Hierarchical Generation of Human Pose With Part-Based Layer Representation

Xian Wu, Chen Li, Shi-Min Hu, *Senior Member, IEEE*, and Yu-Wing Tai, *Senior Member, IEEE*

***Abstract*— Human pose transfer has been becoming one of the emerging research topics in recent years. However, state-of-the-art results are still far from satisfactory. One main reason is that these end-to-end methods are often blindly trained without the semantic understanding of its content. In this paper, we propose a novel method for human pose transfer with consideration of the semantic part-based representation of a human. In particular, we propose to segment the human body into multiple parts, and each of them represents a semantic region of a human. With the proposed part-based layer generators, a high-quality result is guaranteed for each local semantic region. We design a three-stage hierarchical framework to fuse local representations into the final result in a coarse-to-fine manner, which provides adaptive attention for global consistency and local details, respectively. Via exploiting spatial guidance from 3D human model through the framework, our method can naturally handle the ambiguity of self-occlusions which always causes artifacts in previous methods. With semantic-aware and spatial-aware representations, our method outperforms previous approaches quantitatively and qualitatively in better handling self-occlusions, fine detail preservation/synthesis and a higher resolution result.**

***Index Terms*— Human pose transfer, part-based layer representation, self-occlusion, coarse-to-fine generation.**

I. Introduction

SYNTHESIZING human images under specific settings is an interesting but challenging problem. Owing to the recent development of deep learning techniques and generative adversarial networks (GANs) [1], many works have been dedicated to this area, such as appearance transfer [2], [3], image completion [4], and novel view synthesis [5], [6].

In this paper, we focus on one of the most important tasks in human image synthesis, pose transfer, which aims to transfer the source human image to a certain target pose while preserving one's identity and appearance properly. Human pose transfer allows for many industrial applications, for example, motion video generation. Beyond one's imagination, everyone could dance like a pop star or do some actions that he/she never has done. Moreover, generation of human images in different poses can act as a data augmentation method relieving the time-consuming manual annotations, which speeds up the development of a wide range of human-centric vision tasks.

Many previous pose transfer approaches propose to warp the source image through a spatial transformation module by replacing the extracted source pose feature with the target pose feature [7]–[9]. Some other methods [10], [11] reconstruct the human surface textures and try to inpaint the missing regions in the target pose. However, the lack of semantic information and 3D spatial relationship of different body parts makes their results far from satisfactory, especially where the visibility changes from source pose to target pose. By considering the semantic information of the human body, some interior priors, such as its symmetry, can be exploited more to better preserve/synthesize local content, *e.g.*, for the missing regions in target pose. Besides, these methods only represent the target pose as 2D landmarks and it is difficult to precisely identify the visibility between body parts in transfer result. Such occlusion ambiguity further causes unwished artifacts around the self-occluded regions. Moreover, previous methods are always formulated in a single holistic framework with limited input resolution, so high-quality results of local body components with rich identity and appearance details, *e.g.* faces, are hard to be generated.

In order to address these issues, we design a novel three-stage hierarchical human pose transfer framework by utilizing the semantic part-based layer representation which is illustrated in Fig. 1. We roughly synthesize a coarse transfer result in the first, then generate local representation with fine details for each important body component as an intermediate result, and fuse these results with spatial guidance from 3D representation and produce the final pose transfer image in a coarse-to-fine manner. Through the intermediate part-based layer representation, we individually formulate the synthesis of important body parts, namely *face*, *arm* and *leg*. The part-based layer representation, which is aware of semantic information, not only preserves the facial identity and cloth textures during the generation but also synthesizes fine details for each important part regardless of the occlusions. To fuse these local representations in correct spatial order, we exploit 3D target pose to tackle the limitations in self-occlusion handling of previous methods with 2D target pose. The target pose is represented as several informative maps to guide the entire generation framework to ensure the self-occlusions are handled properly. Via the sematic-aware and spatial-aware representations, the generation of high-quality local content of the human body is guaranteed even when some areas are

Manuscript received July 6, 2020; revised July 17, 2021; accepted August 9, 2021. Date of publication September 15, 2021; date of current version September 20, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Junsong Yuan. *(Corresponding author: Yu-Wing Tai.)*

Xian Wu and Shi-Min Hu are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China.

Chen Li is with Weixin Group, Tencent, Shenzhen 518054, China.

Yu-Wing Tai is with Kwai Inc., Palo Alto, CA 94306 USA, and also with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology (HKUST), Hong Kong (e-mail: yuwing@gmail.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2021.3108023, provided by the authors.

Digital Object Identifier 10.1109/TIP.2021.3108023

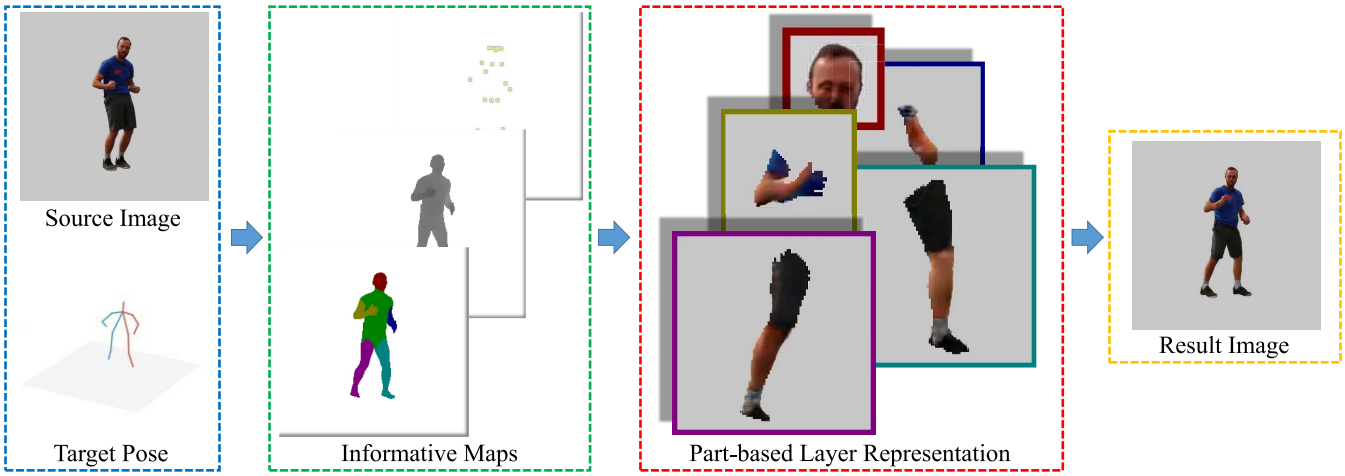


Fig. 1. We propose a part-based layer representation, which generates high-quality results for all important body parts individually with an awareness of their semantic information. Spatial relationships between these local body parts are represented as several informative maps to guide the fusion of these local representations in correct order to produce the final high-resolution pose transfer image.

missing in the source image and makes it a reality to apply our method on high-resolution images.

Extensive experiments show that our method outperforms previous state-of-the-art human pose transfer techniques [9], [12], [13] qualitatively and quantitatively. Our major contributions are summarized as three folds:

- to the best of our knowledge, we are the first to introduce part-based layer representation for generating different body components individually to address the human pose transfer problem;
- we apply the semantic-aware and spatial-aware representation into this task, which can be adapted to various 3D models, such as SMPL [14] and 3D volume;
- we propose a three-stage hierarchical generative framework for fusing the local body parts to compose the final result with fine details in a high-resolution.

II. Related Work

A. Human Pose Transfer

The purpose of this task is to transfer the input human into the target pose while keeping the appearance consistent. Recently, many deep learning based methods have been proposed to solve this problem. Ma *et al.* [15] present a two-stage coarse-to-fine method for pose-guided human synthesis. Si *et al.* [16] introduce the multi-stage adversarial loss to synthesize both the foreground and the background. Esser *et al.* [17] combine a U-Net generator and a variational autoencoder to encode the shape and the appearance, respectively. Ma *et al.* [18] disentangle the pose, appearance and background to synthesize the human image arbitrarily. Several methods apply geometric transformations to the local features for fine details based on masks of body subparts [9], [19] or human parsing [7]. Besides that, some methods [20]–[25] focus on generating human motion video by the guidance of 2D pose sequence. Neverova *et al.* [11] first use the DensePose [26] to guide the human synthesis, which provides the dense correspondence between the image and the 3D human surface. Grigorev *et al.* [10] improve this method by inpainting the coordinates of the textures instead of the colors for smoother generated results. Liu *et al.* [27] leverage a textured 3D character model to render the human actor video. However, this method needs to train the character model for each person and the 3D motion data is hard to acquire. Li *et al.* [8] use 2D keypoints to predict the dense appearance flow for human pose transfer but cannot solve the ambiguity caused by the lack of 3D information. Liquid Warping GAN [12] utilizes HMR [28] to construct the SMPL model [14] for the source and the target images, and then calculate the transformation flow based on the two correspondence maps, which achieves impressive results. However, the 2D correspondence map cannot well-define the self-occluded areas between different body parts, while our method can solve it by semantic-aware and spatial-aware part-based representation.

B. 3D Human Reconstruction

3D human model has shown to be an advantageous representation and with much potential in human synthesis techniques [3], [12], [29]. Reconstructing 3D human shape from the image is a challenging task. Most previous works use parametric body models to represent the 3D human shape, such as the SMPL [14] model, and then predict the parameters of the model. SMPLify [30] estimates the body shape represented by the approximate capsules through minimizing the distance between the projected SMPL [14] model joints and the detected 2D joints. HMR [28] proposes an end-to-end deep learning method to directly predict the SMPL parameters from a single image by optimizing the objective function, which is a combination of the 2D joints error, the 3D joints error, the 3D parameters error and the adversarial loss. Recently, BodyNet [31] and DeepHuman [32] leverage the 3D volume to represent the human body shape without using a parametric

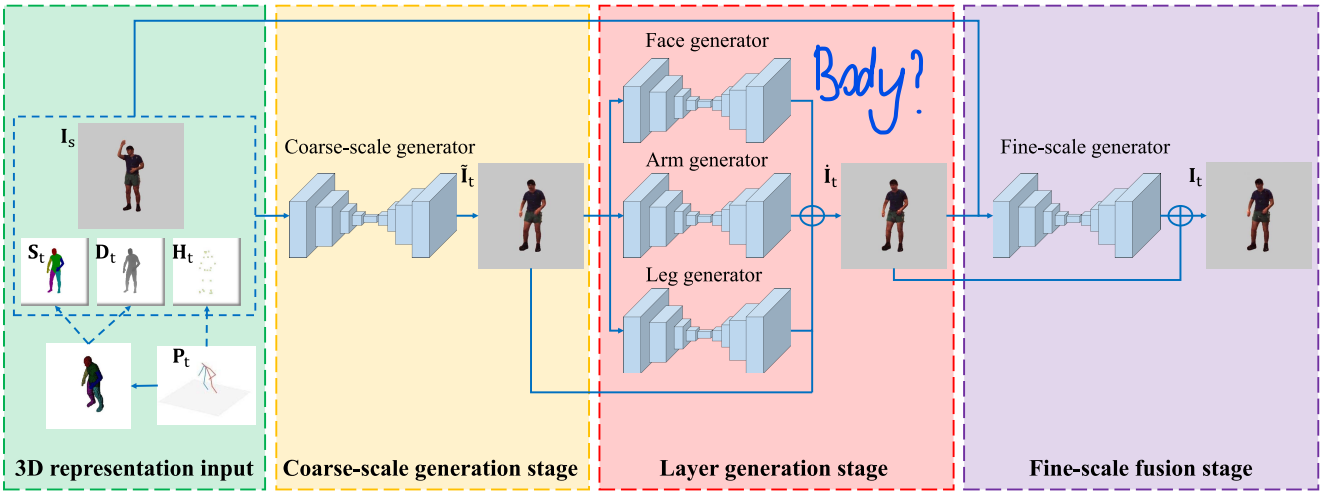


Fig. 2. The pipeline of proposed three-stage hierarchical human pose transfer framework. We use three intermediate part-based layer representations to ensure the high-quality synthesis results for important body components and exploit informative maps from 3D target pose as guidance for coarse-scale generation stage and fine-scale fusion stage to handle the self-occlusion problem properly.

model and demonstrate that the volumetric body shape can be predicted straightforward. We demonstrate that our framework is general to various 3D human representations, including parametric SMPL [14] and 3d volume.

C. Local Generator and Discriminator

Many GAN-based synthesis methods [33], [34] are applying local generators or discriminators to refine the local details. Li and Wand [35] propose the Patch-GAN which determines the local patch to be fake or real for texture synthesis. Several works [13], [36], [37] apply the Patch-GAN for the image-to-image translation task. Iizuka *et al.* [38] and Li *et al.* [39] introduce the local discriminator in the inpainted region for the image completion task to achieve local and global consistency. LADN [40] presents multiple local style discriminators specialized for different facial landmarks to address the facial makeup and de-makeup issue. Chan *et al.* [22] and Wu *et al.* [4] use face GAN to refine the facial areas after the human synthesis process. Our experiments have shown that our diverse part-based generators and discriminators are able to produce fine local details and deal with the self-occlusions through the hierarchical generation.

III. Approach

Our goal is to transfer a high-resolution human image into the target pose while preserving the human identity and body appearance with fine-scale details. To achieve such pleasant results, we propose the semantic part-based layer representation for each body component to better preserve and generate the detail textures. A three-stage hierarchical generation framework is designed in a coarse-to-fine manner to fuse all the part-based representations into a final high-resolution, 1024 $\times$ 1024, result. Through guiding the framework with an existing 3D human model, such as 3D volume [31] or SMPL [14], our method handles the self-occlusion problem better than previous approaches.

A. Overview

We formulate the pose transfer problem as an image-to-image translation task [13], [36], [37] and illustrate the pipeline of our hierarchical human pose transfer network in Fig. 2. Our method takes one source human image $\mathbf{I}_s$ in 1024 $\times$ 1024 resolution and one target 3D pose $\mathbf{P}_t$ as inputs. Through existing 3D human modeling representations, such as 3D volume or SMPL [14], we represent the target 3D pose as three corresponding informative maps, namely a human parsing map $\mathbf{S}_t$, a depth map $\mathbf{D}_t$, and 2D pose heatmaps $\mathbf{H}_t$.

In the first **coarse-scale generation stage**, we concatenate the source image $\mathbf{I}_s$ with three informative maps, $\mathbf{S}_t$, $\mathbf{D}_t$, $\mathbf{H}_t$, together to generate a coarse-scale transfer result $\widetilde{\mathbf{I}}_t$ in 512 $\times$ 512 resolution. After the coarse generation, three part-based layer generators are individually applied to produce high-quality results $\{\mathbf{I}_t^n\}_{n=1,\ldots,5}$ for all the important body components, namely one face, two arms and two legs, respectively. This **layer generation stage** synthesizes the complete body parts without considering the occlusions appearing in target pose, and its results $\{\mathbf{I}_t^n\}$ preserve the facial identity and cloth textures for the source person. Finally, we employ a **fine-scale fusion stage** to fuse the five part-based human layers $\{\mathbf{I}_t^n\}$ with the coarse result $\widetilde{\mathbf{I}}_t$ to achieve the whole body consistency and generate a 1024 $\times$ 1024 high-resolution pose transfer result $\mathbf{I}_t$ in the hierarchical coarse-to-fine manner.

The Gaussian keypoint heatmaps $\mathbf{H}_t$ are directly converted from the input 3D target pose $\mathbf{P}_t$ for ensuring our network captures the pose spatial information effectively [9], [15], [19]. We apply two different human body models, SMPL [14] and 3D volume, to obtain the semantic parsing $\mathbf{S}_t$ and depth map $\mathbf{D}_t$. An ablation study about these two representations and more implementation details are included in Sec. IV. We follow the 7 body component annotations used in BodyNet [31], namely the head, left/right arm, left/right leg, the torso and the background. We conduct a weak-perspective camera projection, an affine transformation, to project the 3D human model and body annotations onto the 2D image.

B. 3D Human Representation

In order to obtain the informative maps to guide the following generation stages, we reconstruct the 3D human model of the source image $\mathbf{I}_s$ in the target pose $\mathbf{P}_t$ at first. To demonstrate the generalization of our method, we apply two types of human model representations in our experiments, namely SMPL [14] and 3D volume.

For the SMPL [14] representation, we first use HMR [28] to reconstruct 3D human model from the source image $\mathbf{I}_s$ with SMPL [14] parameters. We then replace its pose parameters by the 3D target pose $\mathbf{P}_t$. Therefore, we can build the target 3D human model with the source shape and the target pose. By projecting 3D meshes of the SMPL [14] model for each body part onto the image coordinate, we can obtain 7 part masks $\mathbf{S}_t$ as well as a depth map $\mathbf{D}_t$.

For the 3D volume representation, we reconstruct a 3D segmented volume by a volume generation network. The volume generation network takes the source human image $\mathbf{I}_s$ and the target pose $\mathbf{P}_t$ as input, then predicts a segmentation map with 7 labels on a voxel grid. Similarly, we can project the 3D segmented volume to obtain $\mathbf{S}_t$ and $\mathbf{D}_t$. Following recent 3D human shape reconstruction works [31], [41], we adopt the stacked hourglass networks [42] as our volume generation network. It first encodes the source image $\mathbf{I}_s$ and target 3D pose $\mathbf{P}_t$ independently and then concatenates the extracted features as the input of two stacked hourglass networks to decode a 3D segmented volume. We found a narrow depth resolution along z-axis does not affect the final result so much, so we set the voxel grid resolution as $256 \times 256 \times 64$ to accelerate the training. We apply the cross-entropy loss to train the volume generation network.

C. Part-Based Layer Representation

The key to a pleasant pose transfer result is handling the visibility ambiguity between body parts correctly as well as preserving local details properly. We propose to formulate the synthesis of important human parts, namely faces, arms and legs individually via a part-based layer representation. These regions always contain informative content, especially faces, and their visibility changes easily from source pose to target pose. While noticing the symmetry of human body, we let the left parts and right parts share the same generator. Specifically, we apply three part-based layer generators to handle the generation of important human body parts (face, arm, and leg) more precisely. These generators are critical to a high-quality result in both preserving the texture details in source pose and synthesizing missing regions in target pose. To better leverage the body symmetry, we let the layer generators for arm and leg also take the images of the other side as an extra input. The extra body part may provide valuable appearance information, especially when a certain part is occluded in the source image. Owing to this part-based layer representation, we can even synthesize proper texture locally though a body part is totally occluded in the source image but become visible in the target pose. We further force the part-based generator to synthesize a complete body component regardless of the occlusions in this stage and fuse the visible regions in the later fine-scale fusion stage. Thus more realistic results around occlusion boundary can be guaranteed.

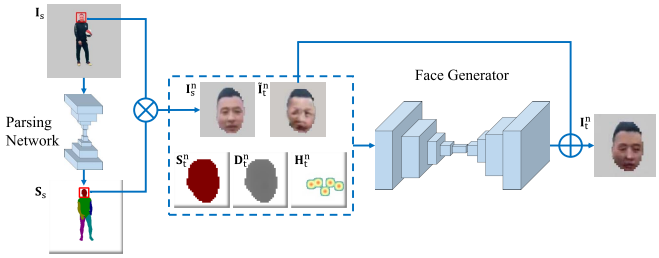


Fig. 3. An example of the proposed part-based layer representation. It takes the local patches of informative maps, coarse transfer result and source image as input to produce the layer result. Please note that arm/leg generator additionally takes source part of the other side as an extra input.

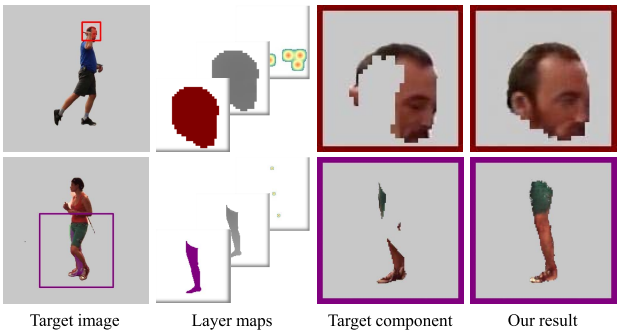


Fig. 4. Our part-based layer generator not only preserves local textures for a certain body component but also recovers the image content in the invisible areas. Even though one body component is partially occluded in the ground-truth target image (the first column), our layer generator still synthesizes a complete result (the last column).

Fig. 3 illustrates the pipeline of our face layer generator as an example. We segment the source image $\mathbf{I}_s$ with the same 7 body component labels as $\mathbf{S}_s$ by using a human parsing network [43] and crop the source layer image $\mathbf{I}_s^n$ accordingly. Three local patches of informative maps, $\mathbf{S}_t^n$, $\mathbf{D}_t^n$, $\mathbf{H}_t^n$, are generated directly from the 3D space to preserve their completeness regardless of the occlusion. To better maintain the personal characteristics and accelerate the training stage, we also take the cropped coarse-scale layer $\widetilde{\mathbf{I}}_t^n$ as input. Therefore, we concatenate $\mathbf{I}_s^n$, $\mathbf{S}_t^n$, $\mathbf{D}_t^n$, $\mathbf{H}_t^n$ and $\widetilde{\mathbf{I}}_t^n$ all together, then feed them into the specific part-based layer generator to synthesize a complete target layer $\mathbf{I}_t^n$ temporarily.

As shown in Fig. 4, we restrict the layer generator to always generate a complete body part instead of considering the appearing occlusions in the target pose. Upon the complete body part result, more proper self-occluded effects can be synthesized in the subsequent fine-scale fusion stage. To achieve this, we introduce an adversarial loss L_{adv}^n to identify whether a certain body part is complete or not:

$$L_{adv}^n = \log(D(\mathbf{I}_r^n, \mathbf{I}_s^n)) + \log(1 - D(\mathbf{I}_t^n, \mathbf{I}_s^n)), \tag{1}$$

where $\{\mathbf{I}_r^n\}$ denotes a random sample from a complete body part set with no occlusion.

Besides, we also apply a perceptual loss L^n_{per} to make the generated results perceptually similar to the ground truth. The perceptual loss measures the distance of the feature maps extracted by a pre-trained perception network, *e.g.* VGG-19 [44]. Aiming to synthesize an occlusion-free result in agreement with L^n_{adv}, we only calculate the perceptual loss on the visible regions to eliminate the effects of occluded unknown areas. So the perceptual loss for the n-th layer generator is formulated as:

$$L^n_{per} = \sum_{l=1}^{L} \left\| \Phi_l(\hat{\mathbf{I}}^n_t) - \Phi_l(\mathbf{I}^n_t) \right\|_1 \odot \mathbf{S}^n_t, \tag{2}$$

where $\hat{\mathbf{I}}^n_t$ denotes the ground-truth of target layer image. L is the number of selected feature layers and Φ_l is the l-th feature layer of the pre-trained perception network Φ.[1] Therefore, the total training loss for the proposed part-based layer generator is $L^n = L^n_{adv} + L^n_{per}$.

In our implementation, we let these layer generators predict an offset image $\check{\mathbf{I}}^n_t$ between the target pose image $\mathbf{I}^n_t$ and the coarse image $\widetilde{\mathbf{I}}^n_t$, which is shown to be more efficient in training. Considering the resolution of the final result is 1024, we determine the resolution of each layer generator by its relative length to the whole body height, namely 128 for face, 256 for arm and 512 for leg.

D. Hierarchical Generation Framework

In this section, we mainly introduce the design of our three-stage hierarchical framework, especially the first coarse-scale generation stage and the final fine-scale fusion stage.

1) Coarse-Scale Generation Stage: generates a rough pose transfer image $\widetilde{\mathbf{I}}_t$ with a lower 512×512 resolution in a coarse scale by using the aforementioned three informative maps $\mathbf{S}_t$, $\mathbf{D}_t$, and $\mathbf{H}_t$ as guidance. We train this network by adopting a perceptual loss $\widetilde{L}_{per}$, an adversarial loss $\widetilde{L}_{adv}$ and a feature matching loss $\widetilde{L}_{FM}$ as:

$$\widetilde{L} = \widetilde{L}_{adv} + \lambda_{FM}\widetilde{L}_{FM} + \lambda_{per}\widetilde{L}_{per}, \tag{3}$$

where $\widetilde{L}_{FM}$ defines the feature matching loss proposed in pix2pixHD [13]. $\widetilde{L}_{adv}$ is multi-scaled [13] and conditioned on $\mathbf{I}_s$ to preserve the human identity in $\widetilde{\mathbf{I}}_t$. So it is represented as:

$$\widetilde{L}_{adv} = \log(D(\hat{\mathbf{I}}_t, \mathbf{I}_s)) + \log(1 - D(\widetilde{\mathbf{I}}_t, \mathbf{I}_s)), \tag{4}$$

where $\hat{\mathbf{I}}_t$ denotes the ground-truth of transfer image. The perceptual loss $\widetilde{L}_{per}$ measures the perceptual distance between the coarse-scale result and the ground-truth of transfer image. We define $\widetilde{L}_{per}$ as:

$$\widetilde{L}_{per} = \sum_{l=1}^{L} \left\| \Phi_l(\hat{\mathbf{I}}_t) - \Phi_l(\widetilde{\mathbf{I}}_t) \right\|_1, \tag{5}$$

[1]In our experiments, Φ includes *relu1_2*, *relu2_2*, *relu3_2*, *relu4_2* and *relu5_2* layers in VGG-19 [44].

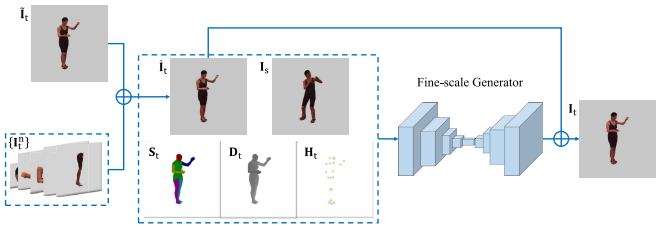


Fig. 5. Our fine-scale generator fuses the coarse-scale result and the five body parts together at first, and then refines this initial result for global consistency with the guidance from informative maps to produce the final high-quality transfer result.

2) Fine-Scale Fusion Stage: fuses the coarse-scale result $\widetilde{\mathbf{I}}_t$ and the offset images $\{\check{\mathbf{I}}^n_t\}$ for individual body parts together to achieve the final transfer result $\mathbf{I}_t$, as shown in Fig. 5. A simple way is to directly add the offset images onto the coarse result according to the parsing map $\mathbf{S}^n_t$ as:

$$\dot{\mathbf{I}}_t = \widetilde{\mathbf{I}}_t + \sum_{n}^{N} \check{\mathbf{I}}^n_t \odot \mathbf{S}^n_t. \tag{6}$$

However, this may generate improper artifacts because each offset image is synthesized independently and necessary global consistency is not guaranteed.

We employ a fine-scale generator to refine the initial result $\dot{\mathbf{I}}_t$ to ensure body consistency for generating a high-quality pose transfer result. We also utilize the informative maps as guidance in this stage to identify the occlusion relationships between all local body parts. With such guidance, our framework handles the self-occlusion problem more precisely than previous methods. Similar to our part-based layer generator, this fusion stage also produces an offset image $\check{\mathbf{I}}_t$ concerning to $\dot{\mathbf{I}}_t$. So the final result is $\mathbf{I}_t = \check{\mathbf{I}}_t + \dot{\mathbf{I}}_t$.

The objective function for training this stage is similar to training the coarse-scale generator and is formulated as:

$$L = L_{adv} + \lambda_{FM} L_{FM} + \lambda_{per} L_{per}, \tag{7}$$

where $L_{adv} = \log(D(\hat{\mathbf{I}}_t, \mathbf{I}_s)) + \log(1 - D(\mathbf{I}_t, \mathbf{I}_s))$ and $L_{per} = \sum_{l=1}^{L} \left\| \Phi_l(\hat{\mathbf{I}}_t) - \Phi_l(\mathbf{I}_t) \right\|_1$. We set $\lambda_{FM} = \lambda_{per} = 10$ in our experiments, both for Eq. (3) and Eq. (7).

IV. Experiments

We demonstrate the advantages of our human pose transfer method through extensive qualitative and quantitative comparisons with three state-of-the-art techniques [9], [12], [13]. Thanks to our hierarchical framework with guidance from 3D target pose as well as the proposed part-based layer representation, our method outperforms previous works in better handling self-occlusions, fine detail preservation/synthesis and a higher resolution result. Various ablation studies are also conducted for validating the effectiveness of important components in our framework, including the 3D guidance from informative maps, the part-based layer representation, and the coarse-to-fine fusion strategy. All the comparisons are evaluated on two datasets, the Human3.6M dataset [45] and our self-collected sport video dataset. Results on video sequences are also presented in the supplementary materials.

A. Datasets

Human3.6M dataset captures 11 actors who perform 15 actions under 4 viewpoints. We uniformly sample the captured videos into individual keyframes and obtain images for the same person with large pose variations. We crop the sampled images and remove the background through a human segmentation approach [46] to make sure that only the human foreground is considered in our method. Besides the ground truth annotation provided in this dataset, we further use OpenPose [47] to estimate the 2D poses and use HMR [28] to obtain the 3D ground truth to ensure the consistency between training and testing stages. We choose 'Greeting', 'Posing', 'SittingDown' and 'Walking' actions for training, and 'Directions', 'TakingPhoto' actions for testing.

Sport video dataset is our self-collected dataset in high resolution with annotated 3D poses for validating the generalization of our approach. We first download 87 high-resolution sport videos from the Internet, including basketball, running, football, etc. Each video contains a sequence of images for the same person with large pose variations. We again use OpenPose [47] and HMR [28] to estimate the corresponding poses for each frame and manually eliminate the inaccurate predictions. The background is also removed by using [46]. We split the dataset randomly into the training and testing set at the ratio of 9:1 according to the person identification.

B. Implementation Details

We adopt the architecture of Johnson *et al.* [48] for all the generators in our method. It contains several downsample and upsample convolutional layers with multiple residual blocks in the middle. In the coarse-scale generation and the fine-scale fusion stages, we use multi-scale 70 $\times$ 70 Patch-GAN [35] discriminators with two scales and three scales, respectively. For the completeness discriminator used in training the part-based layer representation, we adopt the architecture provided by DCGAN [49]. We replace the inner batch normalization [50] layers with spectral normalization [51] in arm and leg discriminators for training stability. The average inference time of our entire pipeline is 0.44s on a GTX 1080Ti GPU. Please refer to the supplementary materials for more hyperparameter details.

During the training stage, we train the coarse-scale generator at first. We then use the result from it to train the part-based layer generators. Finally, we finetune all the above sub-networks with the fine-scale generator together in an end-to-end manner. We train the coarse-scale and fine-scale generators for 10 epochs, and the part-based layer generators for 20 epochs. As we train all the generators with batch size 1, we use the instance normalization [52] instead of origin batch normalization. All the networks are trained by Adam [53] solver and the learning rate is set to 0.0002 which is constant during the first half epochs and is linearly decayed in the latter half training epochs.

The body component label is annotated by segmenting the uniform mesh topology of the SMPL [14] model. We project 3D meshes of each body part into image coordinate to obtain component masks as the 2D human parsing ground truth. The ground truth of 3D segmented volume is obtained through voxelizing a SMPL [14] 3D mesh into the voxel grid by using binvox [54].

TABLE I

QUALITATIVE EVALUATION WITH STATE-OF-THE-ART METHODS [9], [12], [13] ON HUMAN3.6M [45] DATASET AND OUR SELF-COLLECTED SPORT VIDEO DATASET. WE CONDUCT A USER STUDY AND THE PARTICIPANTS ARE ASKED TO PICK THEIR FAVOURITE RESULTS FROM THE FOUR METHODS

Dataset	Human3.6M [42]	Sport video
Pix2pixHD [13]	20.00%	14.29%
DSC [9]	4.29%	1.90%
LW-GAN [12]	4.29%	5.48%
Ours-S	**71.43%**	**78.33%**

C. Comparison With Previous Works

We compare the results of our approach with three state-of-the-art methods, including pix2pixHD [13], DSC [9] and LW-GAN [12]. Pix2pixHD [13] is a general framework to handle the high-resolution image-to-image translation task. DSC [9] and LW-GAN [12] are two state-of-the-art human pose transfer approaches but are limited in low image resolution. Pix2pixHD [13] and DSC [9] use 2D pose as input while LW-GAN [12] use SMPL [14] model as input. We re-train all these methods on the two datasets. Because the informative maps of our method can be generated from either the SMPL [14] model or the segmented 3D volume, we denote the corresponding results as Ours-S and Ours-V, respectively.

The qualitative comparisons on the two evaluated datasets are presented in Fig. 6 and Fig. 7, respectively. It is clear that our method significantly outperforms others, especially for the regions where the self-occlusion happens. Because we consider the semantic representation of human and synthesize each important body component with its own part-based generator, so local details of these regions are well preserved during transferring its pose, *e.g.* the face regions (see the first two rows in Fig. 6 and all results in Fig. 7). Besides preserving the local details, the content in new areas appearing in target pose is also synthesized more precisely by the separated layer generators. Furthermore, thanks to our informative guidance from the 3D target pose, our method identifies the visibility in target pose correctly and the upper layers (visible parts) are obviously improved. Our method can handle the extremely difficult self-occluded situations when legs are cross (see the last two rows in Fig. 6 and the first two rows in Fig. 7) or arms are laid in the front (see the last two rows in Fig. 6 and Fig. 7), while other methods fail.

We also conduct a user study to compare our method with other three human pose transfer techniques [9], [12], [13]. We randomly select 20 pairs of source images and target poses from the test set of Human3.6M [45] and the sport video dataset, respectively. We then show each source image and target pose, along with the four results (Pix2pixHD, DSC, LW-GAN, Ours-S) in a random order to users, who are asked to pick their favourite one. There are 21 participants in total, resulting in 420 votes for each dataset. Tab. I lists

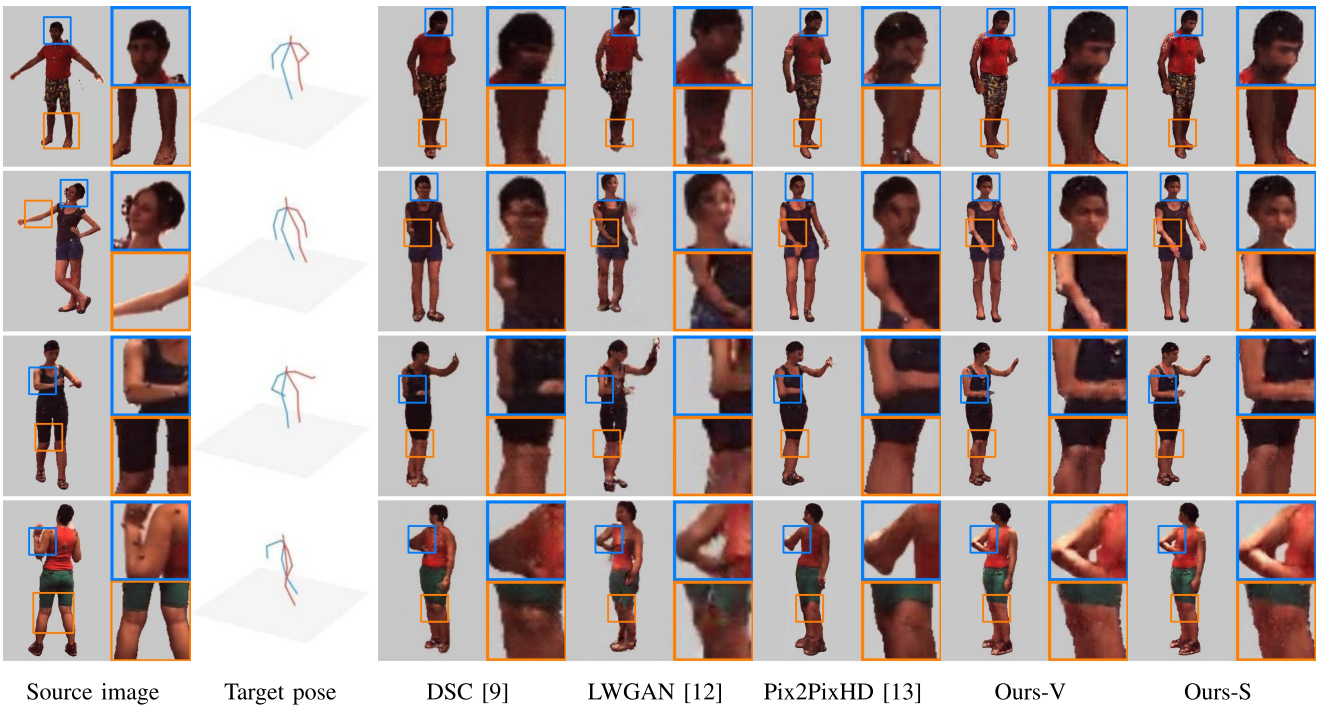


Fig. 6. Qualitative comparison with state-of-the-art methods [9], [12], [13] on Human3.6M [45] dataset. We cropped close-up views for better visualization and please zoom in for details.

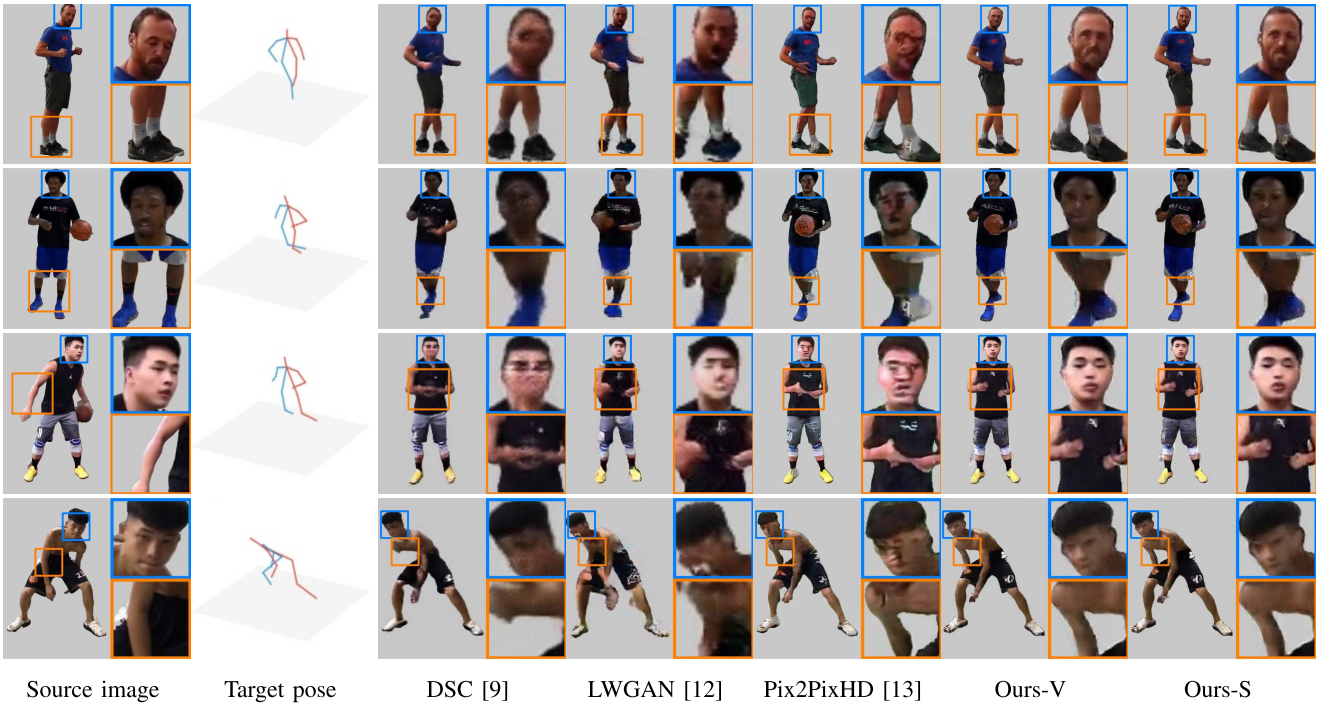


Fig. 7. Qualitative comparison with state-of-the-art methods [9], [12], [13] on our self-collected sport video dataset. We cropped close-up views for better visualization and please zoom in for details.

the proportion of votes for each method and shows that our method receives significantly more votes than other methods.

Besides qualitative comparisons, we employ two quantitative metrics, namely SSIM [56] and Learned Perceptual Similarity (LPIPS) [57] to measure the quality of images generated by all methods. We also design a new metric, Occlusion State Accuracy (OSA), to expressly demonstrate the superiority of our method for addressing the self-occlusion problem. We apply HMR [28] to estimate the 3D human model from the generated image and calculate the accuracy of occlusion states for different body components comparing to the ground truth. We list the corresponding evaluation

TABLE II

QUANTITATIVE EVALUATION WITH STATE-OF-THE-ART METHODS [9], [12], [13] ON HUMAN3.6M [45] DATASET AND OUR SELF-COLLECTED SPORT VIDEO DATASET. THE ARROW AFTER EACH METRIC IDENTIFIES THE IMPROVEMENT DIRECTION

Dataset	Human3.6M [42]			Sport video		
Metric	SSIM (↑)	LPIPS (↓)	OSA (↑)	SSIM (↑)	LPIPS (↓)	OSA (↑)
Pix2pixHD [13]	0.9524	0.0497	0.9143	0.9322	0.0792	0.9132
DSC [9]	0.9223	0.0574	0.9058	0.9026	0.0883	0.9109
LW-GAN [12]	0.8965	0.0800	0.9194	0.8665	0.1280	0.9134
Ours-V	0.9520	0.0471	0.9386	0.9372	0.0735	0.9381
Ours-S	**0.9530**	0.0458	**0.9413**	**0.9373**	**0.0731**	**0.9398**
Ours-S-256	0.9431	**0.0416**	**0.9413**	0.9140	0.0738	**0.9398**

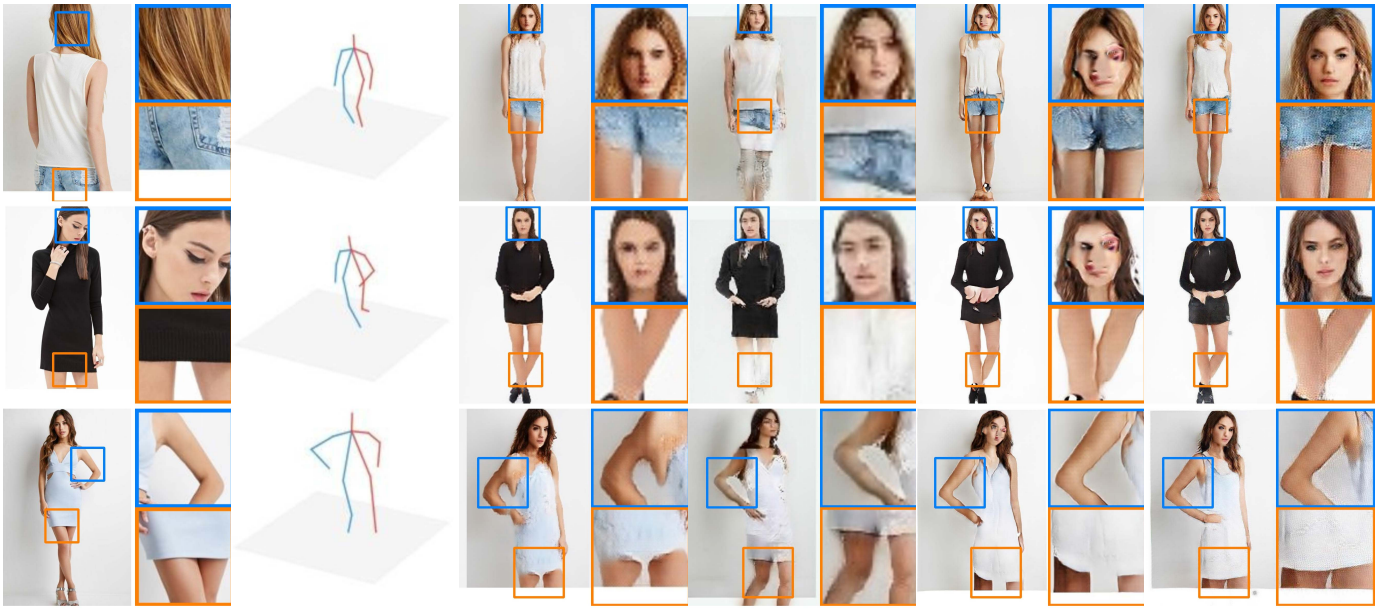


Fig. 8. Qualitative comparison with state-of-the-art methods [9], [12], [13] on DeepFashion [55] dataset. We cropped close-up views for better visualization and please zoom in for details.

in Tab. II. Not surprisingly, our method outperforms state-of-the-art approaches both in the general image quality metrics (SSIM [56] and LPIPS [57]) and self-occlusion handling (OSA). LW-GAN [12] employs HMR [28] to estimate the target 3D model, thus it performs better than the other two methods for self-occlusion handling (OSA). However, the image quality of its result is also unsatisfying because the semantic and spatial relationships of the different body parts are still ignored. Since we use HMR [28] to annotate 3D ground truth for the training set, our method performs better when using SMPL [14] model than 3D volume at the inference time. We resize our results to scale of 256 × 256, denoted as Ours-S-256 in Tab. II. Our method still performs better than LW-GAN [12] and DSC [9], which are both trained and evaluated by the image size of 256 × 256.

We also compare our method with DSC [9], LW-GAN [12] and Pix2pixHD [13] on the DeepFashion [55] dataset. Following the training/test split applied in DSC, we re-train our method and Pix2pixHD on DeepFashion. We also adopt the pre-trained models released publicly by DSC and LW-GAN for testing. Fig. 8 shows the qualitative comparison on the DeepFashion. Our method obviously outperforms other pose transfer techniques, especially for the local details of each body component (faces in the first two rows and lower bodies in the first row) and the regions where self-occlusions happen (cross legs in the second row and the arm in the last row). Our method also achieves the best scores in two quantitative metrics, namely SSIM and LPIPS, as shown in Tab. III.

TABLE III

QUANTITATIVE EVALUATION WITH STATE-OF-THE-ART METHODS [9], [12], [13] ON DEEPFASHION [55] DATASET. THE ARROW AFTER EACH METRIC IDENTIFIES THE IMPROVEMENT DIRECTION

Method	SSIM (↑)[53]	LPIPS (↓)[54]
Pix2pixHD [13]	0.8440	0.2228
DSC [9]	0.8179	0.2258
LW-GAN [12]	0.7472	0.2540
Ours-S	**0.8454**	**0.2204**

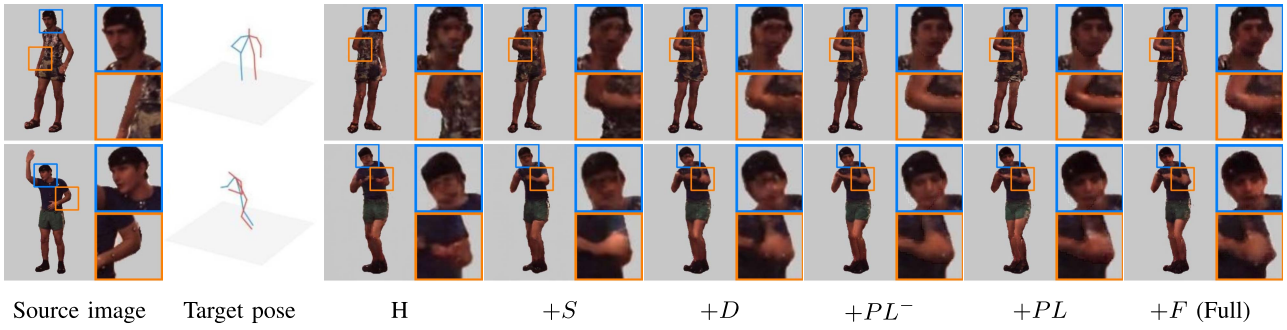


Fig. 9. Qualitative comparison for ablation studies by adding each component one by one. We cropped close-up views for better visualization and please zoom in for details.

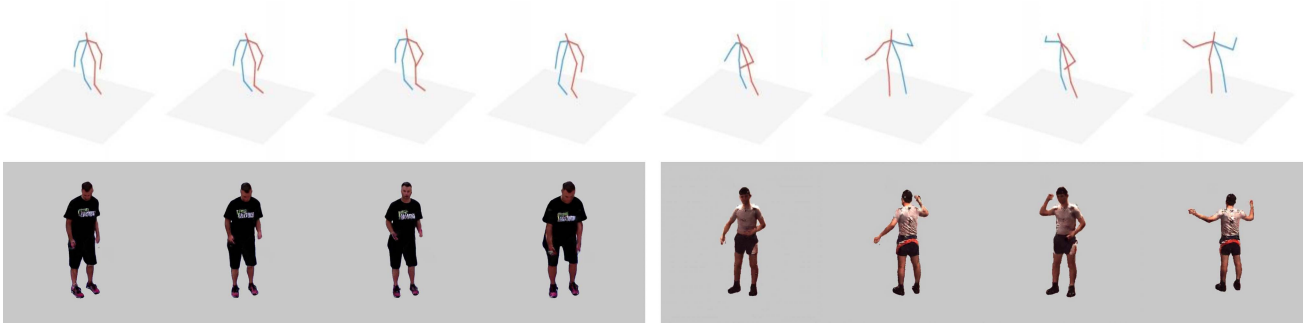


Fig. 10. Selected frames from motion sequences generated by our method. Please zoom in for a better visualization.

TABLE IV

ABLATION STUDY FOR VALIDATING THE EFFECTIVENESS OF EACH IMPORTANT COMPONENT IN OUR FRAMEWORK. H, S, AND D IDENTIFY THE POSE HEATMAPS, THE PARSING MAP, AND THE DEPTH MAP FOR INPUT, RESPECTIVELY. PL^- AND PL DENOTE THE INCOMPLETE/FULL PART-BASED LAYER REPRESENTATION. F DENOTES THE FINE-SCALE FUSION STAGE. THE ARROW AFTER EACH METRIC IDENTIFIES THE IMPROVEMENT DIRECTION

Method	SSIM (↑)[53]	LPIPS (↓)[54]	OSA (↑)
H	0.9402	0.0535	0.9098
$H+S$	0.9424	0.0518	0.9365
$H+S+D$	0.9429	0.0507	0.9405
$H+S+D+PL^-$	0.9521	0.0480	0.9406
$H+S+D+PL$	0.9524	0.0478	0.9405
$H+S+D+PL+F$	**0.9530**	**0.0458**	**0.9413**

D. Ablation Study

In order to validate the effectiveness of each important component in our framework, we conduct an ablation study by incorporating each component one by one. Our baseline model H just uses the coarse-scale generator to produce the target image with source image $\mathbf{I}_s$ and the 2D pose heatmaps $\mathbf{H}_t$ as input. We then incrementally incorporate each important component, namely the human parsing map $\mathbf{S}_t$, the extra depth map $\mathbf{D}_t$, the incomplete and complete part-based layer representation, and the fine-scale fusion stage in this study. We denote these components as S, D, PL^-, PL and F, respectively. So our whole solution can be denoted as $H+S+D+PL+F$. We use the SMPL [14] model as the 3D representation through the entire ablation studies.

We show the quantitative ablation study on Human3.6M dataset [45] using SSIM [56], LPIPS [57] and Occlusion State Accuracy (OSA) of the five variations in Tab. IV. It is obvious that all the studied components contribute to the final high-quality results as the increase of SSIM [56]/OSA and the decrease of LPIPS [57] along the five variations. Among these components, our part-based layer representation significantly improves the quality of generated images which can be explained by the large improvement in SSIM [56] and LPIPS [57]. Qualitative comparison in Fig. 9 also demonstrates that handling important body parts individually by our part-based layer representation is necessary for generating fine-scale local details, *e.g.* faces, and the fine-scale fusion stage further polishes the individual results by ensuring global consistency.

E. Motion Video Results

Our method can be easily applied to generate the motion video with a given target pose sequence for the reference person pose-by-pose. Fig. 10 shows several frames generated by our method from one pose sequence. Please notice that some partial details around the important body components, especially the face and the garments, are well preserved while applying a large pose motion. Please refer to our supplementary materials for more motion video results.

V. CONCLUSION

In this paper, we propose a hierarchical end-to-end human pose generation framework with the consideration of semantic part-based representation of the human body. We segment the human body into multiple parts and formulate individual layer representation for each of them. A high-quality result is fused with these local layer representations in a coarse-to-fine manner. We employ 3D informative guidance through the framework to better identify the occlusion ambiguity and

generate proper results in self-occluded regions. Both qualitative and quantitative evaluations demonstrate that our method significantly outperforms previous methods consistently. Our framework is general and can be adapted to various 3D representations, such as SMPL [14] model and 3D volume.

Limitations: Our method is still not able to handle hand regions well since common joint key points do not include fingers. Our part-based layer representation and the entire framework can be easily extended to such case once more spatial constraints around hand regions are available.

REFERENCES

[1] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[2] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An image-based virtual try-on network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7543–7552.

[3] M. Zanfir, A.-I. Popa, A. Zanfir, and C. Sminchisescu, "Human appearance transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5391–5399.

[4] X. Wu *et al.*, "Deep portrait image completion and extrapolation," *IEEE Trans. Image Process.*, vol. 29, pp. 2344–2355, 2020.

[5] B. Zhao, X. Wu, Z.-Q. Cheng, H. Liu, Z. Jie, and J. Feng, "Multi-view image generation from a single-view," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 383–391.

[6] H. Zhu, H. Su, P. Wang, X. Cao, and R. Yang, "View extrapolation of human body from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4450–4459.

[7] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, "Soft-gated warping-GAN for pose-guided person image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 474–484.

[8] Y. Li, C. Huang, and C. C. Loy, "Dense intrinsic appearance flow for human pose transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3693–3702.

[9] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, "Deformable GANs for pose-based human image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3408–3416.

[10] A. Grigorev, A. Sevastopolsky, A. Vakhitov, and V. Lempitsky, "Coordinate-based texture inpainting for pose-guided human image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12135–12144.

[11] N. Neverova, R. A. Guler, and I. Kokkinos, "Dense pose transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 123–138.

[12] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao, "Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5904–5913.

[13] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[14] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

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

[16] C. Si, W. Wang, L. Wang, and T. Tan, "Multistage adversarial losses for pose-based human image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 118–126.

[17] P. Esser and E. Sutter, "A variational U-Net for conditional appearance and shape generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8857–8866.

[18] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 99–108.

[19] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, "Synthesizing images of humans in unseen poses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8340–8348.

[20] C. Yang, Z. Wang, X. Zhu, C. Huang, J. Shi, and D. Lin, "Pose guided human video generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 201–216.

[21] H. Cai, C. Bai, Y.-W. Tai, and C.-K. Tang, "Deep video generation, prediction and completion of human action sequences," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 366–382.

[22] C. Chan, S. Ginosar, T. Zhou, and A. Efros, "Everybody dance now," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5933–5942.

[23] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1526–1535.

[24] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3332–3341.

[25] Q. Zheng, W. Wu, H. Pan, N. Mitra, and D. Cohen-Or, "Inferring object properties from human interaction and transferring them to new motions," *Comput. Vis. Media*, vol. 7, no. 3, pp. 375–392, 2021.

[26] R. A. Guler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7297–7306.

[27] L. Liu *et al.*, "Neural rendering and reenactment of human actor videos," 2018, *arXiv:1809.03658*. [Online]. Available: http://arxiv.org/abs/1809.03658

[28] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7122–7131.

[29] C. Lassner, G. Pons-Moll, and P. V. Gehler, "A generative model of people in clothing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 853–862.

[30] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 561–578.

[31] G. Varol *et al.*, "BodyNet: Volumetric inference of 3D human body shapes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 20–36.

[32] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, "DeepHuman: 3D human reconstruction from a single image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7739–7749.

[33] X. Wu, K. Xu, and P. Hall, "A survey of image synthesis and editing with generative adversarial networks," *Tsinghua Sci. Technol.*, vol. 22, no. 6, pp. 660–674.

[34] W.-Y. Zhou, G.-W. Yang, and S.-M. Hu, "Jittor-GAN: A fast-training generative adversarial network model zoo based on Jittor," *Comput. Vis. Media*, vol. 7, no. 1, pp. 153–157, 2021.

[35] C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 702–716.

[36] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[37] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[38] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, p. 107, 2017.

[39] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3911–3919.

[40] Q. Gu, G. Wang, M. T. Chiu, Y.-W. Tai, and C.-K. Tang, "LADN: Local adversarial disentangling network for facial makeup and de-makeup," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10481–10490.

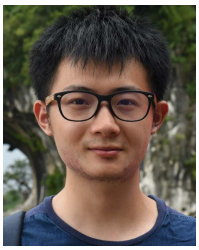
[41] H. Yi *et al.*, "MMFace: A multi-metric regression network for unconstrained face reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7663–7672.

[42] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 483–499.

[43] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[45] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.
[46] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
[47] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," 2018, *arXiv:1812.08008*. [Online]. Available: http://arxiv.org/abs/1812.08008
[48] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 694–711.
[49] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434
[50] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167
[51] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*. [Online]. Available: http://arxiv.org/abs/1802.05957
[52] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: http://arxiv.org/abs/1607.08022
[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[54] F. S. Nooruddin and G. Turk, "Simplification and repair of polygonal models using volumetric techniques," *IEEE Trans. Vis. Comput. Graph.*, vol. 9, no. 2, pp. 191–205, Apr./Jun. 2003.
[55] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1096–1104.
[56] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[57] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.



Xian Wu received the B.S. degree from Tsinghua University in 2015, where he is currently pursuing the Ph.D. degree. His research interests include image synthesis/editing and deep learning in computer graphics.

Chen Li received the B.Eng. degree from Zhejiang University in 2011 and the Ph.D. degree from Zhejiang University in 2017, under the supervision of Prof. Kun Zhou. From February 2012 to December 2016, he worked with Dr. Steve Lin at Internet Graphics Group, Microsoft Research, Beijing, as a Research Intern. He is currently a Senior Researcher with Weixin Group, Tencent. Before he joined Tencent in September 2017, he worked as an Algorithm Engineer at Dajiang Innovation for eight months. His research interests fall in the field of computer vision and computer graphics, specifically the 3D reconstruction, appearance modeling, computational photography, and the relevant VR/AR applications on facial images.

Shi-Min Hu (Senior Member, IEEE) received the Ph.D. degree from Zhejiang University in 1996. He is currently a Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing. His research interests include digital geometry processing, video processing, rendering, computer animation, and computer-aided geometric design. He is the Editor-in-Chief of *Computational Visual Media* and on the Editorial Board of several journals, including *Computer Aided Design* and *Computers and Graphics*.

Yu-Wing Tai (Senior Member, IEEE) received the B.Eng. (Hons.) and M.Phil. degrees from the Department of Computer Science and Engineering, HKUST, in 2003 and 2005, respectively, and the Ph.D. degree from the National University of Singapore in 2009. Since 2017, he has been an Adjunct Professor with the Department of Computer Science and Engineering, HKUST, and the Research Director of Kwai Inc. since 2020. He was the Research Director of YouTu lab of Tencent from 2017 to 2020, the Principle Research Scientist of SenseTime Group Ltd. from 2015 to 2016, and an Associate Professor with Korea Advanced Institute of Science and Technology (KAIST) from 2009 to 2015. His research interests include deep learning, computer vision, and image/video processing. He regularly served as the Area Chair/Program Committees for CVPR/ICCV/ECCV.