 6 human body parts and the background image, as described in Sec. 3.1, to a photo-realistic image $\mathbf{G(X)}$ of

the full character. Our generator network internally consists of an encoder (cf. *Encoder* in Fig. 3) to compress the input into a low-dimensional representation, and a decoder (cf. *Decoder* in Fig. 3) to synthesize the photo-realistic image conditioned on the input character renderings. Each encoder layer comprises a 4×4 convolution with stride 2, batch normalization and a leaky Rectified Linear Unit (ReLU) activation. The decoder reverses the downsampling due to the encoding process with a symmetric 4×4 deconvolution operator with stride 2, which is fed into batch normalization, dropout and ReLU layers. In order to ensure that the final output of the network, i.e., the generated image (cf. *Predicted Output Image in Fig. 3)*, is normalized, we apply a hyperbolic tangent activation at the last layer. In addition, skip connections [Ronneberger et al. 2015] and a cascaded refinement strategy [Chen and Koltun 2017] are used in the generator network to propagate high-frequency details through the generator network. Both the input and output images are represented in a normalized color space, i.e., $[-1, -1, -1]$ and $[1, 1, 1]$ for black and white respectively.

The input to our attentive discriminator network $\mathbf{D}$ is the conditioning input $\mathbf{X}$, and either the predicted output image $\mathbf{G(X)}$ or the ground-truth image $\mathbf{Y}$ (cf. Fig. 3). The employed discriminator is motivated by the PatchGAN classifier [Isola et al. 2017], which we extend to incorporate an attention map to reweigh the classification loss. For more details, please refer to Fig. 3.

Objective Function. In order to achieve high-fidelity character-to-image translation, we base the objective on the expected value of our attentive conditional GAN loss $\mathcal{L}$ and on the ℓ_1-norm loss $\mathcal{L}_{\ell_1}$:

$$\min_{\mathbf{G}} \max_{\mathbf{D}} \mathbb{E}_{\mathbf{X}, \mathbf{Y}} \left[\mathcal{L}(\mathbf{G}, \mathbf{D}) + \lambda \mathcal{L}_{\ell_1}(\mathbf{G}) \right]. \quad (2)$$

The ℓ_1-distance of the synthesized image $\mathbf{G(X)}$ from the ground-truth image $\mathbf{Y}$ is introduced so that the synthesized output is sufficiently sharp while remaining close to the ground truth:

$$\mathcal{L}_{\ell_1}(\mathbf{G}) = \|\mathbf{Y} - \mathbf{G(X)}\|_1. \quad (3)$$

Attentive Discriminator. One of the technical novelties of our approach is an attentive discriminator to guide the translation process. As the attention map is used to reweight the discriminator loss, it is downsampled to the resolution of the discriminator loss map. Similar to the vanilla PatchGAN classifier [Isola et al. 2017], the discriminator $\mathbf{D}$ predicts a $W_D \times H_D$ map, in our case 30×30, where each value represents the probability of the receptive patch for being *real*, i.e., value 1 means that the discriminator decides that the patch is *real*, and the value 0 means it is *fake*. However, in contrast to the PatchGAN approach that treats all patches equally, we introduce the attention map $\mathbf{\Lambda}$, which has the same spatial resolution as the output of $\mathbf{D}$, i.e., $W_D \times H_D$. Its purpose is to shift the focus of areas that $\mathbf{D}$ relies on depending on some measure of importance, which will be discussed below. For the GAN loss, the discriminator $\mathbf{D}$ is trained to classify between *real* and *fake* images given the synthetic character input $\mathbf{X}$, while the generator $\mathbf{G}$ tries to fool the discriminator network by sampling images from the distribution of *real* examples:

$$\mathcal{L}(\mathbf{G}, \mathbf{D}) = \sum \left(\frac{\mathbf{\Lambda}}{\|\mathbf{\Lambda}\|} \odot \left(\log \mathbf{D}(\mathbf{X}, \mathbf{Y}) + \log \left(1 - \mathbf{D}(\mathbf{X}, \mathbf{G}(\mathbf{X})) \right) \right) \right). \quad (4)$$

Here, log is element-wise, by $\|\mathbf{\Lambda}\|$ we denote a scalar normalization factor that sums over all entries of $\mathbf{\Lambda}$, and the outer sum $\sum(\cdot)$ sums over all $W_D \times H_D$ elements due to the matrix-valued discriminator.

We have found that a good option for choosing the attention map $\mathbf{\Lambda}$ is to use the model's per-pixel ℓ_1-norm loss after downsampling it to the resolution $W_D \times H_D$. The idea behind this is to help $\mathbf{G}$ and $\mathbf{D}$ focus on parts where $\mathbf{G}$ still produces low quality results. For instance, $\mathbf{G}$ quickly learns to generate background (since, up to shadows or interactions, it is almost fixed throughout training and can be captured easily through skip-connections.), which leads to a very small ℓ_1-norm loss, and thus fools $\mathbf{D}$. However, there is no explicit mechanism to stop learning $\mathbf{G}$ in these regions, as $\mathbf{D}$ still tries to classify *real* from *fake*. These "useless" learning signals distract the gradients that are fed to $\mathbf{G}$ and even affect the learning of other important parts of the image. In this case, the ℓ_1-norm loss is a good guidance for GAN training.

Training. In order to train the proposed character-to-image translation network, we use approximately 12,000 training pairs, each of which consists of the original monocular video frame $\mathbf{Y}$ as well as the stack of conditioning images $\mathbf{X}$, as described in Sec. 3.1. For training, we set a hyper-parameter of $\lambda = 100$ for the loss function (Eq. 2), and use the Adam optimizer ($lr = 0.0002$, $\beta_1 = 0.5$, $\beta_2 = 0.99$) from which we run for a total of 70,000 steps with a batch size of 10. The number of layers in the generator was empirically determined. We implemented our model in TensorFlow [Abadi et al. 2015].

4 EXPERIMENTS

In order to evaluate our approach, we captured training sequences of 5 subjects performing various motions. Our training corpus consists of approximately 12,000 frames per person. We further recorded 5 separate source video sequences for motion retargeting. We perform pose tracking using Mehta et al. [2017b] to obtain the driving skeletal motion. Training takes about 12h for each subject on a resolution of 256×256 pixels using a single NVIDIA Tesla V100. Template

rendering takes less than 1ms/frame. A forward pass of our network takes about 68ms per frame.

The results presented in the following have been generated with a resolution of 256×256 pixels. Please note, that our approach is also able to generate higher resolution results of 512×512 pixels. Fig. 12 shows a few such examples, which took 24h to train. This further improves the sharpness of the results. Our dataset and code will be made publicly available.

In the following, we evaluate our method qualitatively and quantitatively. We provide a comparison to two state-of-the-art human image synthesis methods trained on our dataset. We also perform an ablation study to analyze the importance of each component of our proposed approach.

4.1 Qualitative Results

Figs. 4 and 5 show example reenactment results. We can see that our method synthesizes faithful imagery of human performances, in which the target characters precisely reenact the motion performed by the source subject. Our final results add a significant level of realism compared to the rendered character mesh from which the conditioning input is computed. Our method generalizes to different types of motions, such as waving, boxing, kicking, rotating and many gymnastic activities. Note that our method generates sharp images with a large amount of fine-scale texture detail. The facial features and the textures on the clothing are both well-preserved. Also note that our results accurately resemble the illumination condition in the real scene. Even the shading due to wrinkles on the garments and the shadows cast by the person onto the ground and wall are consistently synthesized.

Since a forward pass of our character-to-image translation network requires only 68 ms, it can also be used to generate new unseen poses based on interactive user control. Fig. 11 shows a few examples of a live editing session, where a user interactively controls the skeleton pose parameters of a real-world character using handle-based inverse kinematics. Please refer to the accompanying video for the complete video result.

Next, we compare our approach to the state-of-the-art human body image synthesis methods of Ma et al. [2018] and Esser et al. [2018], which we also trained on our dataset. For a fair comparison, we trained one person-specific network per subject for both the method of Ma et al. [2018] and of Esser et al. [2018], as done in our approach. Note that their methods take 2D joint detections as conditioning input. However, using the 2D joint detection of the source subject could make their methods produce inaccurate results, since during training the networks only see the skeleton of the target subject, which may have a different spatial extent than the source skeleton. Hence, to obtain a fair comparison, we use our transferred motion applied to the target subject (see Sec. 3.2) to generate the 2D joint positions. A qualitative comparison is shown in Fig. 6. We can see that the results of Ma et al. [2018] and Esser et al. [2018] exhibit more artifacts than the outputs produced by our approach. In particular, both Ma et al. [2018] and Esser et al. [2018] have difficulties in faithfully reproducing strongly articulated areas, such as the arms, as well highly textured regions, such as the face. In contrast, our method results in shaper reconstructions, preserves

Fig. 4. Qualitative reenactment results. The images in the first row ("Driving motion") are the input to our approach and define the motion, from which the conditioning images are generated. The images in the middle row ("Rendered Mesh") show the 3D mesh of the target person driven by the source motion. The images in the bottom row ("Output") are synthesized by our approach. Our method is able to faithfully transfer the input motion onto the target character for a wide range of human body performances. For full results we refer to the supplementary video.

more details in highly textured regions such as the face, and leads to fewer missing parts in strongly articulated areas, such as the arms.

4.2 User Study

In order to evaluate the user perception of the motion reenactment videos synthesized by our approach, we have conducted a user study that compares our results with the results obtained by Ma et al. [2018] and Esser et al. [2018]. To this end, we present pairs of short video clips, approximately of length between 4 and 25 seconds, to a total of 25 users, recruited mainly from Asia and Europe. We used a total of 14 sequence pairs, where for each pair exactly one sequence was produced by our method, whereas the other sequence in the pair was produced either by Ma et al. [2018] or by Esser et al. [2018], each of them being used 7 times. The users were asked to select for each pair the sequence which appears more realistic. In total, in 84% of the 25·14 = 350 ratings our method was preferred, whereas in 16% of the ratings one of the other methods was preferred.

4.3 Ablation Study

Next, we evaluate our design choices and study the importance of each component of our proposed approach. The reported errors are always computed on the foreground only, which we determine based on background subtraction.

Conditioning Input: First, we analyze the effect of different conditioning inputs. To this end, we compare the use of the following input modalities:

1) rendered skeleton (*skeleton*),
2) rendered RGB mesh and semantic masks (*RGB+mask*),

3) per-body-part rendered mesh RGB only (*RGB parts*)
4) rendered mesh RGBD and semantic masks (*RGBD+mask*),
5) per-body part rendered mesh RGBD (*RGBD parts*, **ours**).

In Figs. 8 and 7 we show the quantitative and qualitative results, respectively, where it is revealed that using the rendered RGB mesh in conjunction with semantic masks (*RGB+mask*, red dashed-dotted line) is superior compared to using only a sparse skeleton for conditioning (*skeleton*, solid blue line). Moreover, explicitly applying the semantic masks to the rendered images (*RGB parts*, dashed yellow line), i.e. breaking the image into its semantic parts, significantly improves the results. The results with depth information *RGBD+mask* (pink dotted line) and *RGBD parts* (black line) are consistently better than the RGB-only results. As can be also seen in Fig. 7 the depth information improves the image synthesis results in general. Moreover, in frames where body-part occlusions exist the depth information helps to reduce the artifacts. We also find that using only the part-based rendered mesh (*RGB parts*, dashed yellow line) in the conditioning input, without additional rendered depth images, is inferior to our final approach. We observe that the additional depth information (*ours*, solid black line) improves the quality of the synthesized imagery, since the network is better able to resolve which body parts are in front of each other. Hence, to achieve better robustness for the more difficult occlusion cases, we decided to use the depth channel in all other experiments. Tab. 1 also confirms the observations made above quantitatively in terms of both L2 error and SSIM comparison to ground truth. Note that there is a large improvement from the *RGBD+mask* to *Ours* (*RGBD parts*), which shows that the part-based representation plays a critical role in the
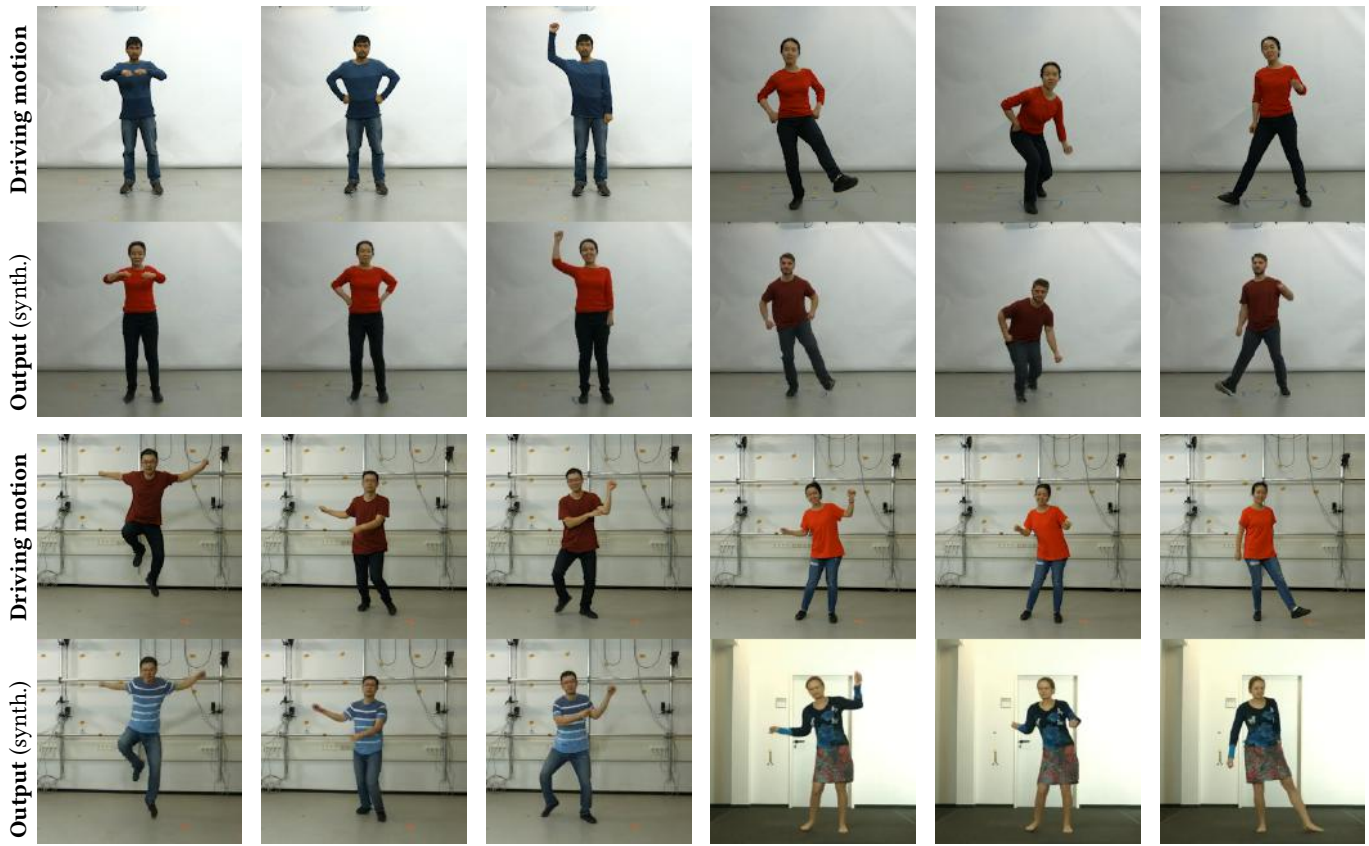
Fig. 5. Qualitative results for the reenactment of human characters. The images in the "Driving motion" row are used to track the motion, from which the conditioning images are generated. The images in the "Output" row are synthesized by our approach. Our method is able to faithfully transfer the input motions onto the target character for a wide range of human body performances. For full results we refer to the supplementary video.

improvement. The intuition behind this is that the separation of body parts can enforce the network to distinguish body parts in different ordinal layers, which helps to improve the results in the presence of occlusion. Comparing to semantic masks, the separation functionality of our part-based representation is more explicit. Comparing RGB parts with Ours (RGBD parts), we can see that, while the influence of the depth input to the visual quality in the example in the last row of Fig. 7 is relatively subtle, the depth information helps in general as indicated by the comparisons in both L2 error and SSIM metrics (see Table. 1). This shows that the part-based representation plays a critical role in the improvement. The intuition behind this is that the separation of body parts can enforce the network to distinguish body parts in different ordinal layers, which helps to improve the results in the presence of occlusion. The separation functionality of our part-based representation is more explicit than semantic masks. This is supported by our experimental results both quantitatively and qualitatively.

Attentive Discriminator: Moreover, we study the effect of using the proposed attention map mechanism in our attentive discriminator. In Fig. 9, it can be seen that using the attention GAN (solid black

Table 1. Quantitative evaluation. We calculate the L2 error and SSIM for the region of the person in the foreground in each image and report the mean value for the whole sequence. Our full approach obtains the best scores.

	L2 error	SSIM
skeleton	17.64387	0.60176
RGB+mask	16.82444	0.63102
RGB part	16.12499	0.64023
RGBD+mask	16.25977	0.64199
Ours	**15.67433**	**0.65328**

line) yields better results than a network trained without the attentive discriminator (dotted pink line). We show these improvements visually in Fig. 10. We also confirm this observation quantitatively in terms of L2 and SSIM errors, see Tab. 2.

5 LIMITATIONS

Despite the fact that we have presented compelling full body reenactment results for a wide range of motion settings, many challenges and open problems remain, which we hope will be addressed in the future. Synthesizing severely articulated motions is very challenging for all kinds of learned generative models, including ours, due
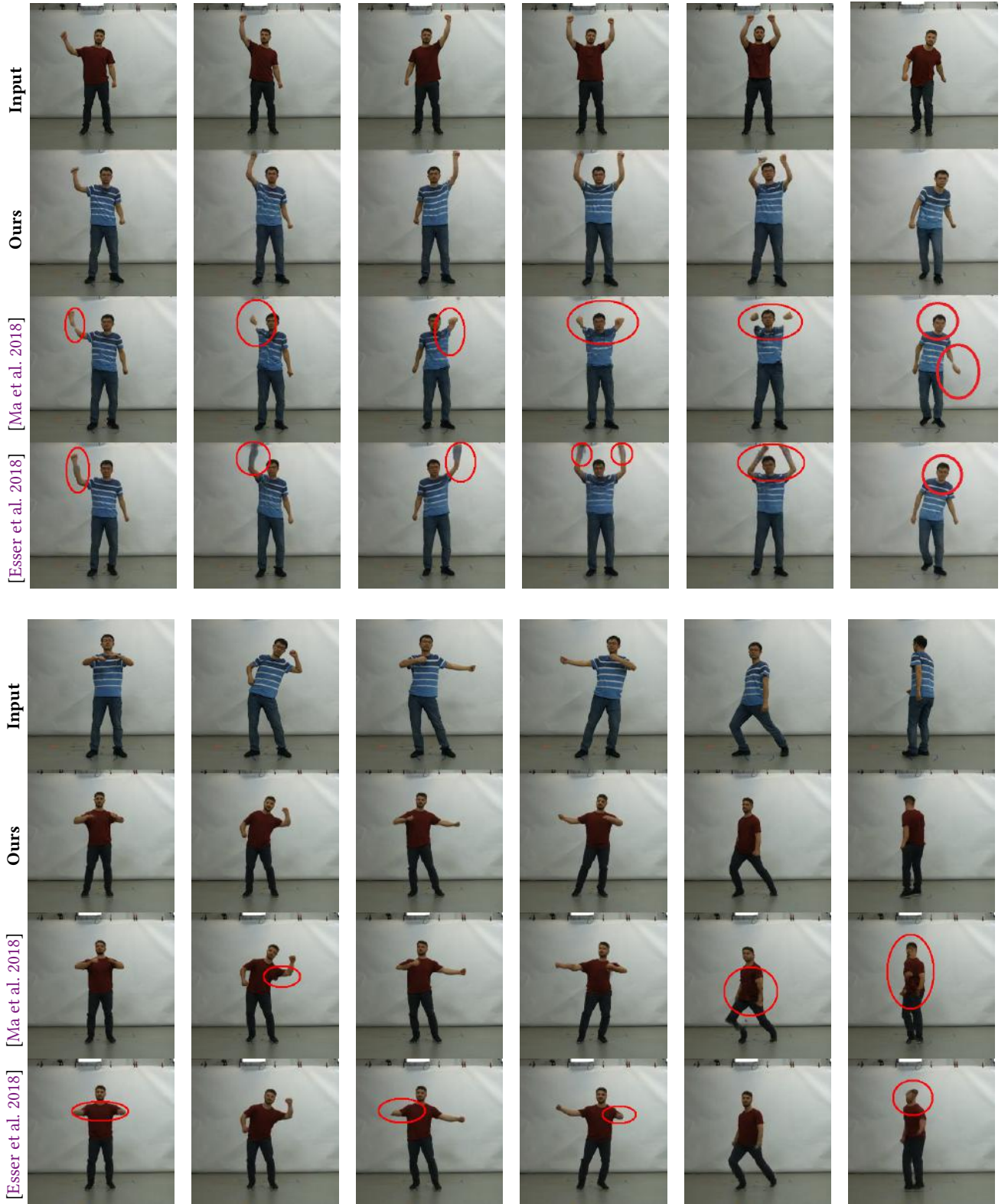
Fig. 6. Qualitative re-targeting comparison of our approach to the state-of-the-art human image synthesis methods of Ma et al. [2018] and Esser et al. [2018]. The encircled areas show regions of the body where the other methods clearly fail (missing parts of arms, wrongly shaped body parts etc.) in comparison to our approach. For the full results we refer to the supplementary video.
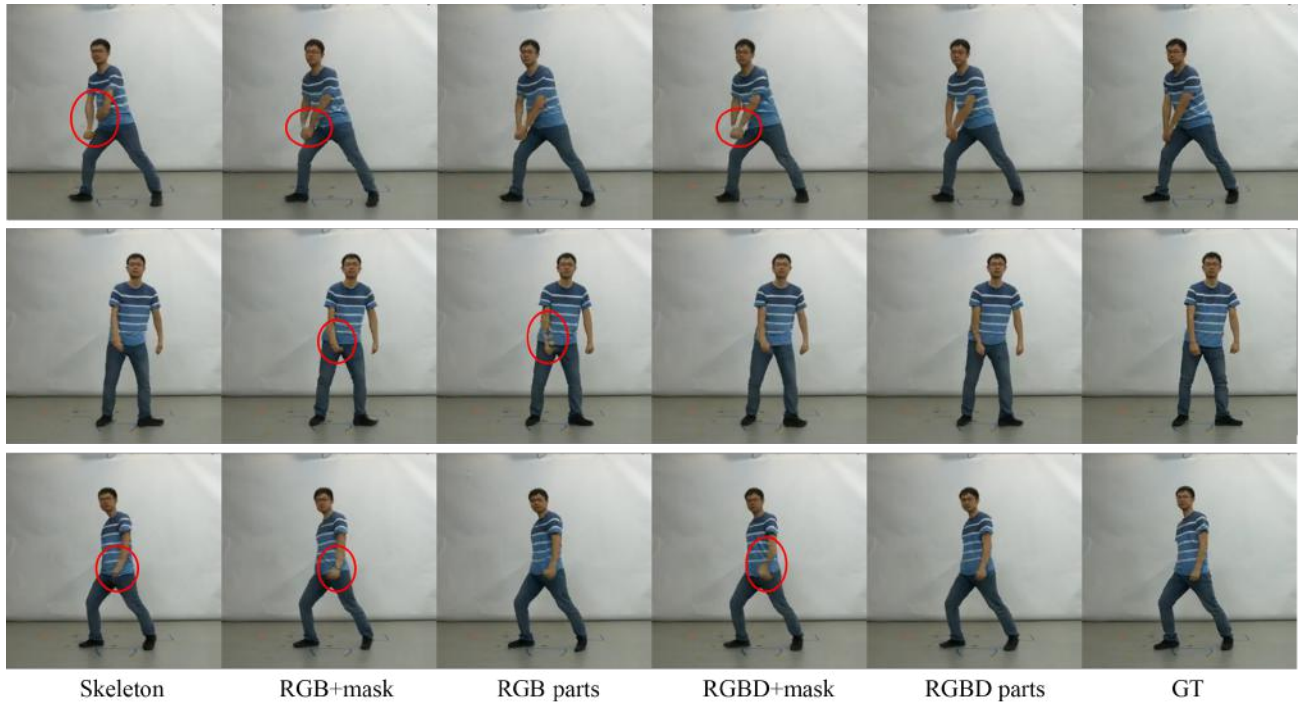
Skeleton · RGB+mask · RGB parts · RGBD+mask · RGBD parts · GT

Fig. 7. Qualitative validation of different conditioning inputs.
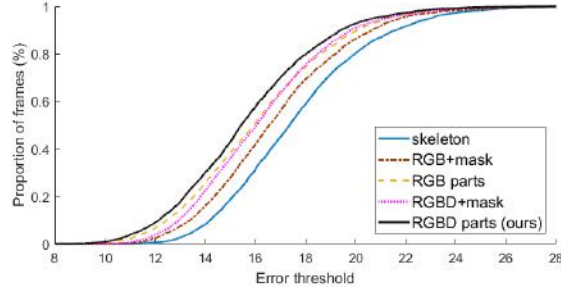
Fig. 8. Quantitative validation of different conditioning inputs. We show the proportion of frames on the vertical axis that have an error mean for the region of the person in the foreground smaller than the threshold on the horizontal axis. Our full approach consistently obtains the best results.
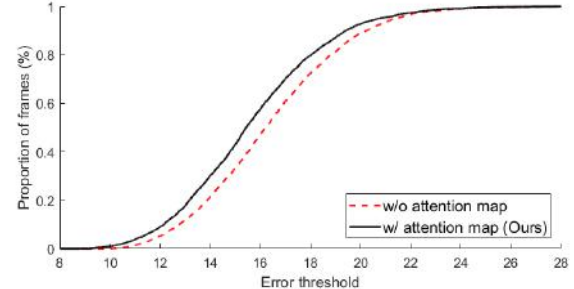
Fig. 9. Quantitative validation of attention map mechanism. We show the proportion of frames on the vertical axis that have an error mean for the region of the person in the foreground smaller than the threshold on the horizontal axis. Consistently better results are obtained with our attention map.

Table 2. Quantitative evaluation of attentive discriminator. We calculate the L2 error and SSIM for the region of the person in the foreground in each image and report the mean value for the whole sequence. Our full approach obtains the best scores.

	L2 error	SSIM
No attention	16.39865	0.64320
Ours	15.67433	0.65328

to multiple reasons: (i) articulated motions are highly non-linear, (ii) self-occlusions in human performances introduce discontinuities,

(iii) monocular tracking imperfections degrade the quality of the training corpus, and (iv) challenging poses are often underrepresented in the training data. Fig. 13 shows several typical failure cases of our method. Artifacts arise in particular at the end-effectors, e.g., hands or feet, since they undergo strong changes in spatial position and rotation. One potential solution could be to split the network into different branches for each body part, possibly into further sub-branches depending on the pose or view-point, while jointly learning a differentiable composition strategy.

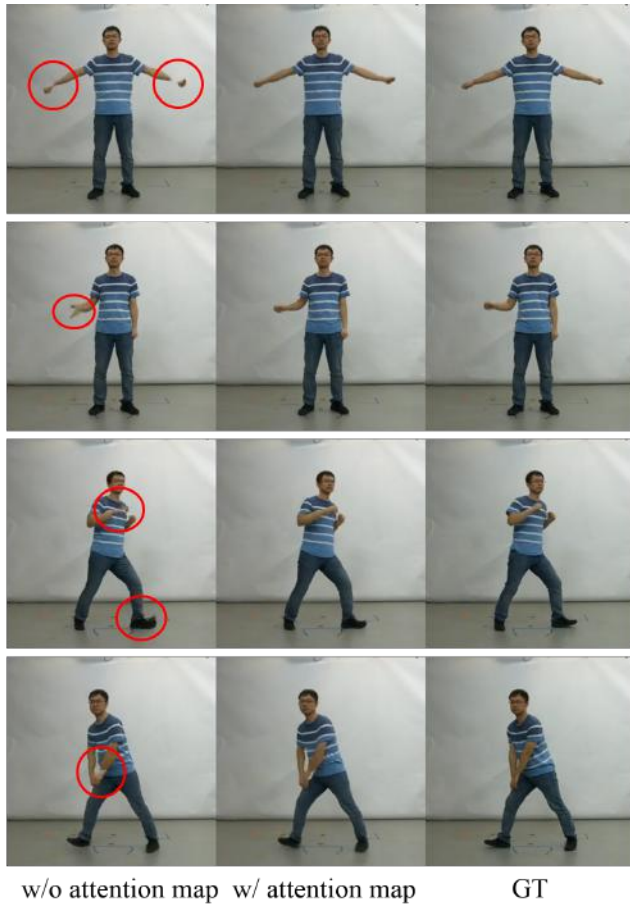w/o attention map   w/ attention map   GT

Fig. 10. Qualitative validation of attention map mechanism in our attentive discriminator. The comparison shows that our attention map based synthesis leads to a clear quality improvement (particularly visible in regions encircled in red), as seen by comparison to the ground truth.

Interactions with objects are challenging to synthesize for our as well as related techniques. This would require to jointly capture body pose as well as object position and shape at high accuracy, which is currently an unsolved problem.

Occasional local high-frequency artifacts are due to the specific choice of the used GAN architecture. Completely removing such local patterns, which are often observed in outputs of GANS, remains an open challenge.

Even though our results exhibit high quality, a temporally coherent synthesis of human performances that is free of temporal aliasing is highly challenging. This is also due to the non-linearities of articulated motion, which is particularly noticeable for fine-scale texture details.

We have conducted experiments on incorporating temporal information by concatenating several adjacent frames as input to the network. However, the results are not significantly better than with our proposed method. We still believe that a more sophisticated integration of temporal information might further improve the results.

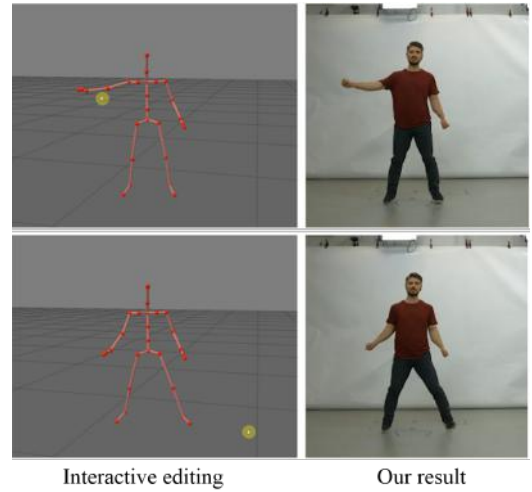Interactive editing   Our result

Fig. 11. Interactive editing results. Since a forward pass of our network only takes about 68ms, our approach allows for interactive user-controlled pose editing in video.
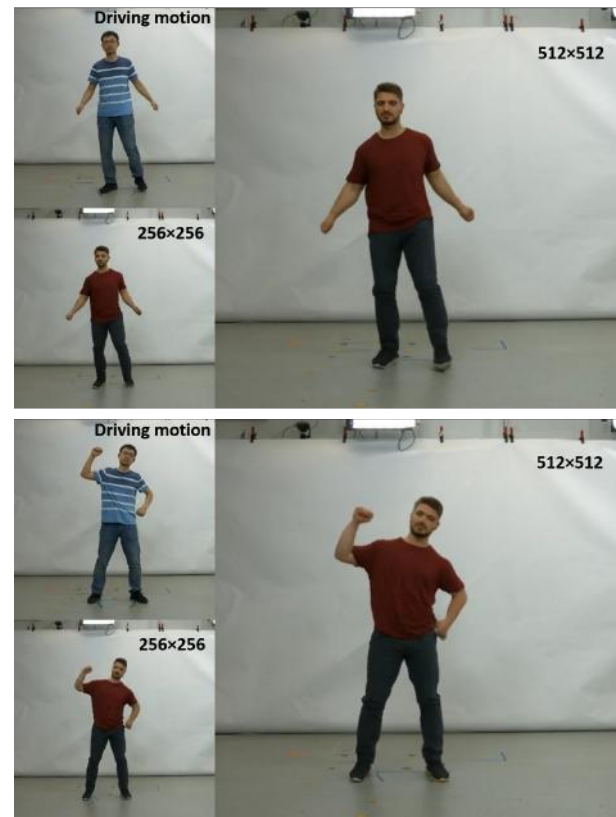

Fig. 12. Our approach can also synthesize high resolution images (512×512 pixels). This further improves the sharpness of the results but comes at the cost of an increased training time of the network.

For example, a space-time adversarial consistency loss, which operates on a small time slice, could help to alleviate local temporal flickering.
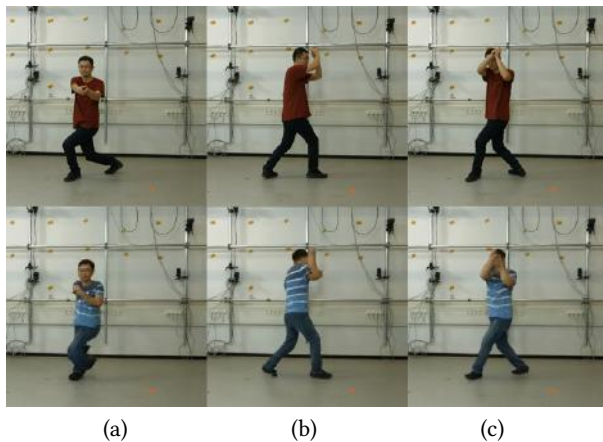
(a) (b) (c)

Fig. 13. Failure cases due to tracking error (a), rarely seen poses in the training corpus (b) and strong occlusion (c).

Another possible solution are recurrent network architectures, such as RNNs or LSTMs.

Currently, our networks are trained in a person-specific manner based on a long training sequence. Generalizing our approach, such that it works for arbitrary people given only a single reference image as input is an extremely challenging, but also very interesting direction for future work.

6 CONCLUSION

In this paper we have proposed a method for generating video-realistic animations of real humans under user control, without the need for a high-quality photorealistic 3D model of the human. Our approach is based on a part-based conditional generative adversarial network with a novel attention mechanism. The key idea is to translate computer graphics renderings of a medium-quality rigged model, which can be readily animated, into realistic imagery. The required person-specific training corpus can be obtained based on monocular performance capture. In our experiments, we have considered the reenactment of other people, where we have demonstrated that our approach outperforms the state-of-the-art in image-based synthesis of humans.

We believe this is a first important step towards the efficient rendition of video-realistic characters under user control. Having these capabilities is of high importance for computer games, visual effects, telepresence, and virtual and augmented reality. Another important application area is the synthesis of large fully annotated training corpora for the training of camera-based perception algorithms.

ACKNOWLEDGMENTS

REFERENCES

Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: Shape Completion and Animation of People. ACM Trans. Graph. 24, 3 (July 2005), 408–416. https://doi.org/10.1145/1073204.1073207

Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Fredo Durand, and John Guttag. 2018. Synthesizing Images of Humans in Unseen Poses. In Computer Vision and Pattern Recognition (CVPR), 2018.

Pascal Bérard, Derek Bradley, Maurizio Nitti, Thabo Beeler, and Markus Gross. 2014. High-quality Capture of Eyes. ACM Trans. Graph. 33, 6 (2014), 223:1–223:12.

Volker Blanz and Thomas Vetter. 1999. A Morphable Model for the Synthesis of 3D Faces. In Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '99). ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 187–194.

Federica Bogo, Michael J. Black, Matthew Loper, and Javier Romero. 2015. Detailed Full-Body Reconstructions of Moving People from Monocular RGB-D Sequences. In International Conference on Computer Vision (ICCV). 2300–2308.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. 2016. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In European Conference on Computer Vision (ECCV).

Cedric Cagniart, Edmond Boyer, and Slobodan Ilic. 2010. Free-form mesh tracking: a patch-based approach. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 1339–1346.

Joel Carranza, Christian Theobalt, Marcus A. Magnor, and Hans-Peter Seidel. 2003. Free-viewpoint Video of Human Actors. ACM Trans. Graph. 22, 3 (July 2003).

Dan Casas, Marco Volino, John Collomosse, and Adrian Hilton. 2014. 4D Video Textures for Interactive Character Appearance. Comput. Graph. Forum 33, 2 (May 2014), 371–380. https://doi.org/10.1111/cgf.12296

Qifeng Chen and Vladlen Koltun. 2017. Photographic Image Synthesis with Cascaded Refinement Networks. 1520–1529. https://doi.org/10.1109/ICCV.2017.168

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In Computer Vision and Pattern Recognition (CVPR), 2018.

Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015a. High-quality Streamable Free-viewpoint Video. ACM Trans. Graph. 34, 4, Article 69 (July 2015), 13 pages. https://doi.org/10.1145/2766945

Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015b. High-quality streamable free-viewpoint video. ACM Transactions on Graphics (TOG) 34, 4 (2015), 69.

Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. 2008. Performance capture from sparse multi-view video. In ACM Transactions on Graphics (TOG), Vol. 27. ACM, 98.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An Open Urban Driving Simulator. In Proceedings of the 1st Annual Conference on Robot Learning. 1–16.

Mingsong Dou, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. 2017. Motion2Fusion: Real-time Volumetric Performance Capture. ACM Trans. Graph. 36, 6, Article 246 (Nov. 2017), 16 pages.

Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, and Shahram Izadi. 2016. Fusion4D: Real-time Performance Capture of Challenging Scenes. ACM Trans. Graph. 35, 4, Article 114 (July 2016), 13 pages. https://doi.org/10.1145/2897824.2925969

Ahmed Elhayek, Edilson de Aguiar, Arjun Jain, Jonathan Tompson, Leonid Pishchulin, Micha Andriluka, Chris Bregler, Bernt Schiele, and Christian Theobalt. 2015. Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3810–3818.

Patrick Esser, Ekaterina Sutter, and Björn Ommer. 2018. A Variational U-Net for Conditional Appearance and Shape Generation. In Conference on Computer Vision and Pattern Recognition (CVPR).

Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. 2009. Motion capture using joint skeleton tracking and surface estimation. In Computer Vision and Pattern Recognition, 2009. CVPR 2009.

IEEE Conference on. IEEE, 1746–1753.

Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor S. Lempitsky. 2016. DeepWarp: Photorealistic Image Resynthesis for Gaze Manipulation. In *ECCV*.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets.

Thomas Helten, Meinard Muller, Hans-Peter Seidel, and Christian Theobalt. 2013. Real-Time Body Tracking with One Depth Camera and Inertial Sensors. In *The IEEE International Conference on Computer Vision (ICCV)*.

Geoffrey E. Hinton and Ruslan Salakhutdinov. 2006. Reducing the Dimensionality of Data with Neural Networks. *Science* 313, 5786 (July 2006), 504–507. https://doi.org/10.1126/science.1127647

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. 5967–5976. https://doi.org/10.1109/CVPR.2017.632

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation.

H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, N. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. 2018. Deep Video Portraits. *ACM Transactions on Graphics 2018 (TOG)* (2018).

Diederik P. Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. *CoRR* abs/1312.6114 (2013). http://dblp.uni-trier.de/db/journals/corr/corr1312.html#KingmaW13

Christoph Lassner, Gerard Pons-Moll, and Peter V. Gehler. 2017. A Generative Model of People in Clothing. In *Proceedings IEEE International Conference on Computer Vision (ICCV)*. IEEE, Piscataway, NJ, USA. http://files.is.tuebingen.mpg.de/classner/gp/

Guannan Li, Yebin Liu, and Qionghai Dai. 2014. Free-viewpoint Video Relighting from Multi-view Sequence Under General Illumination. *Mach. Vision Appl.* 25, 7 (Oct. 2014), 1737–1746. https://doi.org/10.1007/s00138-013-0559-0

Kun Li, Jingyu Yang, Leijie Liu, Ronan Boulic, Yu-Kun Lai, Yebin Liu, Yubin Li, and Eray Molla. 2017b. SPA: Sparse Photorealistic Animation Using a Single RGB-D Camera. *IEEE Trans. Cir. and Sys. for Video Technol.* 27, 4 (April 2017), 771–783. https://doi.org/10.1109/TCSVT.2016.2556419

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017a. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics* 36, 6 (Nov. 2017), 194:1–194:17. Two first authors contributed equally.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised Image-to-Image Translation Networks.

Yebin Liu, Carsten Stoll, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. 2011. Markerless motion capture of interacting characters using multi-view image segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 1249–1256.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. 2018. Disentangled Person Image Generation. (2018).

Liqian Ma, Qianru Sun, Xu Jia, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. 2017. Pose Guided Person Image Generation.

Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J Gortler, and Leonard McMillan. 2000. Image-based visual hulls. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 369–374.

Dushyant Mehta, Helge Rhodin, Dan Casas, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. 2017a. Monocular 3D Human Pose Estimation Using Transfer Learning and Improved CNN Supervision. In *3DV*.

Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017b. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2017)* 36, 4, 14.

Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. (2014). https://arxiv.org/abs/1411.1784 arXiv:1411.1784

Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2017. GANerated Hands for Real-time 3D Hand Tracking from Monocular RGB. *CoRR* abs/1712.01057 (2017).

Aäron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. 2016. Conditional Image Generation with PixelCNN Decoders. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., USA, 4797–4805. http://dl.acm.org/citation.cfm?id=3157382.3157633

Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. 2017. Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. In *Computer Vision and Pattern Recognition (CVPR)*.

Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. 2016. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (Nov. 2017), 245:1–245:17. http://doi.acm.org/10.1145/3130800.3130883

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Rómer Rosales and Stan Sclaroff. 2006. Combining generative and discriminative models in a framework for articulated pose estimation. *International Journal of Computer Vision* 67, 3 (2006), 251–276.

Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. 2011. Deformable Model Fitting by Regularized Landmark Mean-Shift. 91, 2 (2011), 200–215. https://doi.org/10.1007/s11263-010-0380-4

Arno Schödl and Irfan A. Essa. 2002. Controlled Animation of Video Sprites. In *Proceedings of the 2002 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '02)*. ACM, New York, NY, USA, 121–127. https://doi.org/10.1145/545261.545281

Arno Schödl, Richard Szeliski, David H. Salesin, and Irfan Essa. 2000. Video Textures. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 489–498. https://doi.org/10.1145/344779.345012

A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. 2017. Learning from Simulated and Unsupervised Images through Adversarial Training. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2242–2251. https://doi.org/10.1109/CVPR.2017.241

Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuiliere, and Nicu Sebe. 2018. Deformable GANs for Pose-based Human Image Generation. In *CVPR 2018*.

Jonathan Starck and Adrian Hilton. 2007. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications* 27, 3 (2007), 21–31.

Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. 2008. Articulated mesh animation from multi-view silhouettes. In *ACM Transactions on Graphics (TOG)*, Vol. 27. ACM, 97.

Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. 2009. Dynamic shape capture using multi-view photometric stereo. *ACM Transactions on Graphics (TOG)* 28, 5 (2009), 174.

Marco Volino, Dan Casas, John Collomosse, and Adrian Hilton. 2014. Optimal Representation of Multiple View Video. In *Proceedings of the British Machine Vision Conference*. BMVA Press.

Ruizhe Wang, Lingyu Wei, Etienne Vouga, Qixing Huang, Duygu Ceylan, Gerard Medioni, and Hao Li. 2016. Capturing Dynamic Textured Surfaces of Moving Targets. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs.

Michael Waschbüsch, Stephan Würmlin, Daniel Cotting, Filip Sadlo, and Markus Gross. 2005. Scalable 3D video of dynamic scenes. *The Visual Computer* 21, 8-10 (2005), 629–638.

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional Pose Machines. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

E. Wood, T. Baltrusaitis, L. P. Morency, P. Robinson, and A. Bulling. 2016. A 3D morphable eye region model for gaze estimation. In *ECCV*.

Chenglei Wu, Derek Bradley, Pablo Garrido, Michael Zollhöfer, Christian Theobalt, Markus Gross, and Thabo Beeler. 2016. Model-based Teeth Reconstruction. *ACM Trans. Graph.* 35, 6, Article 220 (2016), 220:1–220:13 pages.

Chenglei Wu, Carsten Stoll, Levi Valgaerts, and Christian Theobalt. 2013. On-set Performance Capture of Multiple Actors With A Stereo Camera. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2013)*, Vol. 32. 161:1–161:11. https://doi.org/10.1145/2508363.2508418

Feng Xu, Yebin Liu, Carsten Stoll, James Tompkin, Gaurav Bharaj, Qionghai Dai, Hans-Peter Seidel, Jan Kautz, and Christian Theobalt. 2011. Video-based Characters: Creating New Human Performances from a Multi-view Video Database. In *ACM SIGGRAPH 2011 Papers (SIGGRAPH '11)*. ACM, New York, NY, USA, Article 32, 10 pages. https://doi.org/10.1145/1964921.1964927

Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. 2018. MonoPerfCap: Human Performance Capture from Monocular Video. *ACM Transactions on Graphics* (2018). http://gvv.mpi-inf.mpg.de/projects/wxu/MonoPerfCap

Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. 2868–2876. https://doi.org/10.1109/ICCV.2017.310

T. Yu, K. Guo, F. Xu, Y. Dong, Z. Su, J. Zhao, J. Li, Q. Dai, and Y. Liu. 2017. BodyFusion: Real-Time Capture of Human Motion and Surface Geometry Using a Single Depth

Camera. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 910–919. https://doi.org/10.1109/ICCV.2017.104

Qing Zhang, Bo Fu, Mao Ye, and Ruigang Yang. 2014. Quality Dynamic Human Body Modeling Using a Single Low-cost Depth Camera. In *CVPR*. IEEE, 676–683.

Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. 2016. Deep Kinematic Pose Regression.

Hao Zhu, Hao Su, Peng Wang, Xun Cao, and Ruigang Yang. 2018. View Extrapolation of Human Body from a Single Image. In *CVPR 2018*.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. 2242–2251. https://doi.org/10.1109/ICCV.2017.244

C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. 2004. High-quality video view interpolation using a layered representation. In *ACM Transactions on Graphics (TOG)*, Vol. 23. ACM, 600–608.