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, a deep learning based framework for 3D human reconstruction from a single RGB image. Since this problem is highly intractable, we adopt a stage-wise, coarse-to-fine method consisting of three steps, namely body shape/pose estimation, surface reconstruction and frontal surface detail refinement. Once a body is estimated from the given image, our method generates a dense semantic representation from it. The representations not only encodes body shape and pose but also bridges the 2D image plane and 3D space. An image-guided volume-to-volume translation CNN is introduced to reconstruct the surface given the input image and the dense semantic representation. 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 details of the subject's surface geometry. The details on the frontal areas of the surface are further refined through a normal map refinement network. The normal refinement network 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 approximately 7000 models. The whole 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 reconstruction 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 [6], image and video editing and re-enactment [21, 52], holoportation [48] and virtual dressing [51]. To perform full-body 3D reconstruction, currently available methods require the fusion of multiview images [7, 29, 22] or multiple temporal images [2] of the target. Recovering a human model from a single RGB image remains a challenging task not only because of the requirement to predict the shape of invisible parts but also due to the need for the depth estimation for visible surface. Using only a single image, available human parsing studies have covered popular topics starting from 2D pose detection [50, 5, 47], advancing to 3D pose detection [41, 53, 78], and finally expanding to body shape capture [31] using a human statistic model like SMPL [39]. However, the statistic template models can capture only the shape and pose of a minimally clothed body but lack the ability to represent a human body under a normal clothing layer. Although the most recent BodyNet[64] has pioneered research towards this goal, it only generates SMPL-like 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 be a next hot topic in the coming years.

Figure 1. Example results reconstructed by our method.

In this paper, we propose DeepHuman, a deep learning based framework aiming to address the aforementioned challenges. Specifically, we achieve 3D human reconstruction in a coarse-to-fine manner by decomposing this task into three subtasks: a) body estimation from the input image, b) surface reconstruction from the image and the estimated body, and c) frontal surface detail refinement according to the image. The goal of body estimation is to infer the shape and pose parameters of a parametric body template (e.g., SMPL[39]) from the input image. The body template, SMPL[39], is a watertight 3D mesh model that provides a good approximation for the human surface geometry with only a small amount of free parameters (10 for body shape and 72 for body pose). Therefore, it can be taken as the

prior knowledge about the subject in the input image in the second step, where we reconstruct the surface in the form of 3D occupancy grid. In the final step we refine geometrical details of the frontal surface that occupancy voxels fail to recover.

Since several accurate methods are available for SMPL estimation from a single image[4, 31], we focus on the latter two subtasks. For surface reconstruction, our method leverages the parametric body model by generating a 3D semantic volume and a corresponding 2D semantic map, which serve as a dense representation encoding shape and pose information, bridge the 2D image plane and 3D space and are suitable as the input of neural network. Further, to capture geometrical details like the hairstyle or cloth contours to the maximum possible extent, we propose a multi-scale volumetric feature transformation so that different scales of image guidance information can be fused into the 3D volumes. Finally, we introduce a volumetric normal projection layer in order to further refine and enrich surface geometrical details in 2D image space according to the input image. This layer is designed to concatenate the volume generation network with the normal refinement network and enables end-to-end training/fine-tuning.

The available 3D human dataset [65] used for the network training in BodyNet [64] is essentially a set of synthesized images textured over SMPL models [39]. 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 [76] technique for real-time human model 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 approximately 7000 human models with about 230 kinds of clothes under randomly sampled poses.

Our proposed network learns from the training corpus synthesized from our THuman dataset. Benefiting from the model diversity of our dataset, our 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 [30, 58, 36]. Shape cues like silhouette [42, 38], stereo and shading cues have been integrated in both passive [58, 36, 71] and active illumination [69, 67] modes to improve the reconstruction performance. State-of-the-art real-time [10, 9] and extremely high-quality [7] reconstruction results have also been demonstrated with tens or even hundreds of cameras using binocular [11] or multi-view stereo matching [12] algorithms. To capture detailed motions of multiple interacting characters, more than six hundred cameras have been used to overcome the occlusion challenges [28, 29]. However, all these multi-camera 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 recently been investigated by using CNNs for learning silhouette cues [14] and stereo cues [22]. 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 (usually with about eight cameras) [66, 8, 13] have been developed for dynamic human model reconstruction using skeleton tracking [66, 37] or human mesh-based template deformation [8], 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. Since only a single-view camera is needed, methods in this category require the aggregation of multiple temporal frames for full-body model generation.

For RGBD images, DynamicFusion [46] breaks the static scene assumption and deforms the non-rigid target for TSDF fusion on a canonical static model. Many of the following approaches have tried to improve the robustness by adding color features [23], shading constraints [19] and articulated prior [75] and dealing with topology changes [56, 57]. The recently appeared DoubleFusion [76] method 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 [35]. 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 [79], or keeping the camera static while the target rotates [2, 1]. Recently, human performance capture that can reconstruct dynamic human models using only a single RGB camera has been proposed [72] and sped up to be run in real-time [20], however, similar to the multicamera scenario [66, 8, 13], such approaches require a pre-scanned human model obtained using a laser scanner or temporal images-based like [79, 2, 1].

Human Parsing from a Single Image. Parsing human from a single image has recently been a popu-

lar topic in computer vision. The research can be categorized into sparse 2D parsing (2D skeleton estimation) [70, 50, 5, 47], sparse 3D parsing (3D skeleton estimation) [41, 49, 53, 78, 59, 62, 44, 74], dense 2D parsing [18, 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 [3] and SMPL [39]. For example, by fitting the SCAPE or SMPL model to the detected 2D skeleton and other shape cues of an image [4, 34], or by regressing [31, 60, 63] 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 two recent works by Varol et al. [64] and Jackson et al. [26]. In the former study, the 3D human datasets used for the training process are essentially synthesized human imagery textured over SMPL models (lacking geometry details), leading to SMPL-like voxel geometries in their outputs. The latter 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.

3D Human Body Datasets. Most of the available 3D human datasets are used for 3D pose and skeleton detection. Both the HumanEva [55] and Human3.6M [24] datasets contain multiview human video sequences with ground-truth 3D skeleton motion obtained from a marker-based motion capture system. Because of the need to wear markers or special suits, both of these datasets have limited apparel divergence. MPI-INF-3DHP [43] dataset enriches the cloth appearance by using a multiview markerless mocap system. However, all the above datasets lack a 3D model of each temporal frame. Recently, with the emergence of the requirement of pose and shape reconstruction from a single image, the synthesized SURREAL [65] 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 [34] 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 input image of a person in arbitrary clothes, denoted by $\mathbf{I}$, our method aims to reconstruct his/her full-body 3D surface with plausible geometrical details. Directly reconstructing a 3D surface model of the subject from the given image is very challenging because of depth ambiguities, body self-occlusions and high degree of freedom of the output space. For this reason we perform 3D human reconstruction in a coarse-to-fine manner. Our method starts with body estimation, then steps into surface reconstruction and finally recovers the details on the frontal areas of the surface.

To estimate a body from $\mathbf{I}$, we exploit the state-of-the-art methods, HMR[31] and Simplify[4], both of which are able to infer the shape and pose parameters of SMPL[39] from a single image. We found that they have complementary characteristics: HMR's prediction are always plausible but not well-aligned with the color image, while Simplify aligns the SMPL model with detected keypoint very well but relies on initialization to output plausible results. To obtain an accurate SMPL estimation for the image, we first use HMR to obtain an initial estimation and then improve its accuracy using Simplify.

The estimated body shape and pose parameters determine a polygon mesh representation of the body through linear shape blending and pose skinning[39]. However, it is hard to feed the polygon mesh representation into a deep neural network. Therefore, inspired by "Vitruvian Manifold" [61], we introduce a dense semantic representation generated from SMPL. Specifically, we pre-define a *semantic code* (a 3-dimensional vector) for each vertex on SMPL according to its spatial coordinate at rest pose. Given a SMPL model corresponding to a human image, 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 voxel grid and then propagating the semantic codes into the occupied voxels. Our dense semantic representation has three advantages: it encodes information about both the shape and the pose of the body, provides clues about the corresponding relationship between 3D voxels and 2D image pixels, and is easy to be incorporated into neural networks.

For the surface geometry reconstruction, we adopt an occupancy volume to represent the surface[64]. Specifically, we define a 3D occupancy voxel grid $\mathbf{V}_o$, where the values of voxels inside the surface are set to 1 and others are set to 0. All occupancy volumes have a fixed resolution of $128 \times 192 \times 128$, where the resolution of the y-axis is set to a greater value because we observed that a 3D human model usually has a major axis. 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 multi-scale 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 or recover fine details like wrinkles on the clothes. To further enrich 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 frontal surface using 2D normal maps, and

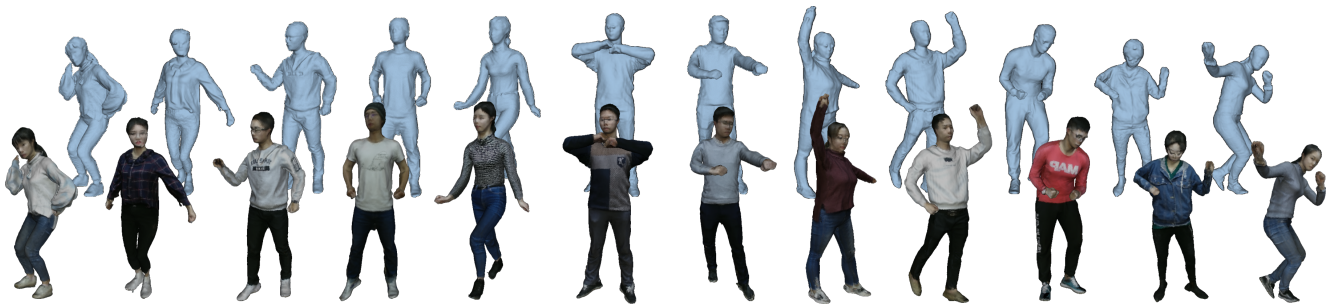


Figure 2. Example meshes sampled from our dataset.

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 training corpus from the dataset. Once the network is trained, it can be used to predict an occupancy volume and a frontal normal map given an RGB image of a person. We obtain the final polygon mesh model for the subject by firstly extracting a triangular polygon mesh from the occupancy volume using Marching Cube algorithm and then refining the mesh according to the frontal normal map using the method in [45].

4. Approach

4.1. Network Architecture

Our network consists of 3 components, namely an image feature encoder $\mathcal{G}$, a volume-to-volume (vol2vol) translation network $\mathcal{H}$ and a normal refinement network $\mathcal{R}$, as shown in Fig.3. The image feature encoder $\mathcal{G}$ aims to extract multi-scale 2D feature maps $\mathbf{M}_f^{(k)}(k = 1, \ldots, K)$ from the combination of $\mathbf{I}$ and $\mathbf{M}_s$. The vol2vol network is a volumetric U-Net [73], which takes $\mathbf{V}_s$ and $\mathbf{M}_f^{(k)}(k = 1, \ldots, K)$ as input, and outputs an occupancy volume $\mathbf{V}_o$ representing the surface. Our vol2vol network $\mathcal{H}$ fuses multi-scale semantic features $\mathbf{M}_f^{(k)}(k = 1, \ldots, K)$ into its encoder through a *multi-scale volumetric feature transformer*. After generating $\mathbf{V}_o$, a normal refinement U-Net [54] $\mathcal{R}$ 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.

4.1.1 Multi-scale Volumetric Feature Transformer

In this work, we extend the Spatial Feature Transformer (SFT) layer [68] to handle 2D-3D data pairs in the multi-scale feature pyramid, and propose multi-scale Volumetric Feature Transformer (VFT). SFT was first used in [68] 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 $\mathbf{F}$ is carried out as:

$$\mathcal{SFT}(\mathbf{F}) = \alpha \odot \mathbf{F} + \beta \tag{1}$$

where $\odot$ is Hadamard product.

In our network, at each level k, a feature volume $\mathbf{V}_f^{(k)}$ (blue cubes in Fig.3) and a feature map $\mathbf{M}_f^{(k)}$ (orange squares in Fig.3) are provided by previous encoding layers. Similar to [68], we first map the feature map $\mathbf{M}_f^{(k)}$ to modulation parameters (α_k, β_k) through convolution+activation layers (see the second row of Fig.4). Note that the operation in Eqn.(1) cannot be applied directly on $\mathbf{V}_f^{(k)}$ and $\mathbf{M}_f^{(k)}$ because of dimension inconsistency ($\mathbf{V}_f^{(k)}$ 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:

$$\mathcal{VFT}\left(\mathbf{V}_f^{(k)}(z_i)\right) = \alpha_k \odot \mathbf{V}_f^{(k)}(z_i) + \beta_k \tag{2}$$

where $\mathbf{V}_f^{(k)}(z_i)$ is the feature slice on plane $z = z_i, z_i = 1, 2, \ldots, Z$ and Z is the maximal z-axis coordinate. The output of a VFT layer is the re-combination of transformed feature slices. The operations applied by VFT layers are illustrated in Fig.4.

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

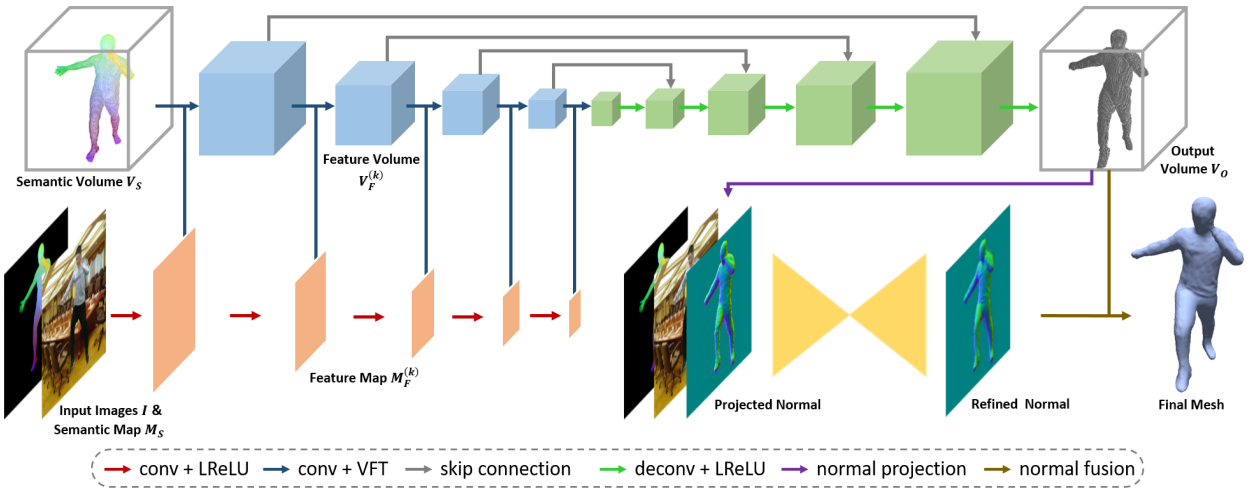


Figure 3. Network architecture. Our network is mainly composed of an image feature encoder (orange), a volume-to-volume translation network (blue & green) and a normal refinement network (yellow).

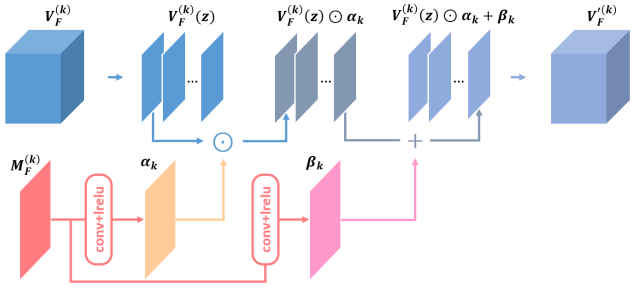


Figure 4. Illustration of volumetric feature transformation at level k.

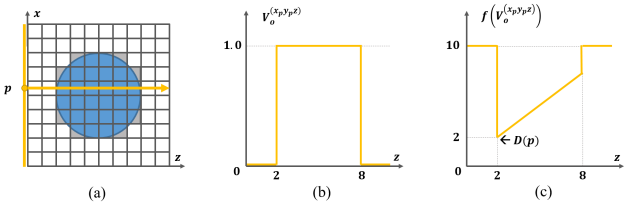


Figure 5. Illustration of differentiable depth projection.

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.3 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 preserve geometry details compared to directly concatenating latent variables at the network bottleneck.

4.1.2 Volume-to-normal Projection Layer

Our goal is to recover geometrical details (e.g. wrinkles and cloth boundary) on the visible surface of the human model. However, a volume-based representation is unable to recover such fine-grain details due to resolution limitations. Thus, we encode the frontal 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.5 is a 2D illustration explaining how the layer projects depth maps. In Fig.5(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 $\mathbf{D}(p)$ 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.5(b)). Afterwards, we can determine $\mathbf{D}(p)$ by finding the nearest occupied voxel. Formally, $\mathbf{D}(p)$ is obtained according to

$$\mathbf{D}(p) = \inf\left\{z | \mathbf{V}_o^{(x_p y_p z)} = 1\right\} \tag{3}$$

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:

$$\mathbf{V}_d^{(xyz)} = f(\mathbf{V}_o^{(xyz)}) = M(1 - \mathbf{V}_o^{(xyz)}) + z\mathbf{V}_o^{(xyz)} \tag{4}$$

where M is a sufficiently large constant. Then $\mathbf{D}(p)$ can be computed as:

$$\mathbf{D}(p) = \min_z f(\mathbf{V}_d^{(x_p y_p z)}), \tag{5}$$

as illustrated in Fig.5(c).

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:

$$\mathbf{N}^{(x_p y_p)} = \mathbf{G}_x(p) \times \mathbf{G}_y(p), \tag{6}$$

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 parameters 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 [25]:

$$\mathcal{L}_V = -\frac{1}{\left|\hat{\mathbf{V}}_o\right|} \sum_{x,y,z} \gamma \hat{\mathbf{V}}_o^{(xyz)} \log \mathbf{V}_o^{(xyz)} + (1-\gamma)\left(1 - \hat{\mathbf{V}}_o^{(xyz)}\right) \log\left(1 - \mathbf{V}_o^{(xyz)}\right) \tag{7}$$

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 [64], we use a multi-view re-projection loss on the silhouette as additional regularization:

$$\mathcal{L}_{FS} = -\frac{1}{\left|\hat{\mathbf{S}}_{fv}\right|} \sum_{x,y} \mathbf{S}_{fv}^{(xy)} \log \mathbf{S}_{fv}^{(xy)} + \left(1 - \hat{\mathbf{S}}_{fv}^{(xy)}\right) \log\left(1 - \mathbf{S}_{fv}^{(xy)}\right) \tag{8}$$

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 [64]: $\mathbf{S}_{fv}^{(xy)} = \max_z \mathbf{V}_o^{(xyz)}$. The side-view re-projection loss $\mathcal{L}_{SS}$ is defined similarly.

Table 1. Network Architecture Details.

Net	Layer	Kernel	Stride	Output
$\mathcal{G}$	conv+lrelu	4	2	$96 \times 64 \times 8$
	conv+lrelu	4	2	$48 \times 32 \times 16$
	conv+lrelu	4	2	$24 \times 16 \times 32$
	conv+lrelu	4	2	$12 \times 8 \times 64$
	conv+lrelu	4	2	$6 \times 4 \times 128$
$\mathcal{H}$	conv+lrelu	4	2	$64 \times 96 \times 64 \times 8$
	conv+lrelu	4	2	$32 \times 48 \times 32 \times 16$
	conv+lrelu	4	2	$16 \times 24 \times 16 \times 32$
	conv+lrelu	4	2	$8 \times 12 \times 8 \times 64$
	conv+lrelu	4	2	$4 \times 6 \times 4 \times 128$
	transconv+lrelu	4	2	$8 \times 12 \times 8 \times 64$
	transconv+lrelu	4	2	$16 \times 24 \times 16 \times 32$
	transconv+lrelu	4	2	$32 \times 48 \times 32 \times 16$
	transconv+lrelu	4	2	$64 \times 96 \times 64 \times 8$
	transconv+lrelu	4	2	$128 \times 192 \times 128 \times 4$
	conv+sigmoid	3	1	$128 \times 192 \times 128 \times 1$
$\mathcal{R}$	conv+lrelu	4	2	$192 \times 128 \times 16$
	conv+lrelu	4	2	$96 \times 64 \times 32$
	conv+lrelu	4	2	$48 \times 32 \times 32$
	conv+lrelu	4	2	$24 \times 16 \times 32$
	conv+lrelu	4	2	$12 \times 8 \times 32$
	transconv+lrelu	4	2	$24 \times 16 \times 32$
	transconv+lrelu	4	2	$48 \times 32 \times 32$
	transconv+lrelu	4	2	$96 \times 64 \times 32$
	transconv+lrelu	4	2	$192 \times 128 \times 16$
	transconv+lrelu	4	2	$384 \times 256 \times 8$
	conv+tanh	3	1	$384 \times 256 \times 3$

* The term "conv" is convolution for short, "transconv" is transposed convolution and "lrelu" is Leakly ReLU.

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

$$\mathcal{L}_N = \frac{1}{\left|\hat{\mathbf{N}}\right|} \sum_{x,y} 1 - \frac{< \mathbf{N}^{(xy)}, \hat{\mathbf{N}}^{(xy)} >}{|\mathbf{N}^{(xy)}| \cdot |\hat{\mathbf{N}}^{(xy)}|} \tag{9}$$

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

$$\mathcal{L} = \mathcal{L}_V + \lambda_{FS}\mathcal{L}_{FS} + \lambda_{SS}\mathcal{L}_{SS} + \lambda_N \mathcal{L}_N, \tag{10}$$

where $\lambda_\bullet$s are scalar weights of the loss terms.

4.3. Implementation details

The volume-to-volume network $\mathcal{H}$ takes as input a semantic volume with $128 \times 192 \times 128$ resolution, and outputs an occupancy with the same shape. The image encoder $\mathcal{G}$ concatenates as input the given RGB image and the corresponding semantic map, both of which have a resolution of 192×128. Our normal refinement U-Net $\mathcal{R}$ takes as input the concatenation of the RGB image, semantic map and up-sampled normal projection result, and the input/output resolution of $\mathcal{R}$ is 384×256. The architecture details are shown in Tab.1.

During network training, the parameters are set to $\lambda_{FS} = \lambda_{SS} = 0.1, \lambda_N = 0.01, \gamma = 0.7$. We exploit a two-stage training procedure: first pre-train the vol2vol network and the normal refinement network, and then fine-tune them jointly with the combined loss. We used Adam [33] with default parameters as the optimizer. The learning rate is fixed to 2e-4 during the whole training procedure, and the batch size is set to be 4.

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 real-time human performance capture system using only a single depth camera [76, 77]. In this section we present the method to capture a 3D real-world human dataset called "THuman dataset" for later supervised deep learning on single-image human reconstruction.

Our capture system is based on the single-view RGB-D DoubleFusion [76] technique. DoubleFusion utilizes a double-layer representation and incorporates a motion prior derived from the SMPL [39]. It simultanuously solves skeleton motions and non-rigid deformation according to the depth observation at the current frame. After getting the motion field, depth pixels in the current frame are fused into a reference volume as described in [46]. As the observed surface is gradually fused and deformed, the shape and pose parameters of the body layer are also gradually optimized through volumetric shape-pose optimization. In this way the two layers can benefit from each other, leading to robust tracking and accurate reconstruction.

The available DoubleFusion technique performs only robust fusion of detailed surface geometries. To obtain full-body texture, we can directly perform color or albedo fusion in a similar way to depth fusion. However, the fused texture blurs when fast body motion occurs. Thus we develop a two-stage capture procedure. In the first stage, the subject actors are required to rotate slowly and perform some basic surface completion motions to obtain a surface geometry that is as complete as possible and clear texture recovery of the surface as well. After that, in the second stage, we disable geometry fusion and texture update, but still perform the non-rigid surface registration based on the input depth information. In this way, we still capture non-rigid motion details of the subject's surface.

In order to obtain human mesh data under natural but diverse poses, our system presents to the subject a reference pose randomly sampled from MOSH[40] dataset every 6 seconds and requires the performer to imitate the reference pose in the second stage. Note that the 6-second interval is usually long enough for subjects to recognize the presented pose and prepare for imitation. At the end of every 6-second interval, the system automatically saves the RGBD image, the 3D surface mesh and its corresponding SMPL model in the current live pose. After data capture, we post-process the raw meshes through hole filling [32], remeshing [27] and isolated artifact removal.

After approximately 70 hours of data capture using only one capture setup, we achieve capturing and reconstruction of 230 subject characters, with each character corresponding to about 30 poses. This data leads to 7000 data items in our THuman dataset; some examples are shown in Fig. 2. Each item contains a textured surface mesh, a RGBD image from the Kinect sensor, and an accompanying well-aligned SMPL model.

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 [65]; 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 a range of photos of people in Fig.6. In this figure, image (A) and (B) are testing images synthesized from our THUman dataset, while image (C)-(H) are natural images sampled from the LIP dataset[15]. Note that the subjects in image (A) and (B) do not appear in our training data. As shown in Fig.6 our approach is able to reconstruct both the 3D human models and surface details like cloth wrinkles (see Fig.6(a,b,c)). From Fig.6(d,e,f) we can also see that our method can not only reconstruct the 3D meshes according to input images and detected SMPL models, but also recover discontinuities on the surface like a belt and the hem of a dress. In Fig.7 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 video for more details.

6.2. Comparison

6.2.1 Competing Approaches

We compare our method against two two state-of-the-art deep learning based approaches for single view 3D human

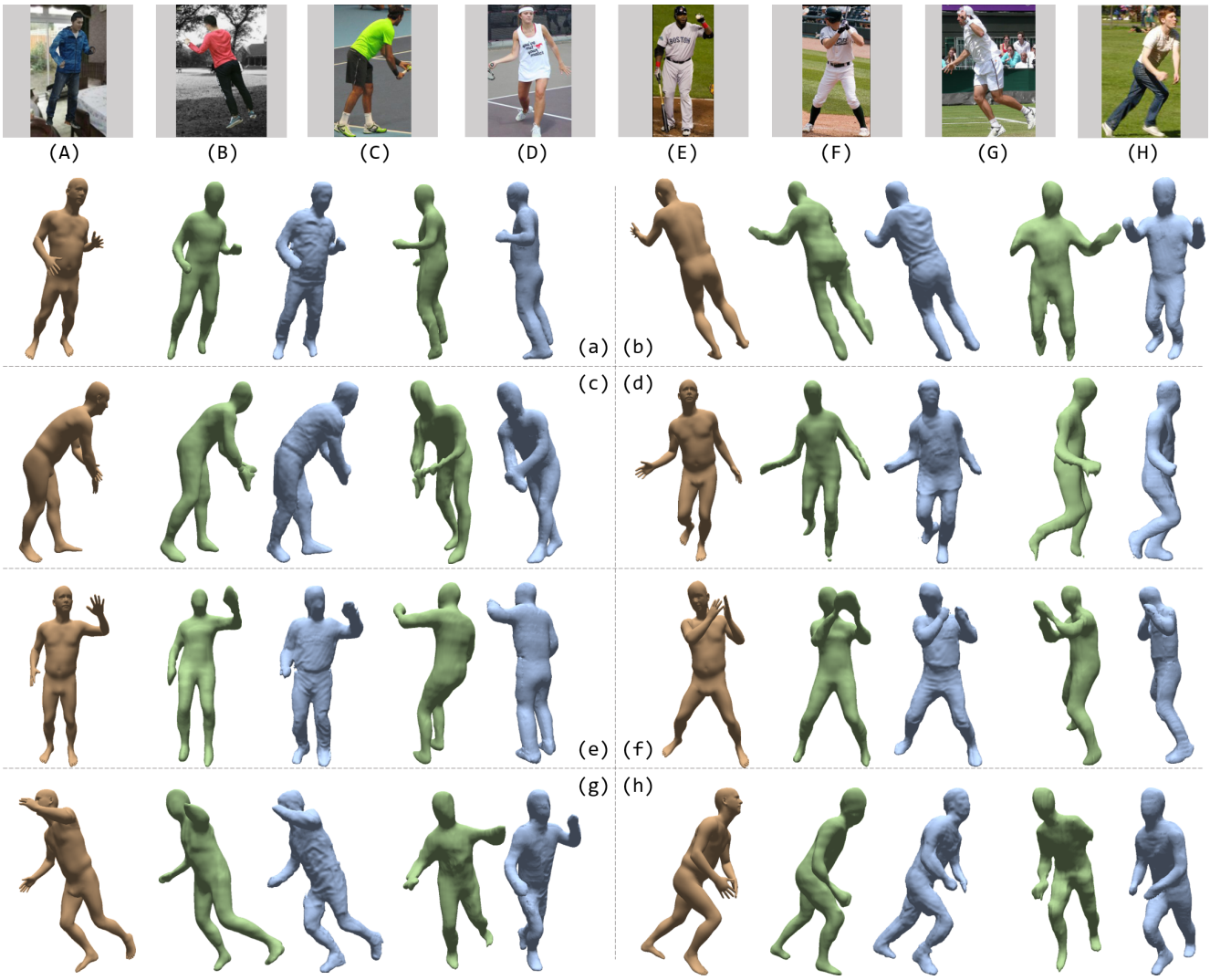


Figure 6. Reconstruction results on synthesis images and natural images. The input images are presented in the first row, while last four rows show the results of HMR[31] (in orange), BodyNet[64] (in green) and our method (in blue). For BodyNet and our method we render the results from two views, i.e., a front view and a side view.

reconstruction. 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.

(1) **HMR**. In [31], Kanazawa et al. proposed a neural network to directly regress the shape and pose parameter of SMPL from an RGB image. The output of HMR is a 75-D vector, which can be used to generate a triangular mesh of SMPL through linear shape blending and pose skinning[39]. It is the state-of-the-art among available methods for single-view pose and shape estimation[31, 4, 16, 60, 63]. We fine-tuned the pretrained HMR model using the color images and the corresponding ground-truth shape/pose parameters in our synthesis training data.

(2) **BodyNet**. In [64], Varol et al. proposed a neural network for direct inference of volumetric body shape from a single image. The output of BodyNet is a $128 \times 128 \times 128$ occupancy volume with similar definition in Sec.3. BodyNet is the most related work to this paper. We fine-tuned the whole network of BodyNet using the color images and the ground-truth occupancy volumes in our training set.

6.2.2 Comparison Results

We compare against HMR and BodyNet both qualitatively and quantitatively.

(1) For **qualitative comparison**, we feed all the networks with the same images and convert the network output into a triangular mesh. The results are rendered in Fig.6. As shown in the figure, our method is able to achieve much more detailed reconstruction than HMR and BodyNet (See Fig.6(a∼f)). In addition, our method has higher robust-

Figure 7. 3D reconstruction from a monocular video using our method.

ness than BodyNet when some body parts are occluded (See Fig.6(b,g,h)).

(2) The **quantitative comparison** is conducted on the testing set of our synthetic data. We convert the output of HMR to occupancy volumes with a resolution of $128 \times 192 \times 128$. We also upsampled the output of BodyNet by a factor of 1.5 using trilinear interpolation, and then crop the volume to make it have the same resolution. After that we use the mean Intersection-over-Union (IoU score) between predicted 3D volumes and their ground-truth as the comparison metric. It should be noted that the predicted volume and the ground-truth may be unaligned along the z-axis because of depth ambiguities. Therefore, we shift the predicted volume along z-axis to search for the best alignment (i.e., to maximize IoU score between the ground-truth volumes and the predict ones), and regard the maximum IoU score as the final score. The results are presented in Tab.2. As shown by the numerical results, our method achieves the most accurate reconstruction among all the approaches. BodyNet occasionally produces broken bodies and consequently gets the lowest score.

Table 2. Quantitative comparison using 3D IOU score.

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

Table 3. Numerical evaluation of semantic volume/map representation.

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

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[64]. 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.8 shows the evaluation results. The figure shows that compared to sparse joints, a network taking

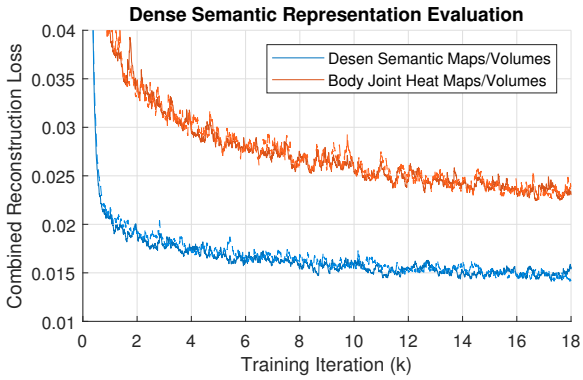


Figure 8. Evaluation of semantic volume/map representation. We evaluate two different inputs for the image-guided vol2vol network: dense semantifc maps/volumes and body joints heat maps/volumes. We show the combined reconstruction losses ($\mathcal{L}_V + \lambda_{FS}\mathcal{L}_{FS} + \lambda_{SS}\mathcal{L}_{SS}$) for different inputs. Solid lines show training error, while dashed lines show validation error (they almost overlap with each other).

dense semantic maps/volumes is able to learn to reconstruct the 3D model more accurately. In Tab.3, we also test these two methods on the testing portion of our synthetic dataset and measure the reconstruction error by calculating the IoU score between 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 may be 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 volumetric feature transformation at the finest scale, while Baseline (B) performs transformation at the coarest 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.

Results. Fig.9 shows the reconstruction loss for different fusing methods. Here, we found 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.9). The same conclusion can be drawn from the visual comparison shown in Fig.10. Using coarsest VFT (Baseline (B)) or latent code concantenation (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.9). The result generated by Baseline (A) is 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.9).

6.3.3 Normal Refinement

Table 4. Numerical normal errors with/without normal refinement.

Error Metric	Cosine Distance	$\ell 2$-norm
Without Refinement	0.0941	0.336
With Refinement	0.0583	0.262

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.4 and Fig.11. In Tab.4 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.11.

7. Discussion

Limitations. Our method relies on HMR to generate a dense semantic representation from SMPL model. Consequently, we cannot give an accurate reconstruction if the estimation of SMPL model is erroneous; see Fig.12 for an example. 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. Our method also fails to recover fine-scale details such as facial expression and hands' motion. 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 further research on 3D vision for humans.

References

[1] Thiemo Alldieck, Marcus A. Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *3DV*, 2018.

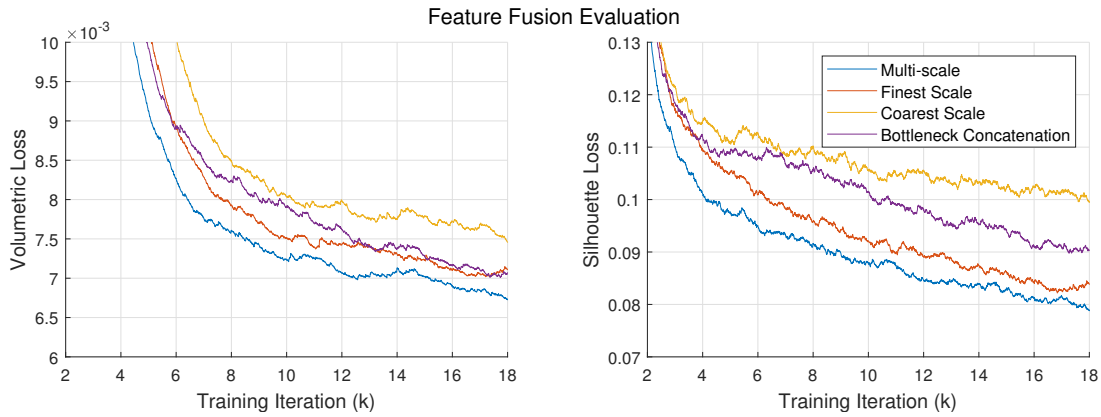


Figure 9. Evaluation of multi-scale volumetric feature transformation (VFT). We evaluate several ways to fuse 2D features into 3D volumes, and show the volumetric loss ($\mathcal{L}_V$) and the silhouette loss ($\mathcal{L}_{FS} + \mathcal{L}_{SS}$) in the figure. For clarity we do not show the validation loss.

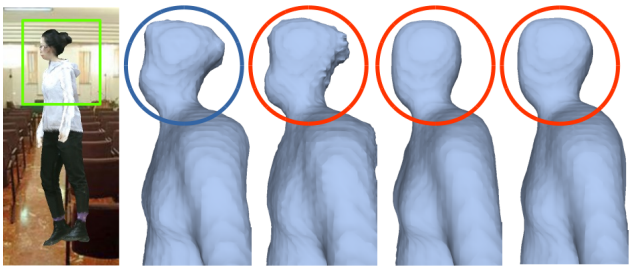
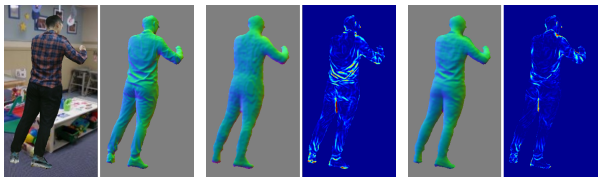


Figure 10. Visual evaluation of multi-scale volumetric feature transformation (VFT). From left to right: input image, head reconstruction result by our method, baseline (A), baseline (B) and baseline (C).



(a) (b) (c)

Figure 11. 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.

Figure 12. A failure case. HMR gives wrong prediction of the subject's upper body pose (middle), which results into wrong reconstruction by our network (right).

[2] Thiemo Alldieck, Marcus A. Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] 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.

[4] 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 *European Conference on Computer Vision (ECCV)*, pages 561–578, 2016.

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, 2017.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: real-time performance capture of challenging scenes. *ACM Trans. Graph*, 35(4):114, 2016.

[11] 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.

[12] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 32(8):1362–1376, 2010.

[13] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1746–1753, 2009.

[14] Andrew Gilbert, Marco Volino, John Collomosse, and Adrian Hilton. Volumetric performance capture from minimal camera viewpoints. In *European Conference on Computer Vision (ECCV)*, pages 591–607, 2018.

[15] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[16] Peng Guan, Alexander Weiss, Alexandru O. Balan, and Michael J. Black. Estimating human shape and pose from a single image. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1381–1388, 2009.

[17] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[18] Riza Alp Güler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 5, 2017.

[19] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. Real-time geometry, albedo and motion reconstruction using a single rgbd camera. *ACM Trans. Graph*, 2017.

[20] 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.

[21] 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.

[22] 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 *European Conference on Computer Vision (ECCV)*, pages 351–369, 2018.

[23] Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 362–379, 2016.

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

[25] 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 International Conference on Computer Vision (ICCV)*, 2017.

[26] Aaron S. Jackson, Chris Manafas, and Georgios Tzimiropoulos. 3d human body reconstruction from a single image via volumetric regression. *CoRR*, abs/1809.03770, 2018.

[27] Wenzel Jakob, Marco Tarini, Daniele Panozzo, and Olga Sorkine-Hornung. Instant field-aligned meshes. *ACM Trans. Graph (Proc. SIGGRAPH ASIA)*, 34(6), Nov. 2015.

[28] 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 International Conference on Computer Vision (ICCV)*, pages 3334–3342, 2015.

[29] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[30] Takeo Kanade, Peter Rander, and P. J. Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE MultiMedia*, 4(1):34–47, 1997.

[31] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *Acm Trans. Graph*, 32(3):1–13, 2013.

[33] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[34] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4704–4713, 2017.

[35] 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.

[36] 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.

[37] Yebin Liu, Juergen Gall, Carsten Stoll, Qionghai Dai, Hans-Peter Seidel, and Christian Theobalt. Markerless motion cap-

ture of multiple characters using multiview image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(11):2720–2735, 2013.

[38] Charles T. Loop, Cha Zhang, and Zhengyou Zhang. Real-time high-resolution sparse voxelization with application to image-based modeling. In *High-Performance Graphics*, pages 73–80, 2013.

[39] 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.

[40] Matthew M. Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Tran. Graph. (Proc.SIGGRAPH Asia)*, 33(6):220:1–220:13, Nov. 2014.

[41] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2659–2668, 2017.

[42] Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J. Gortler, and Leonard McMillan. Image-based visual hulls. In *ACM Trans. Graph. (Proc. SIGGRAPH)*, pages 369–374, 2000.

[43] 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.

[44] 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.

[45] 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.

[46] Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[47] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 483–499, 2016.

[48] 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. Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 741–754, 2016.

[49] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1263–1272, 2017.

[50] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4929–4937, 2016.

[51] 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.

[52] 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.

[53] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1216–1224, 2017.

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[55] 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.

[56] Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5474–5483, 2017.

[57] M. Slavcheva, M. Baust, and S. Ilic. SobolevFusion: 3D Reconstruction of Scenes Undergoing Free Non-rigid Motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[59] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2621–2630, 2017.

[60] Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. In *BMVC*, 2017.

[61] Jonathan Taylor, Jamie Shotton, Toby Sharp, and Andrew Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2012.

[62] Denis Tomè, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5689–5698, 2017.

[63] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *NIPS*, pages 5242–5252, 2017.

[64] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *European Conference on Computer Vision (ECCV)*, 2018.

[65] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4627–4635, 2017.

[66] 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.

[67] 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.

[68] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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

[70] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4732, 2016.

[71] 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 International Conference on Computer Vision (ICCV)*, pages 1108–1115, 2011.

[72] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, 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.

[73] B. Yang, S. Rosa, A. Markham, N. Trigoni, and H. Wen. Dense 3d object reconstruction from a single depth view. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2018.

[74] 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.

[75] Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *IEEE International Conference on Computer Vision (ICCV)*, October 2017.

[76] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[77] Zerong Zheng, Tao Yu, Hao Li, Kaiwen Guo, Qionghai Dai, Lu Fang, and Yebin Liu. Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In *European Conference on Computer Vision (ECCV)*, Sept 2018.

[78] 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 International Conference on Computer Vision (ICCV)*, pages 398–407, 2017.

[79] 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.