Detailed Avatar Recovery from Single Image

Hao Zhu, Member, IEEE, Xinxin Zuo, Member, IEEE, Haotian Yang, Sen Wang, Xun Cao, Member, IEEE, and Ruigang Yang, Senior Member, IEEE

Abstract—This paper presents a novel framework to recover detailed avatar from a single image. It is a challenging task due to factors such as variations in human shapes, body poses, texture, and viewpoints. Prior methods typically attempt to recover the human body shape using a parametric-based template that lacks the surface details. As such resulting body shape appears to be without clothing. In this paper, we propose a novel learning-based framework that combines the robustness of the parametric model with the flexibility of free-form 3D deformation. We use the deep neural networks to refine the 3D shape in a Hierarchical Mesh Deformation (HMD) framework, utilizing the constraints from body joints, silhouettes, and per-pixel shading information. Our method can restore detailed human body shapes with complete textures beyond skinned models. Experiments demonstrate that our method has outperformed previous state-of-the-art approaches, achieving better accuracy in terms of both 2D IoU number and 3D metric distance.

Index Terms—human avatar, 3D reconstruction, texture completion, deep neural network.

1 INTRODUCTION

Building a human avatar from a single image is a challenging problem and has drawn much attention in recent years. A large number of approaches [3], [10], [12], [15], [28], [29], [30], [32], [44], [47], [60], [62] have been proposed in which the human body shapes are reconstructed by predicting the parameters of a statistical skinned model, such as SMPL [38] and SCAPE [7]. The parametric shape is of low-fidelity, and unable to capture clothing details. Though several works [3], [12], [47] attempts to recover more details than a skinned model, they did not go far on the issue of detailed geometry recovery. Another collection of methods [63], [65] estimate volumetric human shape directly from the image using neural networks, while the resulting volumetric representation is fairly coarse and does not contain shape details.

The limited performance of previous methods is caused by the large variations of the human shape and pose. Parametric or volumetric shapes are not expressive enough to represent the inherent complexity. Besides, most of those previous methods focus on shape recovery while neglecting surface texture or appearance which is another important aspect to build a human avatar.

In this paper, we propose a novel framework to reconstruct detailed human avatar from a single image. The key idea is to combine the robustness of the parametric model with the flexibility of free-form deformation. In short, we build on top of the SMPL model to obtain an initial parametric mesh model and perform non-rigid 3D deformation on the mesh to refine the surface shape. We design a coarse-to-fine refinement scheme in which a deep neural network is used in each stage to estimate the 3D mesh vertex movement by minimizing its 2D projection error in the image space. We feed window-cropped images instead of the full image to the network, which leads to a more accurate and robust prediction of deformation. In addition, we integrate a photometric term to allow high-frequency details to be recovered. These techniques combined lead to a method that significantly improves, both visually and quantitatively, the recovered human shape from a single image as shown in Figure 1. Finally, we regress the complete texture of the reconstructed human model...
from the input image with our proposed texture synthesis network.

Different from our previous version [79], we extend our system by texturing the reconstructed 3D human body and recovering its complete appearance from a single image. This is the so-called human avatar which includes both the human shape and its texture. The major issue for the texture recovery problem is that for the predicted topologically-uniformed human mesh, only less than half of the texture can be seen from the single input image, and the visible textures are also semantically misaligned due to the deviation between the predicted mesh and the image. Therefore, synthesizing complete and realistic textures is quite challenging in this case. In this paper, combining the advantages of both flow-based warping networks and image generation networks, we design a novel framework to synthesize the missing texture with high tolerance to image misalignment and background interference.

The contributions of this paper include:

- We develop a novel project - predict - deform strategy to predict the deformation of the 3D mesh model by using 2D features.
- We carefully design a hierarchical update structure, incorporating body joints, silhouettes, and photometric-stereo to improve shape accuracy without losing robustness.
- We propose a texture synthesis network to restore the complete texture from the input image, which is robust to the misalignment caused by the shape reconstruction phase.
- As demonstrated throughout our experiments, the additional free deformation of the initial parametric model leads to quantitatively more accurate shapes with good generalization capabilities to images in the wild. The texture synthesis network generates the complete texture for the recovered shape, after which we will get the final human avatar.

2 Related Work

Previous approaches can be divided into two categories based on the way the human body is represented: parametric methods and non-parametric methods.

As for parametric methods, they rely on a pre-trained generative human model, such as the SCAPE [7] or SMPL [38] models. The goal is to predict the parameters of the generative model. The SCAPE model has been adopted by Guan et al. [15] to recover the human shape and poses from the monocular image as provided with some manually clicked key points and the constraint of smooth shading. Instead of relying on manual intervention, Dibra et al. [12] have trained a convolutional neural network to predict SCAPE parameters from a single silhouette image. Similar to the SCAPE model, Hasler et al. [19] have proposed a multi-linear model of human pose and body shape which was generated by factorizing the measurements into the pose and shape dependent components. The SMPL model [38] has recently drawn much attention due to its flexibility and efficiency. For example, Bogo et al. [10] have presented an automatic approach called SMPLify which fits the SMPL model by minimizing an objective function that penalizes the error between the projected model joints and detected 2D joints obtained from a CNN-based method together with some priors over the pose and shape. Building upon this SMPLify method, Lassner et al. [32] have formed an initial dataset of 3D body fitting with rich annotations consisting of 91 key points and 31 segments. Using this dataset, they have shown improved performance on part segmentation, pose estimation, and 3D fitting. Tan et al. [60] have proposed an indirect learning procedure by first training a decoder to predict body silhouettes from SMPL parameters and then using pairs of real images and ground truth silhouettes to train a full encoder-decoder network to predict SMPL parameters at the information bottleneck. Pavlakos et al. [47] have separated the SMPL parameters prediction network into two sub-networks. Taking the 2D image as input, the first network was designed to predict the silhouette and 2D joints, from which the shape and pose parameters were estimated respectively. The latter network combined the shape and 2D joints to predict the final mesh. Kanazawa et al. [28] have proposed an end-to-end framework to recover the human body shape and pose in the form of SMPL model using only 2D joints annotations with an adversarial loss to effectively constrain the pose. Instead of using joints or silhouettes, Omran et al. [44] believed that a reliable bottom-up semantic body part segmentation was more effective for shape and pose prediction. Therefore, they predicted a part segmentation from the input image in the first stage and took this segmentation to predict SMPL parameter of the body mesh. Alldieck et al. [3] have used a two-stage strategy to recover the human shape. Firstly, they used a part-based model to regress the 3D parametric human model; in the second stage, an iterative refinement was applied using the unwrapped texture to reconstruct the body shape. Kanazawa et al. [29] have proposed a framework that could learn a representation of 3D humans dynamics from a video via the temporal encoding of image features, which has taken advantage of the temporal information to enhance the recovered shape quality. Alldieck et al. [3] have presented a learning-based approach to estimate body shapes including hair and clothing and they also took a monocular video as input. Yu et al. [70] have proposed a multi-layer representation of garments and body to capture human performance using an RGBD camera. The physics-based cloth simulation was incorporated into the performance capture pipeline, to simulate plausible cloth dynamics and cloth-body interactions. Bhattacharjee et al. [9] have proposed to predict body shape and clothing, layered on top of the SMPL model from a single or a few frames. The predicted garment geometry is related to the body shape and can be transferred to new body shapes and poses. Alldieck et al. [5] have proposed to reconstruct the human shape from a single image in UV space. In their framework, a UV transformer firstly transformed the input image to the partially visible UV texture, and then a PatchGAN [23] was used to synthesize the complete normal and displacement map that can be applied to the SMPL model. Mir et al. [39] have focused on transferring textures of clothing images to 3D garments worn on top SMPL, which enables 3D virtual try-on in real-time.

Non-parametric methods directly predict the shape representation from the image. Some researchers have used depth maps as a more general and direct representation of shapes. For example, Varol et al. [64] have trained a convolutional neural network by building up a synthetic dataset of rendered SMPL models to predict the human shape in the form of depth image and body part segmentation. Güler et al. [16], [17] have treated the shape prediction problem as a correspondence regression problem, which would produce a dense 2D-to-3D surface correspondence field for the human body. Another way of representing 3D shapes is to embed the 3D mesh into a volumetric space [63], [65]. For example, Varol et al. [63] have restored volumetric body shape directly from a single image. The mesh model was extracted from
Fig. 2: The flow of our shape-recovering method goes from the bottom left to the top right. The mesh deformation architecture consists of three levels: joint, anchor and per-vertex. In each level, the 3D mesh is projected to 2D space and sent together with the source image to the prediction network. The 3D mesh gets deformed by the predicted results to produce refined human body shapes.

the predicted volumetric model and fitted to an SMPL model as a post-processing procedure. Similarly, Venkat et al. [65] have recovered the volumetric grid of the human body from a single image and they have put much effort into texture view synthesis to get a textured 3D model. However, they only performed the test on images captured in the lab environment rather than images in the wild. In addition to depth and volumetric representations, Kulkarni et al. [31] have proposed a probabilistic programming language that can express generative models for arbitrary 2D/3D objects. The language was used to predict pose as well as shape from simple images while the accuracy was not evaluated. Dibra et al. [11] have proposed to learn a mapping from silhouettes to an embedding space from which 3D human body mesh will get restored. Their method focuses more on robust body measurements rather than shape details or poses. Kolotouros et al. [30] have proposed to relax the reliance on the model’s parameter space, and directly regressed the 3D location of the mesh vertices instead of predicting the model’s parameters using a Graph-CNN. Natsume et al. [41] have introduced the implicit representation that uses 2D silhouettes and 3D joints of a body pose to describe the immense shape complexity and variations of clothed people. They used the deep visual hull algorithm to predict 3D shape from the synthesized silhouettes which are consistent with the input segmentation, and also inferred the texture of the back view using a conditional generative adversarial network. Lazova et al. [33] have proposed to firstly predict the dense correspondence and garment segmentation, then a neural network was designed to predict the completed texture and displacement maps respectively. The displacement map and the complete texture were then merged on the base of the SMPL model to form the fully-textured 3D avatar. Habermann et al. [18] have proposed a real-time human performance capture approach that reconstructed dense, space-time coherent deforming geometry of clothed people from a single monocular RGB stream. Compared with our proposed method, this method requires a pre-reconstructed model as a reference. Smith et al. [58] have proposed to use an image-translation network to recover the 3D geometry of a human body, and adopted per-pixel surface normals instead of per-pixel depth for training loss, which has made it possible to estimate detailed body geometry. Zheng et al. [73] have proposed to use the single image together with dense semantic representation generated from SMPL mesh as input, and used a 3D convolutional network to predict the volumetric shape of the target human. Tang et al. [61] have proposed to train a network to predict the depth map of the human body, and Tan et al. [59] further proposed to train the network to predict human depth map in a self-supervised manner. The photometric loss within several frames was used to supervise the regression of the depth map after the human motion was compensated.

In recent years, implicit functions have also shown great potential in single view reconstruction of the human body. Saito et al. [53] have proposed a framework with a pixel-aligned implicit function to estimate the shape of the clothed human shape. As a follow-up, a multi-level architecture has been introduced to reveal the surface details [54], and it has been optimized for real-time performance capture [34]. Similarly, Huang et al. [20] have proposed to firstly estimate the correspondence between the input image and the canonical model, and then reconstructed the human shape using the implicit function in the canonical space. PaMIR [72] proposed to combine the parametric body model with the free-form deep implicit function, which improves the generalization ability of humans with challenging poses. Similarly, both PaMIR and our method utilize SMPL based parametric models. However, PaMIR predicts the implicit function to represent the cloth, while our method expresses the surface details with a displacement map of the parametric model and recovers the clothed human shape through a free-form deformation, thereby maintaining the mesh topology and the ability of rigging.

While significant progress has been made in this very difficult problem, the recovered human shape is still lacking in accuracy.
and details. In contrast to all the above methods, we present a method to predict human body shape from coarse to fine at multiple scales, and we also propose a texture completion method to generate the human avatar with complete texture.

3 Hierarchical Deformation Framework

In this section, we will present our hierarchical deformation framework to recover detailed human body shapes by refining a template model in a coarse-to-fine manner. As shown in Figure 2, there are four stages in our framework: First, an initial SMPL mesh is estimated from the source image. Starting from this, the next three stages serve as refinement phases which predict the deformation of the mesh so as to produce a detailed human shape. We have used the HMR method [28] to predict the initial human mesh model, which has demonstrated state-of-the-art performance on human shape recovery from a single image. However, like other human shape recovery methods [10], [44], [47] that utilize the SMPL model, the HMR method predicts the shape and pose parameters to generate a skinned mesh model with limited flexibility to closely fit the input image or express surface details. For example, the HMR often predicts deflected joint position of limbs when the human pose is unusual. Therefore, we have designed our framework to refine both the human shape and the pose.

The refining stages are arranged hierarchically from coarse to fine. We define three levels of key points on the mesh, referred to as handles in this paper. We will describe exactly how we define these handles in the next section. In each level, we design a deep neural network to refine the 3D mesh geometry using these handles as control points. We train the three refinement networks separately and successively to predict the residual deformation based on its previous phase.

To realize the overall refinement procedure, a challenging problem is how to deform the 3D human mesh from handles in 2D space using deep neural networks. We address this using Laplacian mesh deformation. In detail, the motion vector for each handle is predicted from the network driven by the joints and silhouettes of the 2D image. Then the human mesh will get deformed with the Laplacian deformation approach given the movements of the handles while maintaining the local geometry of the human model. The deforming strategy has been used in multi-view shape reconstruction problem [4], [36], [48], [50], [51], [76], [77], while we are the first to predict the deformation from a single image with the deep neural network.

3.1 Handle Definitions

In this section, we will describe the handles that we have used in each level. They could be predefined in the template model thanks to the uniform topology of SMPL mesh model.

**Joint handles.** We select 10 joints as the control points – head, waist, left/right shoulders, left/right elbows, left/right knees, and left/right ankles. The vertices around the joints under the T-pose SMPL mesh are selected as handles, as shown in Figure 3. We take the geometric center of each set of handles as the position of its corresponding body joint. The motion of each joint handle is represented as a 2D vector, which refers to the vector from the joint position of projected mesh to ground truth joint position on the image plane.

**Anchor handles.** We select 200 vertices on the human template mesh under T-pose as anchor handles. To select the anchors evenly over the template, we firstly build a vector set

\[ C = \{v_1, v_2, \ldots, v_n\} \]

with \( v_i \) concatenated by the position and surface normal of the vertex \( i \) and \( n \) is the number of SMPL model vertices. Then K-means is applied to cluster set \( C \) into 200 classes. Finally, we set the closest vertex to the center of each cluster as the anchor handles. Besides, we have removed the vertices in the face, fingers, toes from the T-pose SMPL model to prevent the interference of high-frequency shape. To be noticed that, for each anchor, it is only allowed to move along the surface normal direction, so we just need to predict a single value indicating the movement of the anchor point along the normal direction.

**Vertex handles.** The vertices in the SMPL mesh are too sparse to apply pixel-level deform, so we subdivide each face of the mesh into 4 faces. The subdivision increases the number of vertices of the template mesh to 27554, and all these vertices are regarded as handles.

3.2 Joint and Anchor Prediction

**Network.** Both joint and anchor prediction networks use the VGG [57] structure which consists of a feature extractor and a regressor. The network takes the mesh-projected silhouette and source image as input, which are cropped into patches as centered with our predefined handles. Basically, as joint and anchor handles are pre-defined on the topologically uniformed SMPL mesh, we can get the 3D position of the joint/anchor handles by indexing on the mesh. Then the 2D position of the handles can be obtained by projecting 3D handle points, and the cropping windows are centered on these points with the predefined square size. For a \( 224 \times 224 \) input image, the image is cropped into patches with the size of \( 64 \times 64 \) for joint prediction, and \( 32 \times 32 \) for anchor prediction. Compared with the full image or silhouette input, the handle cropped input allows the network to focus on the region of interest. We will demonstrate the effectiveness of the cropped input in Section 5.3.

**Loss.** The output of the joint net is a 2D vector representing the joint motion in the image plane. L2 loss is exploited to train the joint net with the loss function formulated as:

\[ L_{joint} = ||p - \tilde{p}||_2 \]  

where \( p \) is the predicted motion vector from the network and \( \tilde{p} \) is the displacement vector from the mesh-projected joint position to its corresponding ground-truth joint. Both vectors are 2-dimensional.

For the anchor net, our immediate goal is to minimize the area of the mismatched part between the projected silhouette and the ground truth silhouette. One strategy is to follow Aldieck et al.’s work [4], where the mesh is optimized to fit the multi-view silhouettes. However, in the single-view silhouette fitting problem,
it requires strong constraints to maintain the global structure stable. Therefore, we take advantage of the Laplacian deforming strategy, and use sparse-sampled anchor handles as control points. We use a transformation vector to represent the deformation of the silhouette. The transformation vector is in the same direction as the normal vector of the anchor vertex, and its length is calculated as the distance from the predicted silhouette to the ground-truth silhouette along the vertex normal direction. The transformation vector is regarded as the movement of the anchor, and L2 loss is used to train the network. In the Laplacian deformation stage, two kinds of anchor handles do not participate in deformation as control points: One kind is the point with deformation distance \( \geq 0.1m \), which is considered as the internal anchor handle; The other kind is the point that is too close to the edge of the silhouette. Since the Laplacian deforming will keep the local geometry as much as possible, the overall shape would be deformed equally.

Besides, instead of using the RGB image as input, the joint and anchor prediction network could also take the ground-truth silhouette of the human figure as input if available. The silhouette provides more explicit information for the human figure, which prevents the network from getting confused by the cluttered background environment. We will demonstrate its effectiveness on joint and anchor deformation prediction in the experiment section. In this paper, we consider the RGB-only as input by default, and use ‘+Sil.’ to indicate the case where the additional silhouette is used.

3.3 Vertex-level Deformation

To add high-frequency details to the reconstructed human models, we exploit the shading information contained in the input image. First, we project the current 3D human model into the image space, from which we will get the coarse depth map. We then train a shading net that takes the color image and current depth map as input and predicts a refined depth map with surface details. We have built up a relatively small dataset that contains color images, over-smoothed depth maps, and corresponding ground truth depth maps that have good surface details. More detailed explanations on this dataset could be found in Section 3.4. We adopt a multi-stage training scheme with limited labeled data.

Following the training scheme proposed in [56], we firstly train a simple U-Net based encoder-decoder network [52] on our captured depth dataset taking the over-smoothed depth map and its corresponding color image as input. The network is trained as supervised by the ground truth depth maps. Next, we adopt this network on the real images of our human body dataset to obtain enhanced depth maps. As we only have limited supervised data, the network may not generalize well to our real images. Therefore, to finally get depth maps with great surface details consistent with the color images, we train our shading net, which is also a U-Net based network on real images. The network is trained with both the flow prediction and image generation networks to synthesize a complete texture map from the visible part. Previous methods used.

\[
\begin{align*}
\mathbf{L}^* &= \arg \min _{l} \| \rho \sum _{k=1}^{9} l_k H_k(n_{\text{coarse}}) - I \|_2^2 \\
\end{align*}
\]

3.4 Implementation Details

We use the pre-trained model in the HMR-net, then train Joint-Net, Anchor-Net, and shading net successively. We use the ‘Adam’ optimizer to train these networks, with the learning rate as 0.0001. The handle weight in Laplacian edit is 10 for joint deforming and is 1 for anchor deforming.

To provide better training data to the shading net, we have captured a small depth dataset with a Kinect V2 sensor. It consists of 2200 depth frames with three human subjects wearing different clothes under various poses. The captured depth maps are further enhanced using traditional shading refinement techniques [45], [80] to recover small surface details, which can be taken as ground truth depth maps for supervised learning. We have magnified the shape details by 10 times during the test time.

4 Texture Completion

Synthesizing complete texture for the reconstructed human model from a single image is also a challenging problem, since only less than half of the texture is visible and can be retrieved from the input image. Besides, the imperfect fitting between the reconstructed human shape and the input image brings more difficulties to this problem.

As shown in Figure 4, our main idea is to take advantage of the flow prediction and image generation networks to synthesize a complete texture map from the visible part. Previous methods...
[46], [75], [78] have demonstrated that the neural networks can be used to synthesize novel views for rigid objects from specific categories, like cars and chairs. However, these methods failed for the synthesis of humans. The reason is that as compared to those rigid objects, very complex body poses are involved for the humans which result in misalignment for the features from different samples and significantly increase the difficulty for the synthesis [78]. Therefore, in this paper, we propose to predict the complete texture in the UV coordinates to avoid the interference of the misaligned features. As the reconstructed meshes are in uniform topology, the texture features are roughly aligned in the UV coordinates. The minor misalignment caused by inaccurate shape recovering can be corrected by the deep neural networks.

### 4.1 Training Data

We use the synthetic models from SURREAL dataset [64] to train our texture synthesizer. Specifically, we use 929 models with different appearances in SURREAL dataset to generate train/test data. For each appearance, we construct 20 SMPL models with randomly generated shape parameters and also with various poses randomly selected from the UP dataset [32]. We render each generated model with 5 views randomly chosen from 54 viewpoints corresponding to 3 pitch angles (−20°, 0°, 20°) and 18 azimuth angles in the range [0°, 340°] with interval 20°. The image background is randomly sampled from the Places dataset [74]. We have generated 92900 images in total using the above-mentioned process. The generated data is divided into the train and validate set, among which the train set accounts for 90%, and the validate set accounts for 10%. We train the network with the train set and evaluate the performance of the trained model using validate set, and finally, show the results on the real image from the WILD dataset in the experiment section.

Since our reconstructed mesh can not be perfectly aligned with the input image, the color spilling artifacts exists in the visible parts. To make our network robust to the misalignment, we augment the training data by introducing random bias. Specifically, we perturb the generated model and viewpoint parameters with uniformly distributed noise proportional to the parameter value. Then we detect visible faces on the mesh based on the perturbed parameters. For each visible face, we project corresponding pixels in the image to the UV space to get the partial texture and binary visibility mask. As demonstrated in the experiments, our texture synthesis network has better generalization ability when tested on wild images with this strategy.

### 4.2 Appearance Flow for Texture Completion

Our goal is to generate a complete and plausible texture map from this partial texture map. To this end, we map the visible texture into the UV coordinates. Different from other image inpainting problems where the masked region is only a small proportion concerning the original image, in this case, the invisible part is typically more than half and irregular. It has been shown that convolutional neural networks process image features with local convolution kernel layer by layer and thus are not effective for borrowing features from distant spatial location [68]. Also, directly inpainting images with large missing parts tend to produce artifacts such as blurriness and color discrepancy. Some methods such as contextual attention [68] and shift operation [67] have been proposed to deal with this problem. However, these approaches are designed for rectangular masks and are not trivial to generalize to masks with arbitrary shapes.

Inspired by recent advances in novel view synthesis [46], [75] and human pose transfer [35], [43], we propose to use appearance flow to complete the texture map. Specifically, taking the partial texture \( I_{\text{part}} \) and binary mask \( M \) as input, our network predicts a dense flow field of the same size as the texture. Then we use the differentiable bilinear sampling layer introduced in [24] to get the pixel value in the output image \( I_{\text{flow}} \) based on the predicted flow field. L1 loss is used to train the flow net:

\[
L_{\text{flow}} = ||I_{\text{gt}} - I_{\text{flow}}||_1
\]  

(4)

where \( I_{\text{gt}} \) is the ground truth complete texture in UV coordinates.

### 4.3 Symmetry Aware Texture Refinement

As our flow net is unable to hallucinate the missing texture, some pixels cannot be filled properly with existing pixels. To address this problem, we use a refinement network to further improve the texture predicted by the flow net. Considering that the human body is usually symmetrical, we flip the texture according to the symmetric correspondence and concatenate it to the original texture and mask, which is the input to our refinement network. We find that this strategy gives our texture a certain degree of symmetry, especially in the areas where the left and right seams are significantly improved. The refinement network generates a complete texture \( I_{\text{final}} \) of the same size as \( I_{\text{flow}} \). Following the state-of-the-art image enhancement [55] and image inpainting methods [42], we use a joint loss which consists of L1 loss \( L_{l1} \), adversarial loss \( L_{\text{adv}} \), perceptual loss \( L_{\text{perc}} \) [13], [25], and style loss \( L_{\text{style}} \) [55]. Our full objective function is formulated as:

\[
\min_G (\lambda_{\text{adv}} \max_D L_{\text{adv}}(G, D)) + \lambda_{\text{perc}} L_{\text{perc}}(G) + \lambda_{\text{style}} L_{\text{style}}(G) + L_{l1}(G)
\]

(5)

where \( G \) is the refinement net, D is the discriminator, \( \lambda_{\text{adv}}, \lambda_{\text{perc}} \) and \( \lambda_{\text{style}} \) are the weights of adversarial loss, perceptual loss, and style loss, respectively.

### 4.4 Implementation Details

#### Network architecture.

We use the network proposed by Nazeri et al. [42] as the backbone of both flow net and refinement net. We use a 70 × 70 PatchGAN architecture [23] for our discriminator. Spectral Normalization [40] is used in the refinement net and discriminator to stabilize the training.

#### Training setup.

Our flow net and refinement net are trained successively. The loss weights are set to \( \lambda_{\text{adv}} = 0.1, \lambda_{\text{perc}} = 0.1 \) and \( \lambda_{\text{style}} = 250 \) in all experiments. We train our networks using 256 × 256 images with batches of size 10. ‘Adam’ optimizer is...
used and learning rate is set to 0.0001. We randomly jitter the image color for data augmentation.

5 EXPERIMENT

5.1 Datasets

We have assembled three datasets for the experiment: the WILD dataset which has a large number of images with sparse 2D joints and segmentation annotated, and two other small datasets for evaluation in 3D metrics.

**WILD Dataset.** We assemble a quite large dataset for training and testing by extracting from 5 human datasets including MPII human pose database (MPII) [6], Common Objects in Context dataset (COCO) [37], Human3.6M dataset (H36M) [21], [22], Leeds Sports Pose dataset (LSP) [26] and its extension dataset (LSPET) [27]. As most of the images are captured in an uncontrolled setup, we call it the WILD dataset. The Unite the People (UP) dataset [32] provides ground truth silhouettes for the images in LSP, LSPET, and MPII datasets. As we focus on human shape recovery of the whole body, images with partial human bodies are removed based on the following two rules:

- All joints exist in the images.
- All joints are inside the body silhouette.

For COCO and H36M dataset, we further filter out the data with low-quality silhouettes. We separate the training and testing data according to the rules of each dataset. The numbers of the data we use are listed in Table 2.

The main drawback of WILD dataset is the lack of 3D ground truth shape. Though the UP dataset provides the fitted SMPL mesh for some data, the accuracy is uncertain. To help evaluate the 3D accuracy, we make two other small datasets with ground truth shape:

**RECON Dataset** We reconstruct 25 human mesh models using the traditional multi-view 3D reconstruction methods [14]. We render each model to 6 views and the views are randomly selected from 54 candidate views, of which the azimuth ranges from 0° to 340° with intervals 20°, and the elevation ranges from −10° to +10° with intervals of 10°. We use various scene images from the Places dataset [74] as background.

**SYN Dataset** We render 300 synthetic human mesh models in PVHM dataset [78] following their rendering setup, with the random scene images from the Places dataset as background. The meshes of PVHM include the inner surface, which is a disturbance for surface accuracy estimation. To filter out the inner surface, we project the mesh to the viewpoints in 6 orthogonal directions and remove the faces which are invisible in all 6 viewpoints.

For RECON dataset and SYN dataset, the reconstructed 3D meshes are scaled so that the mean height of the human body is close to the general body height of the common adult. In this way, we could measure the 3D error in mm.

5.2 Performance Evaluations

We measure the accuracy of the recovered shape with several metrics (corresponding to the second row in Table 1). For all test sets, we report the silhouette Intersection over Union (referred to as sil IoU), which measures the matching rate of the projected silhouette of the predicted 3D shape and the image silhouette. For the WILD dataset, we measure the difference between the projected 2D joints of the predicted 3D shape and the annotated ground truth joints. The joints of the mesh are extracted by computing the geometric center of the corresponding joint handle vertices. For the RECON dataset and SYN dataset, we also report the Chamfer distance (referred to as 3D err), which is the average distance of vertices between the predicted mesh and the ground truth mesh. We find the closest vertices in the resulting mesh for each vertex in the ground truth mesh and compute the mean of their distances as the 3D error.

The results selected based on the rank of silhouette IoU are shown in Figure 7. We could see in columns of the left side, the person with a simple pose like standing yields a pretty good fit. As we go from left to right columns, sports in the images are getting more complicated and the corresponding human shape is harder to predict. And in the right side columns, our method fails to predict humans with accessories (helmet, gloves) and under extremely twisting poses. In summary, the performance of our method is mostly affected by the complexity of the human pose and articulation.

5.3 Staging Analysis

We show the staging results in Figure 5 (right four columns) and report the quantitative evaluation of each stage in Table 1. The results in different phases are named as HMD-joint, HMD-anchor, and HMD-vertex (final result). We can see that the shape has got refined stage by stage. In the joint deformation phase, the joint correction takes effect to correct the displacement of joints. In the anchor deformation phase, silhouette supervision plays a key role in fitting the human shape. In the vertex deformation stage, the shape details are recovered to produce a visually plausible result.

**Ablation study.** We report the result of the ablation experiment in Table 3, where (w) means the window-cropped input, and (f) means the full image input. The evaluation is based on the WILD dataset, which is referred to in Section 5.1. We demonstrate two following statements: (1) By comparing the performance between full image input (No. 2 and 4) and window-crop image input (No. 3 and 5) in the table, we could see that the window-crop input predicts much higher silhouette IoU and lower joint error comparing to full image input, while the model size of the window-crop network is only 41% of the full image network. The reason why it has got a better result is that the window-crop network inherently focuses on the handle as the input center, so the problem turns to predict the local fit for each handle, which is easier to learn. (2) By comparing the performance between the integration of ‘joint + anchor’ deformation (No. 6) and only anchor or joint deformation (No. 3 and 5), we find that the combination achieves the best performance, and shows larger improvement than the pure anchor deformation.

To further validate the photometric loss as we train the ShadingNet, we also did the ablation study for the photometric loss term and computed the error of the refined depth on our captured depth dataset as described in Section 3.3. In total, we have 2272 frames and we randomly select 1818 frames for training and other 455 frames for testing. As shown in the table below, the surface error decreased after incorporating the photometric loss.

**Prediction with the silhouette.** By default our method takes the RGB image as input, and it also can use additional silhouettes as input. The method using additional silhouette shares the same framework with the default setting, and the difference is explained in Section 3.2. We show the qualitative comparison result in the last column in Figure 5 and the quantitative result in the last
Fig. 5: We compare our method on 3D human model reconstruction with previous approaches. The results of our method in different stages are shown: joint deformed, anchor deformed and vertex deformed (final result). Comparing to other methods, our method recovers more accurate joints and the body with shape details. The human body shape fits better to the input image, especially in body limbs. The rightmost column shows we can get more accurate recovered shapes when ground truth human silhouette is enforced (labeled as HMD (+Sil.)). Note that the images are cropped for the compact layout.

three rows in Table 1. As expected, the prediction with silhouette produces better results in all metrics.

5.4 Comparison with Other Methods

We compare our method with other methods with qualitative results as shown in Figure 5 and quantitative results in Table 1. We use the trained model of BodyNet and HMR provided by the authors. As BodyNet requires 3D shapes for training, they don’t use COCO and H36M datasets. To be fair, the evaluation on the WILD datasets only uses the data from LSP, LSPET, and MPII, which are the intersection of datasets used in all estimated methods. Comparing to SMPL based methods (SMPLify and HMR), our method has got the best performance in all metrics on all three datasets. As compared with BodyNet, a volumetric-based prediction method, we have got comparable scores in 3D error on RECON dataset. The reason is that the BodyNet produces more conservative shapes instead of focusing on the recovery of a complete human model. In some cases, the body limbs have not got reconstructed by the BodyNet when they are not visible from the image, while we always have the complete body recovered even though some parts of limbs haven’t appeared in the image. This makes it easy to have a better registration to the ground-truth mesh resulting in smaller 3D error. However, their scores on SYN datasets are lower than the other two datasets, since the human subjects from the SYN dataset generally have slim body shapes in which case the BodyNet results are degraded.

5.5 3D Error Analysis

Figure 8 shows our recovered 3D model on the RECON and SYN datasets together with the ground truth mesh. We show that the inherent pose and shape ambiguities cannot be resolved with the image from a single viewpoint. As we can see in Figure 8, the human shapes seen from the side view are quite different from the ground truth model even though they could fit closely to the input image. The estimated depth cue from a single image is sometimes ambiguous for shape recovery. This observation
Fig. 6: We compare our method with DeepHuman [73], PIFu [53], and Tex2Shape [5]. These three methods are all trained using ground-truth 3D human shapes. It is worth noting that the input image to PIFu has been segmented using the ground-truth silhouette, while the inputs of the other methods are original images.

explains the reason why the improvement of our method in 2D metrics is relatively larger than the improvement in 3D metrics.

We also evaluate MPJPE after rigid alignment as ‘3D joint error’ (defined as ‘Reconst. Error’ in HMR [28]) on H36M dataset. To derive joints from our models, we use the unified joints definition defined in Section 3.1. The results are reported in Table 6. We can see that: (1) Our method is superior in both 2D and 3D joint error comparing to HMR. The improvement of 3D joint error is slightly smaller than that of 2D joint error because there are few changes in the depth direction. (2) The 2D error of H36M dataset is generally smaller than that of WILD dataset, because the poses in H36M are relatively simpler than the other data in WILD dataset. (3) Adding silhouette as input leads to smaller improvement in H36M dataset than in WILD dataset, because the background in H36M images is pure colored, so there is less distraction from varying backgrounds.

5.6 Texture Synthesis

The texture synthesis results are shown in Figure 9, from which we can see that our method can synthesize plausible texture for the invisible part and even for the human face on the reverse side.

Ablation Study. We perform the ablation study to demonstrate the effectiveness of each component in our texture synthesis framework. We randomly choose 836 texture maps in SURREAL dataset as training data, leaving 93 texture maps for testing. The pose parameters in UP dataset are divided into 90\% for training and 10\% for testing. Then we synthesize the training and testing set as explained in Section 4.1. The evaluation results are shown in Table 4, where the explanation of each test are:

- Without perturb. We train our network on a synthetic dataset without adding random noise. Specifically, we detect visible faces on the mesh according to the ground truth SMPL and viewpoint parameters. Then we only keep visible parts in the texture as our partial texture.
Fig. 7: The results selected according to the percentage of the silhouette IoU.

![Image](source image) ![Image](ground truth) ![Image](ours)

source view side view
source image ground truth ours ground truth ours
source view side view
source image ground truth ours ground truth ours

Fig. 8: We show some recovered meshes and the ground truth meshes on the RECON (left) and SYN dataset (right). The meshes are rendered in the side view by rotating the mesh 90° around the vertical axis.

- Without flow. This model has the same architecture as our refinement network. We take the partial texture directly as input and train the network to hallucinate the missing region.
- Without symmetry. This model is identical to our complete method except that the input to the refinement network doesn’t contain the symmetrically flipped texture.
- Full. This is the complete architecture we described in the texture synthesis section.

Table 4 and Figure 10 show the quantitative and qualitative results of the ablation study. From the qualitative comparison, we can see that the results without perturbing contain textures from the background, which means the model cannot tolerant the interfering texture caused by inaccurate shape prediction. The results without flow contain obvious color distortion, which is a common problem for image generation networks. The results without symmetry also show color distortion, and the seam of the UV texture map in the back is inconsistent. The results of full architecture are visually most plausible in the ablation experiments. In the quantitative comparison, we can see that the network trained on the dataset without noise predicts worse results on the testing set, as the recovered mesh cannot be perfectly aligned to the image. Our full architecture scores the highest in all metrics.

Comparison. We compare our texture completion method with Lazova et al.’s work [33] in Figure 11. The source images, models, and the results are provided by the authors. To make the comparison clear, the predicted textures are mapped to the same canonical A-pose meshes. Lazova et al.’s work [33] uses the correspondence estimated by DensePose [16] to generate the input partial texture, while our method uses the projective texture from the recovered human shape as input. By comparison, we can see that Lazova et al.’s work generates slightly better facial texture, as DensePose aligns better texture alignment in the facial region. Our method can preserve more texture details in the main body. For example, the logo on the T-shirt in (a) and (d), and the coat edges in (b) are preserved by our method but are blurred in Lazova et al.’s result. Besides, Lazova et al.’s inpainting model also introduces some artifacts and blurriness into the restored texture. Our method shows better texture quality in the back area due to the sophisticated two-stage flow-refine scheme. Our symmetry-input design alleviates the inconsistent seam in the back that Lazova et al.’s work suffers, as shown in the back view of (b) and (d).

5.7 Comparison with 3D-Supervised Methods

In Figure 6, we show some sample results for qualitative comparison with DeepHuman [73], PIFu [53] and Tex2Shape [5]. It is worth noting that these three methods all take the ground-truth 3D
Fig. 9: We show some texture synthesis results. For each group, we show the source image and render the recovered mesh with predicted texture in the front view and back view. Though there is slight color distortion comparing to the source image, our method is able to predict plausible texture, and can even hallucinate the completely invisible face from the back.

### TABLE 1: Quantitative evaluation

| method            | WILD dataset | SYN dataset |
|-------------------|--------------|-------------|
|                  | sil IoU      | 2D joint err | sil IoU | 3D err full | 3D err visible | sil IoU | 3D err full | 3D err visible |
| SMPlify [10]      | 66.3         | 10.19       | 73.9    | 52.84       | 51.69         | 71.0    | 62.31       | 60.67         |
| BodyNet [63]      | 68.6         | —           | 72.5    | 43.75       | 40.05         | 70.0    | 54.41       | 46.55         |
| HMR - joint       | 67.6         | 9.90        | 74.3    | 51.74       | 42.05         | 71.7    | 53.03       | 47.75         |
| HMR - anchor      | 76.5         | 8.82        | 85.0    | 44.60       | 39.73         | 79.6    | 47.18       | 44.62         |
| HMD - vertex      | —            | —           | —       | —           | —             | 44.75   | —           | 41.90         |
| HMD(+sil) - joint | 73.0         | 8.31        | 79.2    | 50.49       | 40.88         | 77.7    | 48.41       | 45.16         |
| HMD(+sil) - anchor| 82.4         | 8.22        | 88.3    | 43.50       | 38.63         | 85.7    | 44.59       | 42.68         |
| HMD(+sil) - vertex| —            | —           | —       | —           | —             | 43.22   | 40.98       | 39.11         |

* ‘full’ means the full body shape is used for error estimation, and ‘vis’ means only the visible part concerning the input image viewpoint is used for error estimation.

The statistic unit of 3D error is millimeter; the 2D joint error is measured by pixel. The methods beyond the cutting line use only RGB image as input, while the methods under the cutting line use ‘RGB + silhouette’ as input. Some statistic is blank: the joint position cannot be derived directly from the mesh produced by BodyNet; The sil IoU and 2D joint error after vertex deformation stay the same as anchor deformed results, as the vertex deformation is only along the Z-axis, which is vertical to the silhouette in the image plane.

### TABLE 2: WILD dataset components

| data source | LSP | LSPE1 | MPII | COCO | H36M |
|-------------|-----|-------|------|------|------|
| train num   | 987 | 5376  | 8035 | 4004 | 5747 |
| test num    | 703 | 0     | 1996 | 606  | 1320 |

### TABLE 3: Ablation experiments. In this table, (w) means the results when taking the window-cropped as input, and (f) means the results with the full image as input.

| num | method               | sil IoU | 2D joint err/ pixel |
|-----|----------------------|---------|---------------------|
| 1   | baseline(initial shape) | 67.6%   | 9.30               |
| 2   | joint (f)            | 68.3%   | 9.85               |
| 3   | joint (w)            | 70.7%   | 8.81               |
| 4   | anchor (f)           | 70.1%   | 9.89               |
| 5   | anchor (w)           | 71.3%   | 9.75               |
| 6   | joint (w) + anchor (w) | 76.5%   | 8.82               |

### TABLE 4: Ablation study of texture synthesis.

| method       | PSNR | MAE  | SSIM |
|--------------|------|------|------|
| w/o. perturb | 17.62| 0.1710| 0.5919|
| w/o. flow    | 19.48| 0.1223| 0.6699|
| w/o. symmetry| 19.58| 0.1200| 0.6780|

### TABLE 5: Ablation study of photometric loss.

| method                  | error(mm) |
|-------------------------|------------|
| Without Photometric loss| 1.346      |
| With Photometric loss   | 1.129      |

### TABLE 6: Joint error evaluation on H36M dataset. HMD-j means the result after joint deforming stage, and HMD-a means the result after joint and anchor deforming stage. (s) means using the ground-truth silhouette as input.

| method     | HMR  | HMD-j | HMD-a | HMD(s)-j | HMD(s)-a |
|------------|------|-------|-------|----------|----------|
| 3D err (mm)| 56.5 | 49.6  | 50.9  | 49.2     | 50.4     |
| 2D err (mm)| 5.42 | 3.96  | 4.19  | 3.77     | 4.11     |

By contrast, our joint net and anchor net of our method was trained on the dataset collected from in-the-wild images without 3D ground-truth models as supervision, and the shading net of our method requires the wild images and a small number of depth maps for training.
We render images for BUFF dataset, it consists of 100 pairs of models and images, and SYN datasets have been presented in Section 5.1. For the SYN dataset, and BUFF dataset [71]. The details of RECON error in Table 7. We run the evaluation on RECON dataset, revised methods mentioned above and report the surface and IoU with the input image.

Shapes together with the detailed geometric structure aligned well with the input image. However, the reconstructed surface is over-smoothed and maintain the overall shape and could recover middle-scale garment are upright-posed human models. Tex2Shape has a good ability to the limited diversity of poses in RenderPeople, most of which broken arms (line 1), and duplicated limbs (line 3). This is caused in some complex poses, leading to abnormal shape (line 2/3/5/6), detailed wrinkles in the front view, but its performance is poor in some complex poses, and thus achieved stable performance overall. These comparison methods achieved better results with smaller 3D surface error but they all rely on ground-truth 3D models for training while our method only uses 2D supervision, such as 2D joints. We got comparable results on RECON dataset containing humans under complex poses, and a higher score in IoU of the silhouette.

5.8 Performance of 3D-supervised Training

The results of 3D-supervised training are shown in Figure 12. By comparing the 3D-supervised and semi-supervised results, we can see both schemes can recover geometric details like the wrinkles and the belt, but the 3D-supervised results are closer to the real human 3D shape with fewer texture-copying effects. For example, for the second person in the second column of the figure, the semi-supervised method has taken the texture of the shirt as the geometric surface details, and thus created the wrong wrinkles. Besides, the semi-supervised method tends to yield excessive shape change. By contrast, 3D-supervised schemes recover stable 3D detailed shape, and we believe this is because the Twindom dataset provides abundant shape and appearance together with the 3D human shape. But the limitation of the 3D-supervised scheme is that it relies on an expensive commercial high-quality 3D human dataset.

5.9 View Synthesis

In Figure 13, we use the predicted model by our method and HMR [28] for view synthesis task. To this end, we first assign the color of the pixels in the source image to the vertices of the predicted mesh, then render the model in the novel view to generate the result images. To further refine the synthesis result, we expand the texture in the foreground part to avoid artifacts in the boundary region, which is labeled as 'HMD-r' in the figure. As the predicted HMR model couldn’t fit well to the source image, the artifacts exist in the boundary region. By contrast, we get much better results with visually appealing synthesis using our texture completion method. We also compare our results with VSPV [78], which is an image-to-image synthesis method. We find that VSPV fails to track the right pose when the pose is complex, and the rotation of viewpoint is large (60 degrees in

| method        | RECON dataset | SYN dataset | BUFF dataset |
|---------------|---------------|-------------|--------------|
|               | err full     | err visible | sil IoU      | err full     | err visible | sil IoU      | err full     | err visible | sil IoU      |
| PIFu [53]     | 35.67*       | 32.77*      | -*           | 18.90*       | 17.68*      | -*           | 26.65*       | 27.33*      | -*           |
| PIFuHD [54]   | 52.69        | 48.55       | 58.4         | 22.17        | 20.55       | 56.53        | 22.01        | 24.90       | 87.8         |
| DeepHuman [73]| 38.58        | 37.26       | 71.2         | 19.34        | 18.60       | 60.3         | 34.97        | 37.49       | 83.1         |
| Tex2Shape [5] | 44.7         | 44.39       | 61.2         | 26.86        | 27.82       | 54.8         | 58.80        | 61.03       | 55.3         |
| HMD           | 44.1         | 41.76       | 85.0         | 44.75        | 41.90       | 79.6         | 41.30        | 40.02       | 80.7         |

Please refer to Table 1 for the explanation of metrics.

* PIFu requires an additional ground-truth mask as input, while all other methods take only source images as input.

Fig. 10: Qualitative ablation comparison of texture synthesis.
the figure). By contrast, our method synthesizes more plausible images.

6 Conclusion

In this paper, we have proposed a novel approach to reconstruct the detailed human avatar from a single image, which is the 3D human shape with texture. Starting from an SMPL model based human recovery method, we introduce free-form deformations to refine the body shapes with a project-predict-deform strategy. A hierarchical framework has been proposed for restoring more accurate and detailed human bodies under the supervision of joints, silhouettes, and shading information. Furthermore, we use the neural network to synthesize the complete texture for the recovered 3D mesh using the single image. We have performed extensive comparisons with state-of-the-art shape recovering methods and demonstrated significant improvements in both quantitative and qualitative assessments. Our method also generates plausible texture for the recovered 3D model, and can even hallucinate invisible face from the back.

The limitation of our work is that the pose ambiguities are not solved, and there are still large errors in predicted body meshes especially in depth direction. The results for the human in relatively unusual poses are sometimes not good. For the texture completion part, due to the limitation of the training dataset, the predicted texture is inclined to the appearance in SURREAL dataset.

Acknowledgements

This work was supported by the NSFC grant 62025108, 62001213, 61627804, USDA grant 2018-67021-27416, and NSERC Discovery Grant.

References

[1] Renderpeople dataset. https://renderpeople.com/.
[2] Twindom dataset. https://web.twindom.com/.
[3] T. Alldieck, M. Magnor, B. Lal Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In Proceedings of the The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[4] T. Alldieck, M. A. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 8387–8397, 2018.
[5] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2shape: Detailed full human body geometry from a single image. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 2293–2303, 2019.
[6] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3686–3693, 2014.
[7] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. ACM transactions on graphics (TOG), 24(3):408–416, 2005.
[8] S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. ACM Transactions on Graphics (TOG), 33(4):159, 2014.
[9] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. Multigarment net: Learning to dress 3d people from images. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 5420–5430, 2019.
[10] 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 Proceedings of European Conference on Computer Vision (ECCV), pages 561–578, 2016.
[11] E. Dibra, H. Jain, A. C. Oztiireli, R. Ziegler, and M. H. Gross. Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5504–5514, 2017.
[12] E. Dibra, H. Jain, C. Oztiireli, R. Ziegler, and M. Gross. Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In Proceedings of IEEE International Conference on 3D Vision (3DV), pages 108–117, 2016.
[13] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. In Advances in Neural Information Processing Systems (NIPS), pages 658–666, 2016.
[14] Y. Furukawa, C. Hernandez, et al. Multi-view stereo: A tutorial. Foundations and Trends® in Computer Graphics and Vision, 9(1-2):1–148, 2015.
[15] P. Guan, A. Weiss, A. O. Balan, and M. J. Black. Estimating human shape and pose from a single image. In Proceedings of IEEE International Conference on Computer Vision (ICCV), pages 1381–1388, 2009.
[16] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7297–7306, 2018.
[17] R. A. Güler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zaferiou, and I. Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6799–6808, 2017.
[18] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt. Livecap: Real-time human performance capture from monocular video. ACM Transactions on Graphics (TOG), 38(2):1–17, 2019.
Fig. 12: Comparison between semi-supervised results and 3D-supervised results on the Twindom dataset (upper two rows) and WILD dataset (lower two rows). In each set of results, the left side is the rendered mesh model, and the right side is the rendered normal map.

Fig. 13: View synthesis results using our predicted model and other methods.

[19] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormählen, and H.-P. Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1823–1830, 2010.

[20] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung. Arch: Animatable reconstruction of clothed humans. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3093–3102, 2020.

[21] C. Ionescu, F. Li, and C. Sminchisescu. Latent structured models for human pose estimation. In Proceedings of IEEE International Conference on Computer Vision (ICCV), pages 2220–2227, 2011.

[22] 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 Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 36(7):1325–1339, 2014.

[23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1125–1134, 2017.

[24] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In Advances in Neural Information Processing Systems (NIPS), pages 2017–2025, 2015.

[25] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the European Conference on Computer Vision (ECCV), pages 694–711. Springer, 2016.

[26] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In Proceedings of the British Machine Vision Conference (BMVC), volume 2, page 5, 2010.

[27] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1465–1472, 2011.

[28] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7122–7131,
Hao Zhu received the Ph.D. and B.S. degree from the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. He was a visiting scholar in University of Kentucky. He is currently an associate researcher in Nanjing University. His current research interests include computer vision and deep learning, especially 3D reconstruction and 3D vision.

Xinxin Zuo received the M.E. degree from Northwestern Polytechnical University and Ph.D. degree from the University of Kentucky. She is currently a Postdoctoral Fellow at University of Alberta. Her research interests include computer vision and graphics, especially on 3D reconstruction and human modeling.

Haotian Yang received the B.S. degree from Nanjing University, Nanjing, China, in 2018, where he is currently pursuing the M.S. degree with the School of Electronic Science and Engineering. His current research interests include computer vision and computer graphics.

Sen Wang received the B.E. degree and Ph.D. degree from Northwestern Polytechnical University. From 2015 to 2016, he was a Visiting Ph.D. Student at the University of Kentucky. He is currently a Postdoctoral Fellow at University of Alberta. His research interests include computer vision and robotics.

Xun Cao received the B.S. degree from Nanjing University, Nanjing, China, in 2006, and the Ph.D. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2012. He held visiting positions with Philips Research, Aachen, Germany, in 2008, and Microsoft Research Asia, Beijing, from 2009 to 2010. He was a Visiting Scholar with the University of Texas at Austin, Austin, TX, USA, from 2010 to 2011. He is currently a Professor with the School of Electronic Science and Engineering, Nanjing University. His current research interests include computer vision and robotics.

Ruigang Yang is currently a full professor of Computer Science at the University of Kentucky. He obtained his PhD degree from University of North Carolina at Chapel Hill and his MS degree from Columbia University. His research interests span over computer graphics and computer vision, in particular in 3D reconstruction and 3D data analysis. He has published over 100 papers, which, according to Google Scholar, has received more than 10000 citations with an h-index of 52 (as of 2019). He has received a number of awards, including US NSF Career award in 2004 and the Dean’s Research Award at the University of Kentucky in 2013. He is a senior member of IEEE.