

DeepHuman: 3D Human Reconstruction from a Single Image

Zerong Zheng^{*}

Tao Yu^{*} [†]

Yixuan Wei^{*}

Qionghai Dai^{*}

Yebin Liu^{*}

^{*} Tsinghua University

[†] Beihang University

Abstract

We propose DeepHuman, an image-guided volume-to-volume translation CNN for 3D human reconstruction from a single RGB image. To reduce the ambiguities associated with the surface geometry reconstruction, even for the reconstruction of invisible areas, we propose and leverage a dense semantic representation generated from SMPL model as an additional input. One key feature of our network is that it fuses different scales of image features into the 3D space through volumetric feature transformation, which helps to recover accurate surface geometry. The visible surface details are further refined through a normal refinement network, which can be concatenated with the volume generation network using our proposed volumetric normal projection layer. We also contribute THuman, a 3D real-world human model dataset containing about 7000 models. The network is trained using training data generated from the dataset. Overall, due to the specific design of our network and the diversity in our dataset, our method enables 3D human model estimation given only a single image and outperforms state-of-the-art approaches.

1. Introduction

Image-based reconstruction of a human body is an important research topic for VR/AR content creation [7], image and video editing and re-enactment [19, 43], holoporation [40] and virtual dressing [42]. To perform full-body 3D reconstruction, currently available methods require the fusion of multiview images [8, 25, 20] or multiple temporal images [3, 2, 1] of the target. Recovering a human model from a single RGB image remains a challenging task that has so far attracted little attention. Using only a single image, available human parsing studies have covered popular topics starting from 2D pose detection [41, 6, 39], advancing to 3D pose detection [33, 44, 64], and finally expanding to body shape capture [27] using a human statistic template such as SMPL [32]. However, the statistic template can capture only the shape and pose of a minimally clothed body and lack the ability to represent a 3D human model under a normal clothing layer. Although the most recent work, BodyNet[52], has pioneered research towards this goal, it



Figure 1: Given only a single RGB image, our method automatically reconstructs the surface geometry of clothed human body.

only generates nearly undressed body reconstruction results with occasionally broken body parts. We believe that 3D human reconstruction under normal clothing from a single image, which needs to be further studied, will soon be the next hot research topic.

Technically, human reconstruction from a single RGB image is extremely challenging, not only because of the requirement to predict the shape of invisible parts but also due to the need for the geometry recovery for visible surface. Therefore, a method capable of accomplishing such a task should meet two requirements: first, the degrees of freedom of the output space should be constrained to avoid unreasonable artifacts (e.g., broken body parts) in invisible areas; second, the method should be able to efficiently extract geometric information from the input image, such as clothing styles and wrinkles, and fuse them into the 3D space.

In this paper, we propose DeepHuman, a deep learning-based framework aiming to address these challenges. Specifically, to provide a reasonable initialization for the network and constrain the degrees of freedom of the output space, we propose to leverage parametric body models by generating a 3D semantic volume and a corresponding 2D semantic map as a dense representation after estimating the shape and pose parameters of a parametric body template (e.g., SMPL[32]) for the input image. Note that the requirement of inferring a corresponding SMPL model for an image is not strict; rather, several accurate methods are available for SMPL prediction from a single image[5, 27]. The input image and the semantic volume&map are fed into an image-guided volume-to-volume translation CNN for sur-

face reconstruction. To accurately recover surface geometry like the hairstyle or cloth contours to the maximum possible extent, we propose a multi-scale volumetric feature transformation so that those different scales of image guidance information can be fused into the 3D volumes. Finally, we introduce a volumetric normal projection layer to further refine and enrich visible surface details according to the input image. This layer is designed to concatenate the volume generation network and the normal refinement network and enables end-to-end training. In summary, we perform 3D human reconstruction in a coarse-to-fine manner by decomposing this task into three subtasks: a) parametric body estimation from the input image, b) surface reconstruction from the image and the estimated body, and c) visible surface detail refinement according to the image.

The available 3D human dataset [53] used for network training in BodyNet [52] is essentially a set of synthesized images textured over SMPL models [32]. No large-scale human 3D dataset with surface geometry under normal clothing is publicly available. To fill in this gap, we present the THuman dataset. We leverage the state-of-the-art DoubleFusion [63] technique for real-time human mesh reconstruction and propose a capture pipeline for fast and efficient capture of outer geometry of human bodies wearing casual clothes with medium-level surface detail and texture. Based on this pipeline, we perform capture and reconstruction of the THuman dataset, which contains about 7000 human meshes with approximately 230 kinds of clothes under randomly sampled poses.

Our network learns from the training corpus synthesized from our THuman dataset. Benefiting from the data diversity of the dataset, the network generalizes well to natural images and provides satisfactory reconstruction given only a single image. We demonstrate improved efficiency and quality compared to current state-of-the-art approaches. We also show the capability and robustness of our method through an extended application on monocular videos.

2. Related Work

Human Models from Multiview Images. Previous studies focused on using multiview images for human model reconstruction [26, 47, 30]. Shape cues like silhouette, stereo and shading cues have been integrated to improve the reconstruction performance [47, 30, 58, 57, 55]. State-of-the-art real-time [11, 10] and extremely high-quality [8] reconstruction results have also been demonstrated with tens or even hundreds of cameras using binocular [12] or multiview stereo matching [13] algorithms. To capture detailed motions of multiple interacting characters, more than six hundred cameras have been used to overcome the occlusion challenges [24, 25]. However, all these systems require complicated environment setups including camera calibration, synchronization and lighting control.

To reduce the difficulty of system setup, human model reconstruction from extremely sparse camera views has re-

cently been investigated by using CNNs for learning silhouette cues [15] and stereo cues [20]. These systems require about 4 camera views for a coarse-level surface detail capture. Note also that although temporal deformation systems using lightweight camera setups [54, 9, 14] have been developed for dynamic human model reconstruction using skeleton tracking [54, 31] or human mesh template deformation [9], these systems assume a pre-scanned subject-specific human template as a key model for deformation.

Human Models from Temporal Images. To explore low-cost and convenient human model capture, many studies try to capture a human using only a single RGB or RGBD camera by aggregating information from multiple temporal frames. For RGBD images, DynamicFusion [38] breaks the static scene assumption and deforms the non-rigid target for TSDF fusion on a canonical static model. BodyFusion [62] have tried to improve the robustness by adding articulated prior. DoubleFusion [63] introduced a human shape prior into the fusion pipeline and achieved state-of-the-art real-time efficiency, robustness, and loop closure performance for efficient human model reconstruction even in cases of fast motions. There are also offline methods for global registration of multiple RGBD images to obtain a full-body model [29]. To reconstruct a human body using a single-view RGB camera, methods have been proposed for rotating the camera while the target remains as static as possible [65], or keeping the camera static while the target rotates [3, 2, 1]. Recently, human performance capture that can reconstruct dynamic human models using only a single RGB camera has been proposed [59, 18]; however, similar to the multicamera scenario [54, 9, 14], such approaches require a pre-scanned human model as input.

Human Parsing from a Single Image. Parsing human from a single image has recently been a popular topic in computer vision. The research can be categorized into sparse 2D parsing (2D skeleton estimation) [6, 39], sparse 3D parsing (3D skeleton estimation) [33, 44, 64, 48, 50, 35, 61], dense 2D parsing [17] and dense 3D parsing (shape and pose estimation). Dense 3D parsing from a single image has attracted substantial interest recently because of the emergence of human statistical models like SCAPE [4] and SMPL [32]. For example, by fitting the SCAPE or SMPL model to the detected 2D skeleton and other shape cues of an image [5, 28], or by regressing [27, 49, 51] the SMPL model using CNNs, the shape and pose parameters can be automatically obtained from a single image.

Regarding single-view human model reconstruction, there are only several recent works by Varol et al.[52], Jackson et al.[23] and Natsume et al.[36]. In the first study, the 3D human datasets used for network training lacks geometry details, leading to SMPL-like voxel geometries in their outputs. The second study shows the ability to output high-quality details, but their training set is highly constrained, leading to difficulty in generalization, e.g., to different human poses. The concurrent work by Natsume et al.[36] pre-

dicts multiview 2D silhouettes to reconstruct the 3D model, but their reconstruction results have limited pose variation.

3D Human Body Datasets. Most of the available 3D human datasets are used for 3D pose and skeleton detection. Both HumanEva [46] and Human3.6M [21] contain multi-view human video sequences with ground-truth 3D skeleton annotation obtained from marker-based motion capture systems. Because of the need to wear markers or special suits, both datasets have limited apparel divergence. MPI-INF-3DHP [34] dataset enriches the cloth appearance by using a multiview markerless mocap system. However, all these datasets lack a 3D model of each temporal frame. To meet the requirement of pose and shape reconstruction from a single image, the synthesized SURREAL [53] datasets have been created for this task by rendering SMPL models with different shape and pose parameters under different clothing textures. The “Unite the People” dataset [28] provides real-world human images annotated semi-automatic with 3D SMPL models. These two datasets, in contrasts to our dataset, do not contain surface geometry details.

3. Overview

Given an image of a person in casual clothes, denoted by \mathbf{I} , our method aims to reconstruct his/her full-body 3D surface with plausible geometrical details. Directly recovering a surface model of the subject from the image is very challenging because of depth ambiguities, body self-occlusions and the high degree of freedom of the output space. Therefore, we perform 3D human reconstruction in a coarse-to-fine manner. Our method starts from parametric body estimation, then performs full-body surface reconstruction and finally refines the details on the visible areas of the surface.

We exploit the state-of-the-art methods HMR[27] and SMPLify[5] to estimate a SMPL model from \mathbf{I} ; see the supplementary document for more details. To feed the SMPL estimation into the CNN, we predefine a *semantic code* (a 3-dimensional vector) for each vertex on SMPL according to its spatial coordinate at rest pose. Given the SMPL estimation, we render the semantic code onto the image plane to obtain a semantic map \mathbf{M}_s and generate a semantic volume \mathbf{V}_s by first voxelizing the SMPL model into the voxel grid and then propagating the semantic codes into the occupied voxels. Our dense semantic representation has three advantages: (1) it encodes information about both the shape and the pose of the body and thus provides a reasonable initialization for the network and constrain the degrees of freedom of the output space; (2) it provides clues about the corresponding relationship between 3D voxels and 2D image pixels; (3) it is easy to be incorporated into neural networks. More details are presented in the supplementary document.

For the surface geometry reconstruction, we adopt an occupancy volume to represent the surface[52]. Specifically, we define a 3D occupancy voxel grid \mathbf{V}_o , where the voxel values inside the surface are set to 1 and others are set to 0. All occupancy volumes have a fixed res-

olution of $128 \times 192 \times 128$, where the resolution of the y-axis is set to a greater value and it can be automatically adapted to the major axis of the observed human body. To reconstruct \mathbf{V}_o from \mathbf{V}_s with the assistance of \mathbf{I} and \mathbf{M}_s , we propose an image-guided volume-to-volume translation network (Sec.4.1), in which we use multiscale volumetric feature transformation (Sec.4.1.1) to fuse 2D image guidance information into a 3D volume. Accordingly, the network will take advantage of knowledge from both the 2D image and the 3D volume.

Due to resolution limitations, a voxel grid always fails to capture fine details such as clothing wrinkles. To further enrich and refine the geometrical details on the visible part of the surface, we propose to directly project a 2D normal map \mathbf{N} from \mathbf{V}_o (Sec.4.1.2) and refine it with a U-net (Sec.4.1). In other words, we encode the geometrical details of the visible surface using 2D normal maps and consequently lower the memory requirement.

To train the network with supervision, we contribute THuman, a real-world 3D human model dataset (Sec.5). We synthesize the training corpus from the dataset. Once the network is trained, it can predict an occupancy volume and a normal map of the visible surface given an image of a person and the corresponding SMPL estimation. We obtain the final reconstruction result by first extracting a triangular polygon mesh from the occupancy volume using the Marching Cube algorithm and then refining the mesh according to the normal map using the method in [37].

4. Approach

4.1. Network Architecture

Our network consists of 3 components, namely an image encoder \mathcal{G} , a volume-to-volume (vol2vol) translation network \mathcal{H} and a normal refinement network \mathcal{R} , as shown in Fig.2. The image encoder \mathcal{G} aims to extract multi-scale 2D feature maps $\mathbf{M}_f^{(k)}$ ($k = 1, \dots, K$) from the combination of \mathbf{I} and \mathbf{M}_s . The vol2vol network is a volumetric U-Net [60], which takes \mathbf{V}_s and $\mathbf{M}_f^{(k)}$ ($k = 1, \dots, K$) as input, and outputs an occupancy volume \mathbf{V}_o representing the surface. Our vol2vol network \mathcal{H} fuses multi-scale semantic features $\mathbf{M}_f^{(k)}$ ($k = 1, \dots, K$) into its encoder through a *multi-scale volumetric feature transformer*. After generating \mathbf{V}_o , a normal refinement U-Net [45] \mathcal{R} further refines the normal map \mathbf{N} after calculating it directly from \mathbf{V}_o through a *volume-to-normal projection layer*. All operations in the network are differentiable, and therefore, it can be trained or fine-tuned in an end-to-end manner. Implementation details are presented in the supplementary document.

4.1.1 Multi-scale Volumetric Feature Transformer

In this work, we extend the Spatial Feature Transformer (SFT) layer [56] to handle 2D-3D data pairs in the multi-scale feature pyramid, and propose multi-scale Volumet-

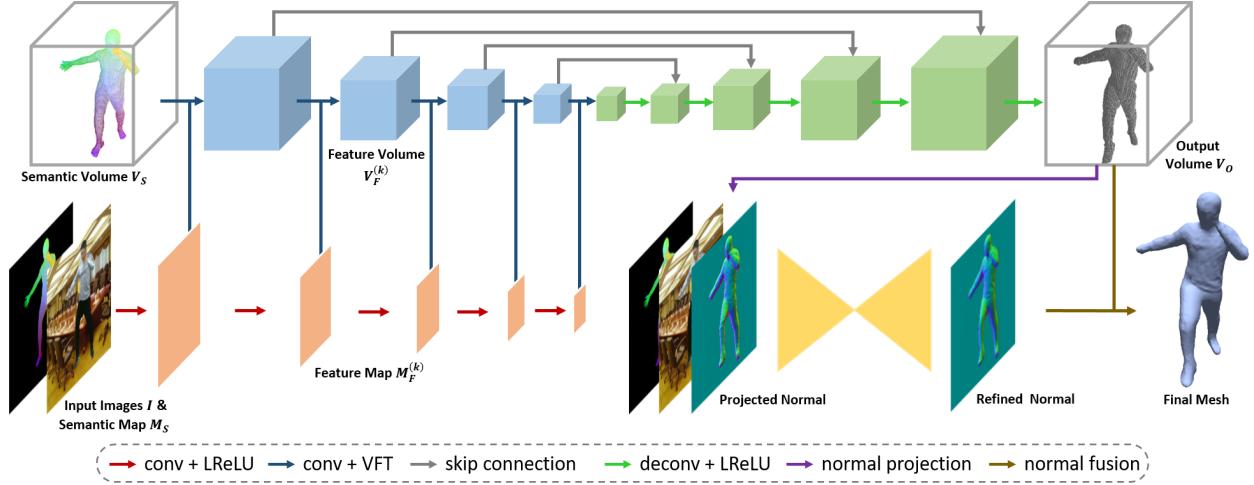


Figure 2: Network architecture. Our network is mainly composed of an image feature encoder (orange), a volume-to-volume translation network (blue & green) and a normal refinement network (yellow).

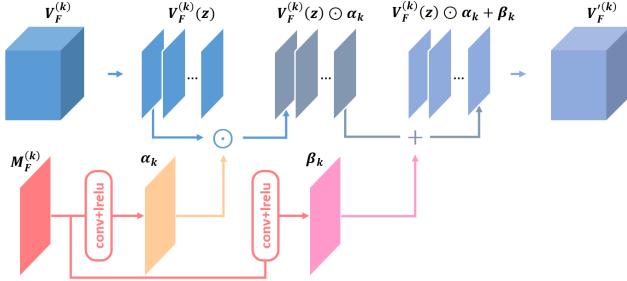


Figure 3: Illustration of volumetric feature transformation (VFT) at level k .

ric Feature Transformer (VFT). SFT was first used in [56] to perform image super-resolution conditioned on semantic categorical priors to avoid the regression-to-the-mean problem. A SFT layer learns to output a modulation parameter pair (α, β) based on the input priors. Then transformation on the feature map F is carried out as: $\mathcal{SFT}(F) = \alpha \odot F + \beta$, where \odot is Hadamard product.

In our network, at each level k , a feature volume $V_f^{(k)}$ (blue cubes in Fig.2) and a feature map $M_f^{(k)}$ (orange squares in Fig.2) are provided by previous encoding layers. Similar to [56], we first map the feature map $M_f^{(k)}$ to modulation parameters (α_k, β_k) through convolution+activation layers (see the second row of Fig.3). Note that the operation in $\mathcal{SFT}(\cdot)$ cannot be applied directly on $V_f^{(k)}$ and $M_f^{(k)}$ because of dimension inconsistency ($V_f^{(k)}$ has a z-axis while (α_k, β_k) doesn't.) Therefore, we slice the feature volume along the z-axis into a series of feature slices, each of which has a thickness of 1 along the z-axis. Then we apply the same element-wise affine transformation to each feature z -

slice independently:

$$\mathcal{VFT}\left(V_f^{(k)}(z_i)\right) = \alpha_k \odot V_f^{(k)}(z_i) + \beta_k \quad (1)$$

where $V_f^{(k)}(z_i)$ is the feature slice on plane $z = z_i$, $z_i = 1, 2, \dots, Z$ and Z is the maximal z-axis coordinate. The output of a VFT layer is the re-combination of transformed feature slices. Fig.3 is an illustration of VFT.

The superiority of VFT is three-fold. First, compared to converting feature volumes/maps into latent codes and concatenating them at the network bottleneck, it preserves the shape primitiveness of image/volume feature and thus encodes more local information. Second, it is efficient. Using VFT, feature fusion can be achieved in a single pass of affine transformation, without requiring extra convolutions or full connection. Third, it is flexible. VFT can be performed on either the original image/volume or downsampled feature maps/volumes, making it possible to fuse different scales of features and enabling much deeper feature transfer.

In order to integrate image features to the maximum possible extent, we perform volumetric feature transformation on the multi-scale feature pyramid; see the blue arrows/lines in Fig.2 for illustration. We only perform VFT in the encoder part of our vol2vol network; however, the transformation information can be propagated to the decoder through skip-connections. As discussed in Sec.6.3, the multi-scale feature transformation helps recover more accurate surface geometry compared to directly concatenating latent variables at the network bottleneck.

4.1.2 Volume-to-normal Projection Layer

Our goal is to obtain geometric details (e.g. wrinkles and cloth boundary) on the visible surface of the human model.

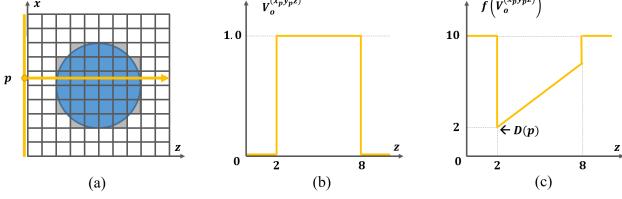


Figure 4: Illustration of differentiable depth projection.

However, a volume-based representation is unable to capture such fine-grain details due to resolution limitations. Thus, we encode the visible geometrical details on 2D normal maps, which can be directly calculated from the occupancy volume using our differentiable volume-to-normal projection layer. The layer first projects a depth map directly from the occupancy volume, transforms the depth map into a vertex map, and then calculates the normal maps through a series of mathematical operations.

Fig.4 is a 2D illustration explaining how the layer projects depth maps. In Fig.4(a), the blue circle is the model we aim to reconstruct, and the voxels occupied by the circle are marked in grey. Consider the pixel $p = (x_p, y_p)$ on the image plane as an example. To calculate depth value $D(p)$ of p according to \mathbf{V}_o , a straightforward method is to consider a ray along the z -axis and record the occupancy status of all voxels along that ray (Fig.4(b)). Afterwards, we can determine $D(p)$ by finding the nearest occupied voxel. Formally, $D(p)$ is obtained according to

$$D(p) = \inf \left\{ z \mid \mathbf{V}_o^{(x_p, y_p, z)} = 1 \right\} \quad (2)$$

where $\mathbf{V}_o^{(x_p, y_p, z)}$ denotes the value of the voxel at coordinate (x_p, y_p, z) . Although this method is straightforward, it is difficult to incorporate the operation, $\inf \{\cdot\}$ into neural networks due to the complexity of differentiating through it. Therefore, we transform the occupancy volume to a *depth volume* \mathbf{V}_d by applying a transformation f :

$$\mathbf{V}_d^{(xyz)} = f(\mathbf{V}_o^{(xyz)}) = M(1 - \mathbf{V}_o^{(xyz)}) + z\mathbf{V}_o^{(xyz)} \quad (3)$$

where M is a sufficiently large constant. Then as illustrated in Fig.4(c), $D(p)$ can be computed as:

$$D(p) = \min_z f(\mathbf{V}_d^{(x_p, y_p, z)}). \quad (4)$$

After depth projection, we transform the depth map to a vertex map \mathbf{M}_v by assigning x and y coordinates to depth pixels according to their positions on the images. Then Sobel operators are used to calculate the directional derivative of the vertex map along both the x and y directions: $\mathbf{G}_x = S_x * \mathbf{M}_v$, $\mathbf{G}_y = S_y * \mathbf{M}_v$, where S_x and S_y are Sobel operators. The normal at pixel $p = (x_p, y_p)$ can be calculated as:

$$\mathbf{N}^{(x_p, y_p)} = \mathbf{G}_x(p) \times \mathbf{G}_y(p), \quad (5)$$

where \times denotes cross product. Finally, \mathbf{N} is up-sampled by a factor of 2 and further refined by a U-Net.

4.2. Loss Functions

Our loss functions used to train the network consist of reconstruction errors for the 3D occupancy field and 2D silhouette, as well as the reconstruction loss for normal map refinement. We use extended Binary Cross-Entropy (BCE) loss for the reconstruction of occupancy volume [22]:

$$\begin{aligned} \mathcal{L}_V = -\frac{1}{|\hat{\mathbf{V}}_o|} \sum_{x,y,z} \gamma \hat{\mathbf{V}}_o^{(xyz)} \log \mathbf{V}_o^{(xyz)} + \\ (1 - \gamma) (1 - \hat{\mathbf{V}}_o^{(xyz)}) \log (1 - \mathbf{V}_o^{(xyz)}) \end{aligned} \quad (6)$$

where $\hat{\mathbf{V}}_o$ is the ground-truth occupancy volume corresponding to \mathbf{V}_o , $\mathbf{V}_o^{(xyz)}$ and $\hat{\mathbf{V}}_o^{(xyz)}$ are voxels in the respective volumes at coordinate (x, y, z) , and γ is a weight used to balance the loss contributions of occupied and unoccupied voxels. Similar to [52], we use a multi-view re-projection loss on the silhouette as additional regularization:

$$\begin{aligned} \mathcal{L}_{FS} = -\frac{1}{|\hat{\mathbf{S}}_{fv}|} \sum_{x,y} \mathbf{S}_{fv}^{(xy)} \log \hat{\mathbf{S}}_{fv}^{(xy)} + \\ (1 - \hat{\mathbf{S}}_{fv}^{(xy)}) \log (1 - \mathbf{S}_{fv}^{(xy)}) \end{aligned} \quad (7)$$

where \mathcal{L}_{FS} denotes the front-view silhouette re-projection loss, \mathbf{S}_{fv} is the silhouette re-projection of \mathbf{V}_o , $\hat{\mathbf{S}}_{fv}$ is the corresponding ground-truth silhouette, and $\mathbf{S}_{fv}^{(xy)}$ and $\hat{\mathbf{S}}_{fv}^{(xy)}$ denote their respective pixel values at coordinate (x, y) . Assuming a weak-perspective camera, we can easily obtain $\mathbf{S}_{fv}^{(xy)}$ through orthogonal projection [52]: $\mathbf{S}_{fv}^{(xy)} = \max_z \mathbf{V}_o^{(xyz)}$. The side-view re-projection loss \mathcal{L}_{SS} is defined similarly.

For normal map refinement, we use the cosine distance to measure the difference between predicted normal maps and the corresponding ground truth:

$$\mathcal{L}_N = \frac{1}{|\hat{\mathbf{N}}|} \sum_{x,y} 1 - \frac{\langle \mathbf{N}^{(xy)}, \hat{\mathbf{N}}^{(xy)} \rangle}{|\mathbf{N}^{(xy)}| \cdot |\hat{\mathbf{N}}^{(xy)}|} \quad (8)$$

where $\mathbf{N}^{(xy)}$ is the refined normal map produced by the normal refiner, $\hat{\mathbf{N}}^{(xy)}$ is the ground-truth map, and similarly $\mathbf{N}^{(xy)}$ and $\hat{\mathbf{N}}^{(xy)}$ denote their respective pixel values at coordinate (x, y) .

Therefore, the combined loss is

$$\mathcal{L} = \mathcal{L}_V + \lambda_{FS} \mathcal{L}_{FS} + \lambda_{SS} \mathcal{L}_{SS} + \lambda_N \mathcal{L}_N. \quad (9)$$

5. THuman: 3D Real-world Human Dataset

Collecting rich 3D human surface model with texture containing casual clothing, various human body shapes and natural poses has been a time-consuming and laborious task

as it always relies on either expensive laser scanners or sophisticated multiview systems in a controlled environment. Fortunately, this task becomes easier with the recently introduced DoubleFusion[63], a real-time human performance capture system using a single depth camera. Based on DoubleFusion, we develop a method to capture a 3D human mesh models, and collect a 3D real-world human mesh dataset called “THuman”. THuman has about 7000 data items; each item contains a textured surface mesh, a RGBD image from the Kinect sensor, and an accompanying well-aligned SMPL model. More details about the capture system and the dataset are presented in the supplementary document.

In this work, we only use the textured surface mesh and the accompanied SMPL model to generate training data. The training corpus are synthesized in the following steps: for each model in our dataset, we first render 4 color images from 4 random viewpoints using a method similar to [53]; after that, we generate the corresponding semantic maps and volumes, occupancy volumes as well as normal maps. By enumerating all the models in our dataset, we finally synthesize $\sim 28K$ images for network training.

6. Experiments

6.1. Results

We demonstrate our approach with various human images in Fig.5. The input images are natural images sampled from the LIP dataset[16]. As shown in Fig.5 our approach is able to reconstruct both the 3D human models and surface details like cloth wrinkles, belts and the hem of a dress. In Fig.6 we show an extended application on 3D human performance capture from a single-view RGB video. It should be noted that the reconstruction results are generated by applying our method on each the video frame independently, without any temporal smoothness involved. The results demonstrate the ability of our method to tackle various human poses and its robust performance. Please see the supplemental materials and video for more results.

6.2. Comparison

We compare our method against two state-of-the-art deep learning based approaches for single view 3D human reconstruction: HMR[27] and BodyNet[52]. To eliminate the effect of dataset bias, we fine-tuned the pre-trained model of both network with the same training data as we use to train our network. The **qualitative comparison** are rendered in Fig.5. As shown in the figure, our method is able to achieve much more detailed reconstruction than HMR and BodyNet (See Fig.5(a~f)) and more robust performance than BodyNet when some body parts are occluded (See Fig.5(b,g,h)). The **quantitative comparison** is conducted on the testing set of our synthetic data, and the results are presented in Tab.1. As shown by the numerical results, our method achieves the most accurate reconstruc-

| Method | HMR | BodyNet | Ours |
|-----------------|-------|---------|-------|
| Averaged 3D IOU | 41.4% | 38.7% | 45.7% |

Table 1: Quantitative comparison using 3D IOU score.

| Representation | IOU score (%) |
|------------------------|---------------|
| Joints Heat Map/Volume | 74.16 |
| Semantic Map/Volume | 79.14 |

Table 2: Numerical evaluation of semantic volume/map representation.

tion among all the approaches. BodyNet occasionally produces broken bodies and consequently gets the lowest score. Please see the supplementary document for more details.

6.3. Ablation Study

6.3.1 Semantic Volume/Map Representation

Baseline. An alternative representation to our semantic volume/map is body joint heat volumes/maps that are used in BodyNet[52]. A joint heat map is a multi-channel 2D image where in each channel a Gaussian with fixed variance is centered at the image location of the corresponding joint. By extending the notion of 2D heat maps to 3D, we can also define the heat volumes for body joints. In order to evaluate our semantic volume/map representation, we implement a baseline network that takes body joints’ heat maps and heat volumes as input and has the identical structure to the network presented in Sec.4. In this experiment we generate input semantic volumes/maps and joint heat volumes/maps from the ground-truth SMPL model to eliminate the impact of inaccurate SMPL estimation.

Results. Fig.7 shows the experimental results. We can see that compared to sparse joints, a network taking dense semantic maps/volumes as input is able to learn to reconstruct the 3D model more accurately. In Tab.2, we also test these two methods on the testing portion of our synthetic dataset and measure the reconstruction error using the IoU score of the network output and the ground-truth volume. The numerical results also show that taking dense semantic maps/volumes as input helps the network achieve higher reconstruction accuracy. We think that it is because our semantic volume/map representation encodes information about the body shape and pose jointly and provides a good initialization for the volumetric reconstruction network.

6.3.2 Multi-scale Volumetric Feature Transformation

Baseline. To evaluate our multi-scale VFT component, we implement 3 baseline networks: Baseline (A) only performs VFT at the finest scale, while Baseline (B) at the coarsest scale; different from the original network and Baseline (A)(B), Baseline (C) first encodes input images/volumes into latent codes, concatenates the latent code of the image with that of the volume and then feeds the concatenation into the volume decoder.

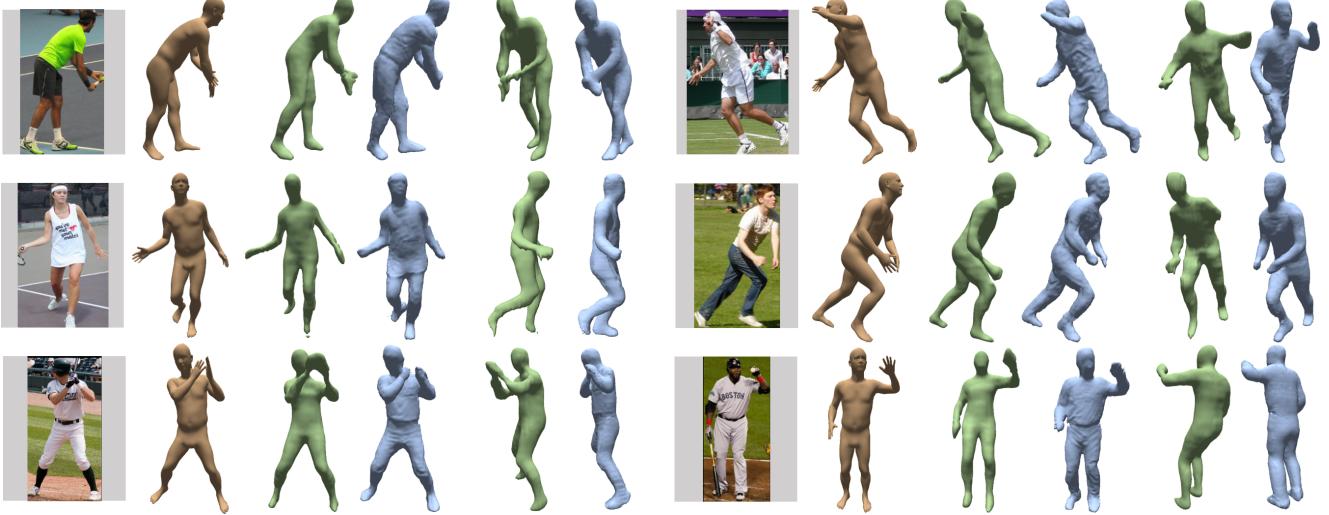


Figure 5: Reconstruction results on natural images. In each panel, the input images are presented in the left column, while last five columns show the results of HMR[27] (in orange), BodyNet[52] (in green) and our method (in blue). For BodyNet and our method we render the results from two views, i.e., the input camera view and a side view.



Figure 6: 3D reconstruction from monocular videos using our method. The reconstruction results are generated by applying our method on each individual video frame independently. The last three video clips come from the dataset of MonoPerfCap[59].

Results. Fig.8 shows the reconstruction loss for different fusing methods. Here we found that by using multi-scale VFT, the network outperforms the baseline method in terms of the reconstruction of the model boundaries (see the second plot in Fig.8). The same conclusion can be drawn from

the visual comparison shown in Fig.9. Using coarsest VFT (Baseline (B)) or latent code concatenation (Baseline (C)) results into over-smooth reconstruction of the girl’s head due to the lack of higher-scale information (see the last two results in Fig.8). The result generated by Baseline (A) is

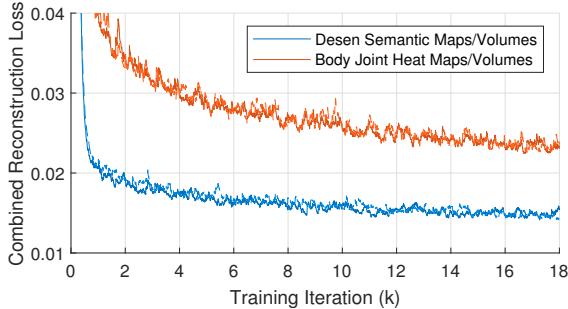


Figure 7: Evaluation of semantic volume/map representation. We evaluate two different inputs for the image-guided vol2vol network, and show the combined reconstruction losses ($\mathcal{L}_V + \lambda_{FS}\mathcal{L}_{FS} + \lambda_{SS}\mathcal{L}_{SS}$). Solid lines show training error and dashed lines show validation error (they almost overlap with each other).

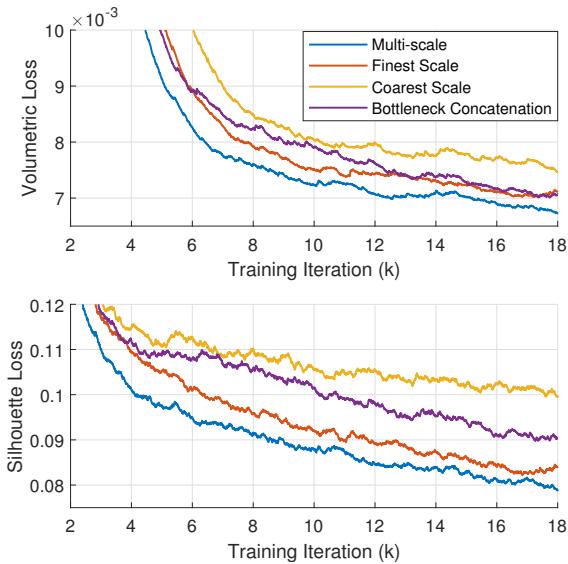


Figure 8: Evaluation of multi-scale volumetric feature transformation (VFT). We evaluate several ways to fuse 2D features into 3D volumes, and show the volumetric loss (\mathcal{L}_V) and the silhouette loss ($\mathcal{L}_{FS} + \mathcal{L}_{SS}$) in the figure. For clarity we do not show the validation loss.

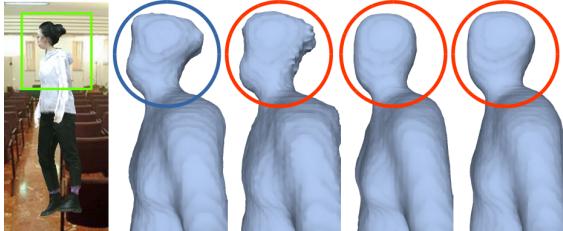


Figure 9: Visual evaluation of multi-scale VFT. From left to right: input image, head reconstruction result by our method, baseline(A), baseline(B) and baseline(C).

much more accurate but contain noises. With the proposed multi-scale VFT component, our network is able to reconstruct the hair bun of the girl (the blue circle in Fig.8).

| Error Metric | Cosine Distance | ℓ_2 -norm |
|--------------------|-----------------|----------------|
| Without Refinement | 0.0941 | 0.336 |
| With Refinement | 0.0583 | 0.262 |

Table 3: Numerical normal errors with/without normal refinement.

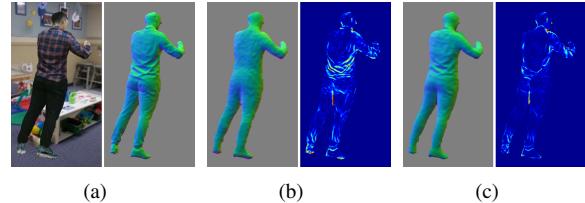


Figure 10: Qualitative evaluation of normal refinement. (a) Reference image and the ground-truth normal. (b) Surface normal and error map without normal refinement. (c) Refined normal and the corresponding error map.

6.3.3 Normal Refinement

Baseline. To evaluate our normal refinement module, we implement a baseline network by removing the volume-to-normal projection layer and the normal refinement U-Net as well from the original network.

Results. The evaluation experiment is conducted using our synthetic dataset and the results are shown in Tab.3 and Fig.10. In Tab.3 we present the prediction error of surface normal with and without normal refinement. This numeric comparison shows that the normal refinement network properly refines the surface normal based on the input image. We can also observe that surface details are enhanced and enriched after normal refinement in Fig.10.

7. Discussion

Limitations. Our method relies on HMR and SMPLify to estimate a SMPL model for the input image. As a result, we cannot give an accurate reconstruction if the SMPL estimation is erroneous. Additionally, the reconstruction of invisible areas is over-smoothed; using a generative adversarial network may force the network to learn to add realistic details to these areas. Due to the limited resolution of the depth maps, DoubleFusion is unable to reconstruct hand geometry and thus all hands are clenched in the THuman dataset. Consequently, our method also fails to recover fine-scale details such as facial expression and hands' shape. This issue can be addressed using methods that focus on face/hand reconstruction.

Conclusion. In this paper, we have presented a deep-learning based framework to reconstruct a 3D human model from a single image. Based on the three-stage task decomposition, the dense semantic representation, the proposed network design and the 3D real-world human dataset, our method is able to estimate a plausible geometry of the target in the input image. We believe both our dataset and network will enable convenient VR/AR content creation and inspire many further researches on 3D vision for humans.

References

- [1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2019. [1](#), [2](#)
- [2] Thiemo Alldieck, Marcus A. Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *3DV*, 2018. [1](#), [2](#)
- [3] Thiemo Alldieck, Marcus A. Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE CVPR*, 2018. [1](#), [2](#)
- [4] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416, 2005. [2](#)
- [5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, pages 561–578, 2016. [1](#), [2](#), [3](#)
- [6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE CVPR*, pages 1302–1310, 2017. [1](#), [2](#)
- [7] Young-Woon Cha, True Price, Zhen Wei, Xinran Lu, Nicholas Rewkowski, Rohan Chabra, Zihe Qin, Hyounghun Kim, Zhaoqi Su, Yebin Liu, Adrian Ilie, Andrei State, Zhenlin Xu, Jan-Michael Frahm, and Henry Fuchs. Towards fully mobile 3d face, body, and environment capture using only head-worn cameras. *IEEE Transactions on Visualization and Computer Graphics*, 24(11):2993–3004, 2018. [1](#)
- [8] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. Graph.*, 34(4):69, 2015. [1](#), [2](#)
- [9] Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. *ACM Trans. Graph.*, 27(3):98:1–98:10, 2008. [2](#)
- [10] Mingsong Dou, Philip L. Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. Motion2fusion: real-time volumetric performance capture. *ACM Trans. Graph.*, 36(6):246:1–246:16, 2017. [2](#)
- [11] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escalano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: real-time performance capture of challenging scenes. *ACM Trans. Graph.*, 35(4):114, 2016. [2](#)
- [12] Sean Ryan Fanello, Julien P. C. Valentin, Christoph Rhemann, Adarsh Kowdle, Vladimir Tankovich, Philip L. Davidson, and Shahram Izadi. Ultrastereo: Efficient learning-based matching for active stereo systems. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, pages 6535–6544, 2017. [2](#)
- [13] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE T-PAMI*, 32(8):1362–1376, 2010. [2](#)
- [14] Juergen Gall, Carsten Stoll, Edilson de Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *IEEE CVPR*, pages 1746–1753, 2009. [2](#)
- [15] Andrew Gilbert, Marco Volino, John Collomosse, and Adrian Hilton. Volumetric performance capture from minimal camera viewpoints. In *ECCV*, pages 591–607, 2018. [2](#)
- [16] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *IEEE CVPR*, July 2017. [6](#)
- [17] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *IEEE CVPR*, 2018. [2](#)
- [18] Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. Reticam: Real-time human performance capture from monocular video. *CoRR*, abs/1810.02648, 2018. [2](#)
- [19] Peng Huang, Margara Tejera, John P. Collomosse, and Adrian Hilton. Hybrid skeletal-surface motion graphs for character animation from 4d performance capture. *ACM Trans. Graph.*, 34(2):17:1–17:14, 2015. [1](#)
- [20] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. Deep volumetric video from very sparse multi-view performance capture. In *ECCV*, pages 351–369, 2018. [1](#), [2](#)
- [21] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natura environments. *IEEE T-PAMI*, 36(7):1325–1339, 2014. [3](#)
- [22] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. *IEEE ICCV*, 2017. [5](#)
- [23] Aaron S. Jackson, Chris Manafas, and Georgios Tzimiropoulos. 3d human body reconstruction from a single image via volumetric regression. *CoRR*, abs/1809.03770, 2018. [2](#)
- [24] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *IEEE ICCV*, pages 3334–3342, 2015. [2](#)
- [25] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *IEEE CVPR*, 2017. [1](#), [2](#)
- [26] Takeo Kanade, Peter Rander, and P. J. Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE MultiMedia*, 4(1):34–47, 1997. [2](#)
- [27] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE CVPR*, 2018. [1](#), [2](#), [3](#), [6](#), [7](#)
- [28] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE CVPR*, pages 4704–4713, 2017. [2](#), [3](#)
- [29] Hao Li, Etienne Vouga, Anton Gudym, Linjie Luo, Jonathan T. Barron, and Gleb Gusev. 3d self-portraits. *ACM Trans. Graph.*, 32(6):187:1–187:9, 2013. [2](#)
- [30] Yebin Liu, Qionghai Dai, and Wenli Xu. A point-cloud-based multiview stereo algorithm for free-viewpoint video.

- IEEE Transactions on Visualization and Computer Graphics*, 16(3):407–418, 2010. 2
- [31] Yebin Liu, Juergen Gall, Carsten Stoll, Qionghai Dai, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. *IEEE T-PAMI*, 35(11):2720–2735, 2013. 2
- [32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1, 2
- [33] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *IEEE ICCV*, pages 2659–2668, 2017. 1, 2
- [34] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved CNN supervision. In *3DV*, pages 506–516, 2017. 3
- [35] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: real-time 3d human pose estimation with a single RGB camera. *ACM Trans. Graph.*, 36(4):44:1–44:14, 2017. 2
- [36] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. *CoRR*, abs/1901.00049, 2019. 2
- [37] Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. Efficiently combining positions and normals for precise 3d geometry. *ACM Trans. Graph.*, 24(3):536–543, July 2005. 3
- [38] Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE CVPR*, 2015. 2
- [39] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hour-glass networks for human pose estimation. In *ECCV*, pages 483–499, 2016. 1, 2
- [40] Sergio Orts-Escolano, Christoph Rhemann, Sean Ryan Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles T. Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien P. C. Valentim, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchny, Cem Keskin, and Shahram Izadi. Holoporation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 741–754, 2016. 1
- [41] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V. Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *IEEE CVPR*, pages 4929–4937, 2016. 1
- [42] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Trans. Graph (Proc. SIGGRAPH)*, 36(4), 2017. 1
- [43] Fabian Prada, Misha Kazhdan, Ming Chuang, Alvaro Collet, and Hugues Hoppe. Motion graphs for unstructured textured meshes. *ACM Trans. Graph.*, 35(4):108:1–108:14, 2016. 1
- [44] Gr  gory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *IEEE CVPR*, pages 1216–1224, 2017. 1, 2
- [45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 3
- [46] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1-2):4–27, 2010. 3
- [47] Jonathan Starck and Adrian Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3):21–31, 2007. 2
- [48] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *IEEE ICCV*, pages 2621–2630, 2017. 2
- [49] Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. In *BMVC*, 2017. 2
- [50] Denis Tom  , Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *IEEE CVPR*, pages 5689–5698, 2017. 2
- [51] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *NIPS*, pages 5242–5252, 2017. 2
- [52] G  l Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018. 1, 2, 3, 5, 6, 7
- [53] G  l Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *IEEE CVPR*, pages 4627–4635, 2017. 2, 3, 6
- [54] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popovic. Articulated mesh animation from multi-view silhouettes. *ACM Trans. Graph.*, 27(3):97:1–97:9, 2008. 2
- [55] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul E. Debevec, Jovan Popovic, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. *ACM Trans. Graph.*, 28(5):174:1–174:11, 2009. 2
- [56] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *IEEE CVPR*, June 2018. 3, 4
- [57] Michael Waschbusch, Stephan W  rmlin, Daniel Cotting, Filip Sadlo, and Markus H. Gross. Scalable 3d video of dynamic scenes. *The Visual Computer*, 21(8-10):629–638, 2005. 2
- [58] Chenglei Wu, Kiran Varanasi, Yebin Liu, Hans-Peter Seidel, and Christian Theobalt. Shading-based dynamic shape refinement from multi-view video under general illumination. In *IEEE ICCV*, pages 1108–1115, 2011. 2
- [59] Weipeng Xu, Avishek Chatterjee, Michael Zollh  fer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Trans. Graph.*, 37(2):27:1–27:15, 2018. 2, 7

- [60] B. Yang, S. Rosa, A. Markham, N. Trigoni, and H. Wen. Dense 3d object reconstruction from a single depth view. *IEEE T-PAMI*, 2018. 3
- [61] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy S. J. Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. *CoRR*, abs/1803.09722, 2018. 2
- [62] Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jian-hui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. Body-fusion: Real-time capture of human motion and surface geometry using a single depth camera. In *IEEE ICCV*, October 2017. 2
- [63] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Double-fusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *IEEE CVPR*, June 2018. 2, 6
- [64] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *IEEE ICCV*, pages 398–407, 2017. 1, 2
- [65] Hao Zhu, Yebin Liu, Jingtao Fan, Qionghai Dai, and Xun Cao. Video-based outdoor human reconstruction. *IEEE Trans. Circuits Syst. Video Techn.*, 27(4):760–770, 2017. 2