Temporal Consistency Loss for High Resolution Textured and Clothed 3D Human Reconstruction from Monocular Video

Akin Caliskan   Armin Mustafa   Adrian Hilton

{a.caliskan, armin.mustafa, a.hilton}@surrey.ac.uk

CVSSP, University of Surrey, UK

Abstract

We present a novel method to learn temporally consistent 3D reconstruction of clothed people from a monocular video. Recent methods for 3D human reconstruction from monocular video using volumetric, implicit or parametric human shape models, produce per frame reconstructions giving temporally inconsistent output and limited performance when applied to video. In this paper we introduce an approach to learn temporally consistent features for textured reconstruction of clothed 3D human sequences from monocular video by proposing two advances: a novel temporal consistency loss function; and hybrid representation learning for implicit 3D reconstruction from 2D images and coarse 3D geometry. The proposed advances improve the temporal consistency and accuracy of both the 3D reconstruction and texture prediction from a monocular video. Comprehensive comparative performance evaluation on images of people demonstrates that the proposed method significantly outperforms the state-of-the-art learning-based single image 3D human shape estimation approaches achieving significant improvement of reconstruction accuracy, completeness, quality and temporal consistency.

1. Introduction

Parsing humans from images is a fundamental task in many applications including AR/VR interfaces [13], character animation [36], autonomous driving, virtual try-on [10] and re-enactment [30]. There has been significant progress in 2D human pose estimation [8, 2], 2D human segmentation [14, 49] and 3D human pose estimation from monocular video [25, 48, 45] to understand the coarse geometry of the human body. Recent research has learnt to estimate full 3D human shape from a single image with impressive results [41, 43, 52, 20, 46, 6]. However temporally consistent textured 3D reconstruction of clothed humans from monocular video remains a challenging problem due
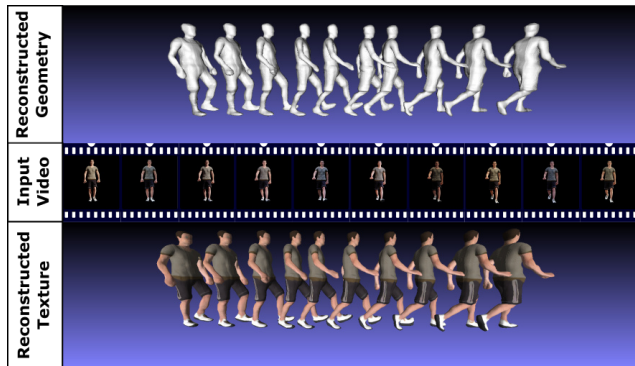

Figure 1. Given an monocular video of a subject (middle), the proposed method creates an accurate and temporally consistent 3D reconstruction (top) with texture (bottom).

to the large variation in action, clothing, hair, camera viewpoint, body shape and pose. This paper addresses this gap in the literature by exploiting a temporally consistent monocular training loss between wide-time separated frames and hybrid implicit-volumetric representation, as shown in Fig. 1.

Traditional multi-view reconstruction methods [34, 28, 12, 50] have demonstrated the advantages of temporally consistent reconstruction [27, 37, 35]. However, temporally consistent 3D human reconstruction from a monocular RGB video remains an open challenge. Parametric model-based 3D human shape estimation methods have been proposed that exploit temporal neural network architectures [23, 24] for a temporally consistent 3D output. Existing parametric models only represent the underlying naked body shape and lack important geometric variation of clothing and hair. Augmented parametric model representations proposed to represent clothing [4] are limited to tight clothing which maps bijectively to body shape and do not accurately represent general apparel such as dresses and jackets.

Recent model-free approaches have achieved impressive results in 3D shape reconstruction of clothed people from a single image using learnt volumetric [52, 20, 38, 39], point cloud [11], geometry image [42] and implicit [43, 44] surface representations. [7] proposed multi-view supervi-

sion to learn complete and view-consistent 3D human reconstruction. ARCH [17] proposed robust 3D reconstruction for arbitrary poses from a single color image and Li et al.[29] proposed 3D human reconstruction from video but process frame-by-frame. These methods are trained only with single-image and 3D model pairs without exploiting temporal information between frames.

We address this problem by proposing a learning framework for textured 3D human reconstruction using temporal consistency for video across wide-timeframes, together with a hybrid 3D volumetric-implicit representation for high resolution textured 3D shape reconstruction. A volumetric shape representation is learnt using a novel temporal loss function between wide-time separated frames which ensures accurate single-view reconstruction of occluded surface regions. The novel loss function learns to incorporate surface photo-consistency cues in the monocular reconstruction which are not present in the observed image or 3D ground-truth shape. Temporal consistency can only be minimized when the predictions of the trained model are consistent and plausible across all temporal views. The proposed method predicts both high resolution 3D geometry and colored texture from a single view for both visible and unseen human parts. The contributions of this work include:

- A novel learning framework for temporally consistent reconstruction of detailed shape and texture for clothed people from monocular video
- Temporal consistency loss based on wide-timeframe coherence of the shape and appearance reconstruction
- A hybrid representation for learning 3D shape which combines the advantages of explicit volumetric representation of occupancy with implicit shape detail
- The first realistic synthesized video dataset of 400 people with ground-truth 3D models

The proposed approach learns a temporally consistent hybrid representations giving significant improvement in the accuracy and completeness of reconstruction compared to the state-of-the-art methods for single image human reconstruction [52, 43, 44, 29, 7].

2. Related Work

2.1. Monocular 3D Human Reconstruction

Parsing 3D humans from a single image can be categorized into model-based and model-free 3D human reconstruction. The first group of approach use parametric human model such as SMPL [32, 3] to estimate the body pose and shape parameters in an iterative manner using either 2D joints locations [22], 2D joints and silhouettes [5] or 3D joints and mesh coordinates [41]. To improve the accuracy of the models, an iterative optimization stage was added to the regression network [26]. Even though parametric model-based methods are able to reliably estimate

the human body from a single image in the wild, estimated shapes are the naked human body without hair, clothing or other surface details. Recent approaches have extended this to tight-fitting clothing [33].

Model-free non-parametric 3D human reconstruction approaches reconstruct clothed people, an overview is given in Table 1. Model-free methods such as Bodynet [46], SiCloPe [38], DeepHuman [52] and MCNet [7] draw a direct inference of volumetric 3D human shape from a single image. However representing 3D human shape in voxels limits the surface resolution for clothing and hair details. Implicit function networks were introduced to obtain high-resolution 3D reconstruction from a single image. PIFU [43] estimates 3D human reconstruction from a single image by proposing an implicit decoder which takes pixel-wise image and depth features and predicts occupancy values of 3D points in the encapsulated volume. Following this, PIFUHD [43] improves the previous method by adding features extracted from surface normal maps to the implicit decoder to reconstruct shape detail on the 3D human. However, both methods can not handle large variations in human pose, clothing and hair. ARCH [17] proposed a variation of the implicit function network using a parametric model fitted to human silhouettes to improve the 3D reconstruction for arbitrary human poses. Among previous methods, PIFU and ARCH not only reconstruct the 3D geometry of human from a single image, but they also predict the complete texture appearance for the reconstruction.

Previous methods consider only a single image without any temporal information or consistency. This results in inconsistent reconstruction of shape and appearance when applied to video sequences. We propose a novel method to learn 3D reconstruction of clothed human from a monocular video using image-3D model pairs together with temporal consistency between video frames and 3D models.

Table 1. Comparison of Single View 3D Reconstruction Methods.

	3D Represent.	Temporal Coherency	Training Data	3D Human Geom. / Text.
Bodynet [46]	Voxel	No	Image	Yes / No
SiCloPe [38]	Voxel	No	Image	Yes / No
DeepHuman [52]	Voxel	No	Image	Yes / No
3DPeople [42]	Geo. Image	No	Image	Yes / No
Mould.Hum. [11]	Point Cloud	No	Image	Yes / No
PIFU [43]	Implicit	No	Image	Yes / Yes
ARCH [17]	Implicit	No	Image	Yes / Yes
PIFU-HD [44]	Implicit	No	Image	Yes / No
MCNet [7]	Voxel	No	Image	Yes / No
Geo-PIFU [15]	Hybrid	No	Image	Yes / No
Proposed	**Hybrid**	**Yes**	**Video**	**Yes / Yes**

2.2. Temporal Consistency in Neural Networks

Previous methods in applications other than 3D shape estimation enforce temporal consistency through a *temporal consistency loss* in the context of style transfer [16], video-to-video synthesis [47] or monocular depth estimation [51, 40]. *Temporal consistency loss* in training or test-
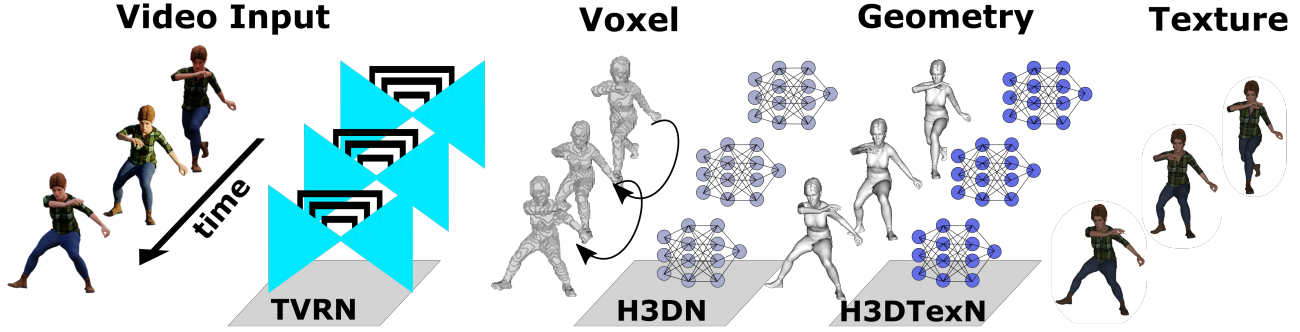
Figure 2. The Proposed Framework for Temporally Consistent 3D Human Reconstruction Learning from Video

ing encourages similar values along the temporal correspondences estimated from the input video. Existing temporally coherent 3D reconstruction methods [27, 37, 35] require multi-view input videos. Applying single-image 3D shape estimation methods independently to each frame in a video often produces flickering results. To address this model-based methods have exploited temporal coherency [23, 24]. However, our aim is to predict temporally consistent model-free 3D clothed human reconstructions from a video, Table 1. A feed-forward network is used to perform single-view 3D human reconstruction for monocular videos and simultaneously maintain 3D temporal consistency between video frames. Our feed-forward network is trained by enforcing the output 3D reconstruction of temporally distant frames to be both accurate and temporally consistent. Monocular video of a moving person provides significant additional information. The body/clothing shape and appearance of the person should be temporally consistent.

2.3. Learning Hybrid 3D Representations

Model-free single-image 3D human reconstruction methods use various 3D representations-voxel, point cloud, geometry image, and implicit, as seen in Table 1. Using voxel representation increases computational cost and limits the 3D surface resolution but it keeps global topology and locality of 3D reconstruction. Implicit surface function representation loses the global topology of the 3D human body during inference, but reconstruct the surface with a high level of shape detail. Recent research has combined multiple 3D representations to exploit their relative advantages in the context of 3D deep learning [31, 15]. [31] represents the 3D input data in points to reduce the memory consumption, while performing the convolutions in voxels to reduce the irregular, sparse data access and improve the locality. On the other hand, [15] combine latent voxel features and implicit function learning for 3D geometry prediction. However [9] shows that using latent features limits ability to learn complex geometry, like articulated shapes, and latent feature approaches do not preserve 3D surface details. Therefore, [9] proposes an improved 3D shape en-

coding which is a rich encoding of 3D data through subsequently convolving it with learned convolutions. In the proposed paper, we design a hybrid implicit-volumetric decoder using the shape encoding from temporally coherent 3D voxel reconstruction and image features to predict both 3D geometry and texture of clothed human in an implicit function learning framework.

3. Temporally Consistent 3D Reconstruction

This section explains the novel proposed method for temporally consistent textured 3D human reconstruction from a monocular video. An overview of the approach is presented in Fig. 2. N frames from a monocular video of a dynamic human with arbitrary pose, clothing, and viewpoint are given as input to the pipeline, and the network predicts the textured 3D human reconstruction in a temporally consistent manner.
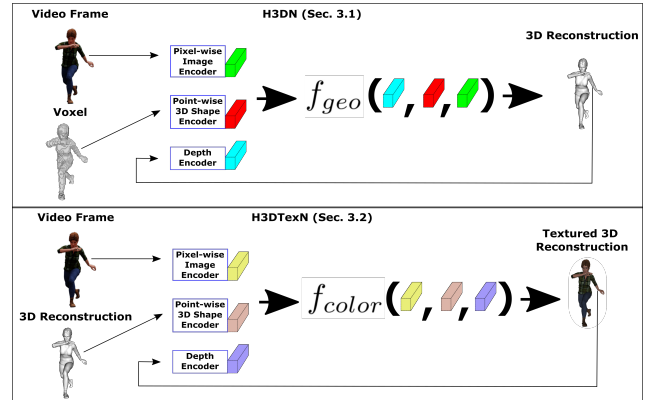

Figure 3. This figure shows Hybrid Implicit 3D Reconstruction Network *(H3DN) and Hybrid Implicit 3D Texture Network (H3dTexN)*.

3.1. Learning 3D Surface Reconstruction

We propose a method to learn 3D surface reconstruction from a monocular video (Fig. 2). In contrast to previous single-view 3D reconstruction approaches, Table 1, that learn 3D human surface reconstruction from single im-

ages, we propose a cascaded network architecture to learn from a monocular video. The proposed architecture consists of *Temporal Voxel Regression Network (TVRN)* and *Hybrid Implicit 3D Reconstruction Network (H3DN). TVRN* reconstructs temporally consistent voxel occupancy grids from a monocular video, and then *H3DN* refines the surface reconstruction. The proposed voxel regression network learns to reconstruct human shape in a temporally consistent manner: N video frames are given to the network such that each frame is passed through its own voxel regression network and parameters are shared between the N networks. The voxel estimation through the TVRN network reduces inconsistency in the reconstruction over time. However, TVRN lacks the high resolution surface details due to the voxel quantisation. Hence the predicted voxel reconstruction is then passed to a hybrid implicit surface function decoder to obtain high-quality surface reconstruction.

3.1.1 Learning Architecture

The proposed learning architecture shown in Fig. 2 consists of two sub-networks, *Temporal Voxel Regression Network (TRVN)* and *Hybrid Implicit 3D Reconstruction Network (H3DN)*. In previous works, voxel regression has been used to address complete 3D reconstruction of people in a wide variety of poses from a single image [52, 7, 20]. Inspired by these approaches we use a voxel regression network architecture to reconstruct the complete topology of the 3D human from single image. To obtain a temporally consistent reconstruction, we introduce a novel learning framework to exploit the temporal consistency between reconstructions from video frames through the proposed *TVRN* architecture (Fig. 2). *TVRN* network consists of multiple parallel stacked hourglass networks with shared parameters. This architecture allows the introduction of a temporal loss function between 3D reconstructions from the input video frames. As seen in the Fig. 2, N frames are used to train the *TVRN* network, including the current frame I_t at time t and $N - 1$ other frames at different time. The *TVRN* network learns temporally consistent 3D shape and predicts the voxel occupancy grid, V for all frames using the proposed loss function which is computed between predicted voxel occupancy grids for different time frames (Sec. 3.1.2).

However, the temporally consistent output of *TVRN* has limited surface detail due to the voxel quantisation. To represent high-resolution shape detail we propose *Hybrid Implicit 3D Reconstruction Network (H3DN)* to refine the temporally consistent voxel occupancy grids. Methods have been proposed in the literature to learn and predict implicit surface representation from a single image [43, 44, 17]. These implicit reconstruction methods give a high-level of details on 3D surface. However, all previous methods lose the complete topology of the human body because of the

sampling scheme during training. In this paper we address this limitation of previous methods by using the voxel occupancy grids as input to the implicit representation instead of a single input image. This allows us to to reconstruct a high-level of surface detail and keep the complete shape topology of the clothed human body due to implicit function learning and input voxel occupancy grids respectively. The implicit surface is obtained using the proposed novel hybrid implicit function network which takes as input multiple feature encodings and predicts the occupancy of a 3D point. In the network, feature encoders from three different inputs (image, voxel and depth) and a Multi-Layer-Perceptron (MLP) as the decoder to predict occupancy values.

As illustrated in Fig. 3, each sampled 3D point ($\mathbf{X}$) is projected on the input image ($\mathbf{x}$) and pixel-wise image features are extracted by concatenating intermediate layer outputs of the hourglass network [43]. We denote the pixel-wise image feature as $\mathcal{H}(I(\mathbf{x}))$. The second input to the decoder are point-wise features extracted from outputs of the *TVRN* network, i.e. voxel occupancy grids. For a sampled 3D point, we apply multi-scale shape encoding [9] in the aligned voxel occupancy grid using trilinear interpolation within the neighborhood of sampled point. Shape encoding for a 3D sampled point ($\mathbf{X}$) is denoted as $\mathcal{S}(\mathbf{X})$. The last input to the proposed decoder is the depth value of sampled 3D points ($\mathbf{X}$) with respect to the camera, denoted $\mathcal{D}(\mathbf{X})$. Overall hybrid implicit surface function is formalized as $f_{geometry}$:

$$f_{geometry}(\mathcal{S}(\mathbf{X}), \mathcal{H}(I(\mathbf{x})), \mathcal{D}(\mathbf{X})) = s : s \in \mathbb{R}, s \in [0, 1] \tag{1}$$

The implicit function predicts occupancy values for the sampled 3D points. Marching cubes is applied to obtain a high-quality surface reconstruction.

3.1.2 Loss Functions

The proposed network is supervised from the ground-truth 3D human models rendered from temporal frames and self-supervised between time distant frames from a monocular video. In order to train the *TVRN* network we combine 3D loss $\mathcal{L}_{3D}^{Voxel}$ and temporal consistency loss $\mathcal{L}_{TC}^{Voxel}$. The 3D loss function $\mathcal{L}_{3D}^{Voxel}$ computes error between the estimated 3D voxel occupancy grid ($\hat{V}_t$) and 3D ground-truth (V_t) for time frame t. As stated in Equation 2, the binary cross entropy [19] is computed after applying a sigmoid function on the network output. In particular, we used weighted binary cross entropy loss and γ is a weight to balance occupied and unoccupied points in the voxel volume:

$$\mathcal{L}_{3D}^{Voxel} = \sum_{t=1}^{N} \mathcal{L}(\mathcal{V}_t, \hat{\mathcal{V}}_t) \tag{2}$$

$$\mathcal{L}(\mathcal{V}_t, \hat{\mathcal{V}}_t) = \sum_x \sum_y \sum_z \gamma \mathcal{V}_t^{xyz} \log \hat{\mathcal{V}}_t^{xyz}$$
$$+ \ (1-\gamma)(1-\mathcal{V}_t^{xyz})(1-\log \hat{\mathcal{V}}_t^{xyz})$$

where $\mathcal{V}^{xyz}$ is the occupancy value of a voxel grid $\mathcal{V}$ at position (x, y, z). Training a network with only binary cross entropy loss gives temporally inconsistent reconstruction for the dynamic parts of the human body as shown in Fig. 2. In order to improve 3D model accuracy and completeness, we propose a second loss function, *temporal consistency loss* ($\mathcal{L}_{TC}$) between reconstructions from multiple video frames. With the *temporal consistency loss*, the representation can learn features robust to temporal changes, self-occlusion and flickering between frames. 3D voxel occupancy grids estimated per frame and the temporal correspondences between vertices are transformed to voxel correspondences as shown in Fig. 4. The *temporal consistency loss* is defined in Equation 3, $L2$ loss is computed between voxel occupancy estimates $\hat{V}$ from one time frame and $N-1$ other frames.

$$\mathcal{L}_{TC}^{Voxel} = \sum_{t=1}^{N} \sum_{\substack{l=1 \\ l \neq t}}^{N} \hat{\mathcal{L}}(\hat{\mathcal{V}}_t, \hat{\mathcal{V}}_l) \qquad (3)$$

$$\hat{\mathcal{L}}(\hat{\mathcal{V}}_t, \hat{\mathcal{V}}_l) = \sum_x \sum_y \sum_z \|\hat{\mathcal{V}}_t^{xyz} - \hat{\mathcal{V}}_l^{\mathcal{P}(xyz)}\|_2$$

where $\mathcal{P}$ is the transformation operator between 3D point correspondences.

In order to train *Hybrid Implicit 3D Reconstruction Network*, we take 3D point samples around the surface of 3D human models and their occupancy values. To create training point samples, we sample a number $n \in \mathbb{N}$ of points $\mathbf{p}_i \in \mathbb{R}^3, i \in 1, \ldots, n$ by sampling points on the ground-truth surfaces for every 3D human model, and adding random displacements $\mathbf{n}_i \sim \mathcal{N}(0, \sigma)$, i.e. $\mathbf{p}_i^s := \mathbf{p}_i + \mathbf{n}_i$. Then we compute the ground truth occupancy values of the sampled points, $o(\mathbf{p}_i^s) \in \{0, 1\}$. If $\mathbf{p}_i^s$ is inside the mesh surface, the corresponding $o(\mathbf{p}_i^s) = 1$ and if not, $o(\mathbf{p}_i^s) = 0$. We train the *geometry prediction network* (Equation 1) by minimizing the average mean squared error:

$$\mathcal{L}_{3D}^{Hybrid} = \frac{1}{n} \sum_{i=1}^{n} \|f_{geo}(\mathcal{S}(\mathbf{X_i}), \mathcal{H}(I(\mathbf{x_i})), \mathcal{D}(\mathbf{X_i})) - o(\mathbf{X_i})\|_2$$

Both the trained networks are then used to estimate temporally consistent 3D reconstruction from a monocular video. The next section describes estimation of temporally consistent texture for each 3D shape estimation.

3.2. Learning Textured 3D Reconstruction

In order to reconstruct complete 3D human models, we propose a temporally consistent texture prediction network in addition to geometry reconstruction, i.e. *Hybrid Implicit 3D Texture Network (H3DTexN)* The proposed method (Fig. 3) predicts the color values for each vertex on the reconstructed surface from a monocular video. Different from
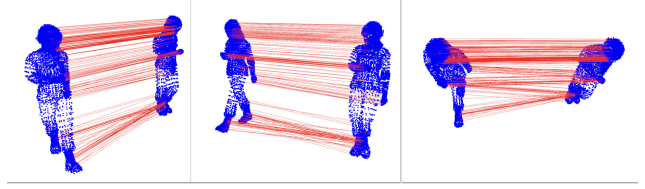

Figure 4. This figure shows temporal 3D vertex correspondences between time distant frames from a monocular video to train the proposed networks: *TVRN* and *H3DTexN* (Sec. 3)

the previous approaches [43, 29], *H3DTexN* learns texture prediction in a temporally consistent manner using the proposed novel hybrid architecture and loss function.

3.2.1 Learning Architecture

The overall method is illustrated in Fig. 3, which contains N stacked hybrid encoder-decoder networks, where N is number of input video frames to the *H3DTexN* network to learn temporally consistent texture. The parameters are shared between the hybrid encoder-decoder networks. Each hybrid encoder-decoder takes multiple feature encodings as input and predicts the RGB color of the sampled 3D point using a Multi-Layer-Perceptron (MLP) decoder to predict color values. During training, as illustrated in Fig. 3, a sampled 3D point ($\mathbf{X}$) is projected on the input image ($\mathbf{x}$) and pixel-wise image feature is extracted by concatenating intermediate layer outputs of a modified Resnet network architecture [21]. We denote the pixel-wise image feature $\mathcal{H}(X)$. The second input to the decoder is point-wise features extracted from output of *Hybrid Implicit 3D Reconstruction Network*, i.e. point-wise occupancy values. For a sampled 3D point, we apply multi-scale shape encoding [9] in the predicted occupancy values within the neighborhood of sampled point using trilinear interpolation. Shape encoding for a sampled point ($\mathbf{X}$) is denoted as $\mathcal{S}(\mathbf{X})$. The last input to the proposed decoder is the depth value of the sampled 3D point ($\mathbf{X}$) with respect to the camera, denoted as $\mathcal{D}(\mathbf{X})$. Overall hybrid implicit texture function is f_{color}:

$$f_{color}(\mathcal{S}(\mathbf{X}), \mathcal{H}(I(\mathbf{x})), \mathcal{D}(\mathbf{X})) = c : c \in \mathbb{R}^{3 \times 1} \qquad (4)$$

The proposed approach learns temporally consistent prediction between different video frames and allows us to use temporal consistency loss (Sec. 3.2.2). In the overall network, we propose a neural network framework to learn temporally consistent 3D textured human reconstruction from video. The details of the image encoders and MLP networks are explained in Sec. 4. In contrast to previous approaches [17, 43], the hybrid decoder captures the global topology of the shape with the shape encoding to predict accurate geometry and texture for 3D reconstruction with minimum computational cost.

3.2.2 Loss Functions

We train *Hybrid Implicit 3D Texture Network (H3DTexN)* by minimizing two loss functions, $\mathcal{L}_{Color}^{Hybrid}$ and $\mathcal{L}_{Color}^{temporal}$. In order to compute $\mathcal{L}_{Color}^{Hybrid}$, we sample a number $m \in \mathbb{N}$ of points $\mathbf{p}_i \in \mathbb{R}^3, i \in 1, \ldots, m$ by sampling points on the ground-truth surfaces for every human model. Then, we obtain the color values of the sampled points, $c(\mathbf{p}_i) \in \mathbb{R}^3$. We train the *texture prediction network* by minimizing the average absolute difference error:

$$\mathcal{L}_{Color}^{Hybrid} = \frac{1}{n} \sum_{i=1}^{n} |f_{color}(\mathcal{S}(\mathbf{X_i}), \mathcal{H}(I(\mathbf{x_i})), \mathcal{D}(\mathbf{X_i})) - c(\mathbf{X_i})|$$
(5)

In order to train the texture prediction network in a temporally consistent manner, we use the temporal correspondences of a sampled point, $\mathbf{p}_i^t$ for $t = \{1, \ldots, N-1\}$, and train the network by minimizing the L_2 loss computed between color estimates, $\hat{c}$, from one time frame and $N-1$ other frames:

$$\mathcal{L}_{Color}^{Temporal} = \sum_{i=1}^{n} \sum_{\substack{l=1 \\ l \neq t}}^{n} \|\hat{c}(\mathbf{X_i}^t) - \hat{c}(\mathbf{X_i}^l)\|_2$$
(6)

$$\hat{c}(\mathbf{X_i}) = f_{color}(\mathcal{S}(\mathbf{X_i}), \mathcal{H}(I(\mathbf{x_i})), \mathcal{D}(\mathbf{X_i}))$$

The overall loss function, $\mathcal{L}$, is the combined loss function for geometry and color:

$$\mathcal{L} = f(\mathcal{L}_{geo}, \mathcal{L}_{color})$$
$$\mathcal{L}_{geo} = h(\mathcal{L}_{3D}^{Voxel}, \mathcal{L}_{TC}^{Voxel}, \mathcal{L}_{3D}^{Hybrid})$$
$$\mathcal{L}_{color} = g(\mathcal{L}_{Color}^{Hybrid}, \mathcal{L}_{Color}^{Temporal})$$

Testing the proposed method with a monocular video input is illustrated 2. Each frame from a monocular video is first passed to *TVRN*. Then, every sample point inside the occupancy volume are given as input to *H3DN* to predict occupancy values for each sample point. Following this, mesh surface reconstruction is estimated from the occupancy volume by applying marching cube algorithm. For appearance, we predict color values of each reconstructed 3D point by using trained *H3DTexN* with the inputs: predicted occupancy volume and a video frame.

4. Experimental Evaluation

This section presents the implementation details and synthetic dataset generation together with qualitative and quantitative result on both images and videos of people with varying pose and clothing. We evaluate the proposed method on monocular videos randomly chosen from the datasets. For each video, we give the network video frames and associated segmentation masks. For a given test video, the proposed method estimates the temporally consistent surface shape and texture appearance reconstruction based on the framework presented in Section 3.

4.1. Datasets

The proposed temporally consistent textured reconstruction from a monocular video is supervised from ground-truth 3D human models and temporal vertex correspondences between video frames. Therefore, we generate a new dataset using similar framework used in the public domain synthetic human image data generation framework *3DVH*[7] (Fig. 7). Since *3DVH* is limited to static images of 3D humans, we generate 30-frame-length video sequences of the 400 human models with large variations in clothing, hair and pose, which are rendered to 100 camera views per frame. This dataset is referred to as *3DVH Video* and will be made available for research.

4.2. Implementation Details

The proposed network is trained on the *3DVH Video* dataset, which is split into training, validation and test sets. In the *temporal voxel regression network (TVRN)* network the size of the input image is $512 \times 512 \times 3$ and output voxel grid resolution is $128 \times 128 \times 128$. In the ground-truth data the points inside and outside the occupied volume are assigned to 1 and 0 values, respectively. During training, batch size is set to 4 and and epochs to 40. With these settings, the network is trained for 3 days using NVIDIA TitanX with 12GB memory. *TVRN* is trained on relatively low memory GPUs restricting the resolution to 128^3; however, we can achieve a higher resolution, because *hybrid implicit 3D reconstruction* module can be trained with 3D shape features extracted from higher voxel resolutions. The Adam optimizer is used with learning rate $lr = 2.5e - 4$ with the decimation of the step-size every 20 epochs.

The *hybrid implicit 3D reconstruction (H3DN)* network is trained with the input set of 3D voxels predicted in *TVRN* and video frames. During training of this network, 3D points are sampled around the surface of the 3D ground-truth model (Sec.3.1.2). To train the network, we use 10000 sample points of inside and outside of 3d ground-truth surface. To extract the point-wise shape features from the voxel, we use the 3D convolution architecture from *IFNET* [9] as it is explained in Sec. 3.1.1. The size of the point-wise features is $[2583 \times 1]$. For the image encoder, we use an hourglass network architecture [43] to get pixel-wise features of size $[256 \times 1]$. For the last feature encoder, *depth encoder*, we normalize the actual depth value of the sample point with respect to the camera. In order to predict the occupancy value for the sampled 3D point, concatenation of these features are passed through a Multi Layer Perceptron (MLP) consisting of 5 linear layers of input/output size $[2849, 1024, 512, 256, 128, 1]$, respectively.

For *texture prediction network*, we use the image encoder adapted from CycleGAN architecture [21] to extract pixel-wise image features and MLP of 5 linear layers of input/output size $[2849, 1024, 512, 256, 128, 3]$. Both the net-
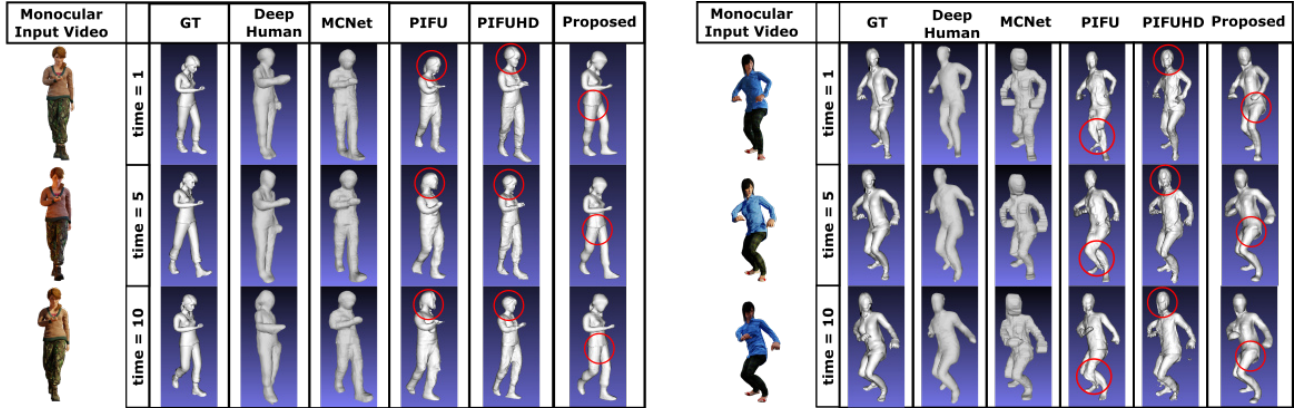
Figure 5. Reconstruction results of Deephuman [52], MCNet [7], PIFU [43], PIFUHD [44] and the proposed method and ground-truth 3D human models.
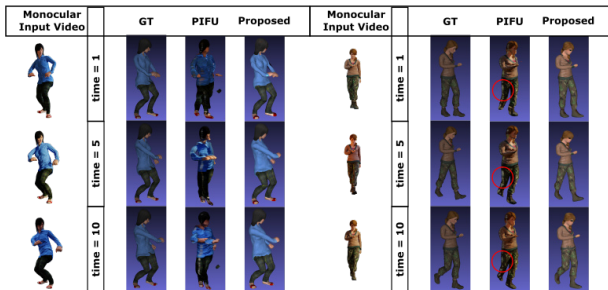

Figure 6. Textured reconstruction results of PIFU [43] and the proposed method and ground-truth 3D human models.
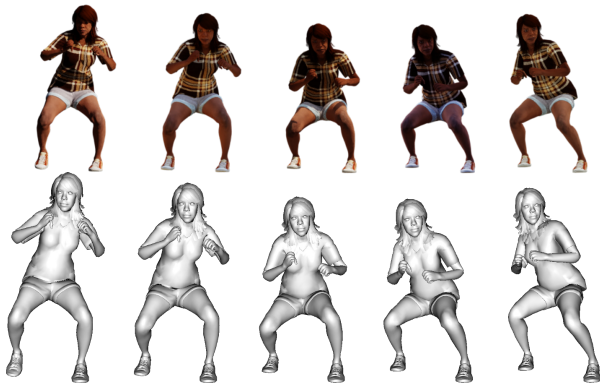

Figure 7. This figure shows a sample sequence from *3DVH Video* Dataset.

works are trained for 5 days using NVIDIA TitanX with 12GB memory, with 2 batch size and 100 epochs. The RM-Sprop optimizer is used with learning rate $lr = 1e - 3$ with the decimation of the step-size every 60 epoch.

4.3. Evaluation

The proposed method is qualitatively and quantitatively evaluated against four recent state-of-the-art deep learning-based methods for single image 3D human reconstruction: DeepHuman [52], PIFU [43], MCNet [7], PIFUHD [44]. To allow fair comparison, we retrain MCNet, PIFU and Deephuman with the *3DVH Video* dataset using the code provided by the authors and use the pre-trained network of PIFUHD (training code unavailable). Qualitative and quantitative comparison of the 3D shape obtained using the proposed approach and the state-of-the-art methods is shown in Fig. 5, 6 and 8, along with the ground-truth. All algorithms are tested with monocular video input, and Fig. 5 illustrates the 3D reconstruction results from side views. These results shows that voxel-based methods, DeepHuman and MCNet, are able to predict the coarse 3D reconstruction without cloth and hair details. On the other hand, implicit surface reconstruction methods, PIFU, PIFUHD can reconstruct better surface details while errors occur in the overall topology of the human body for arbitrary poses. For example, Fig. 5 illustrates that PIFU and PIFUHD predict incorrect reconstructions. Also, previous methods shows temporal inconsistency between adjacent reconstructions over time. For example, the PIFU and PIFUHD methods predict inconsistent 3D reconstructions of clothing, face and hair (Fig. 5). However, the proposed method using a hybrid volumetric-implicit representation trained with a loss-function to enforce temporal consistency results in temporally consistent reconstructions which correctly predicts both body shape and surface details from a single monocular video.

Texture prediction results and comparison of the proposed method with PIFU [43] is shown in Fig. 6. Textured 3D reconstruction results are shown for both visible and unseen part of the person. PIFU shows limited accuracy for unseen parts while the proposed method is able to predict high-quality temporally consistent texture appearance for the complete surface. In PIFU, the texture prediction network uses image features whereas the proposed method en-
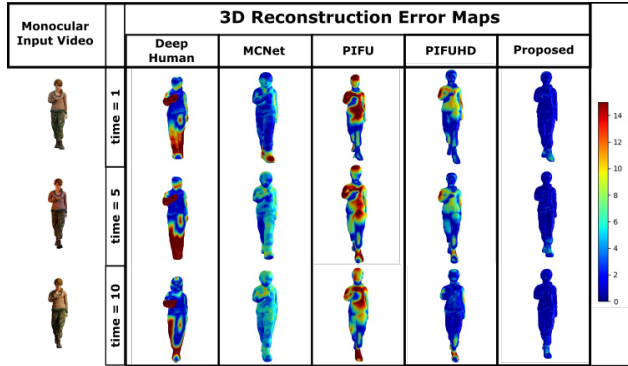
	CD ↓	3D IoU ↑
DeepHuman [51]	7.07	64%
MCNet [6]	1.64	75%
PIFU [42]	6.63	68%
PIFUHD [43]	3.84	70%
Ours	**0.51**	**79%**

Figure 8. **[Top]** Per vertex chamfer distance from reconstruction to ground-truth model is shown. 3D error maps are illustrated for visible and invisible sides.**[Bottom]**Comparison of the proposed method with the state-of-the-art methods for different error metrics. **CD:** Chamfer Distance, **3D IoU**: 3D Intersection of Union For more details, please refer to the text.

codes also the shape features into the hybrid implicit surface decoder. The proposed methods also addresses the temporal consistency between frames of the video in Fig. 6.

In addition to the qualitative results, we compute two error metrics using the ground-truth 3D models to measure the global accuracy of shape reconstruction: Chamfer Distance (CD) and 3D Intersection of Union (3D IoU) [18]. Fig. 8 shows the comparison of results with ground-truth through the error comparison models with the Chamfer distance error coloured from blue to red as error increases (in centimeters). Fig. 8 shows the accuracy of reconstructions and temporal consistency between video frames. The reconstruction obtained using the proposed approach with temporal consistency is significantly better than the 3D shapes obtained using all the previous approaches: DeepHuman, MC-Net, PIFU and PIFUHD. In addition to the improvement in the accuracy of the reconstruction, the 3D shape estimated using the proposed method is temporally consistent and the exploitation of temporal redundancy in the learning framework significantly improves the accuracy and completeness of the estimated 3D shape.

Real Data Evaluation: We evaluate our method against state-of-the-art methods, namely DeepHuman [52], PIFU [43], MCNet [7], PIFUHD [44] on the publicly available TV Presenter [1] dataset, which consists of multiple camera captures of dynamic real humans in a controlled indoor studio. The state-of-the-art methods in figure 9 train their models using real datasets: specifically, DeepHuman [52] is trained on THuman dataset, and PIFu [43] and PIFuHD [44] use the RenderPeople dataset of real human captures. Com-

pared to these methods, the proposed network is trained on the synthetic *3DVH* dataset.

In figure 9, while DeepHuman [52] can recover the coarse shape of human body, the coarse-to-fine approach fails in DeepHuman [52] as limbs are missing. Although, PIFu [43] and PIFuHD [44] can recover the surface details, they reconstruct legs in wrong position. For textured 3D reconstruction, while PIFu [43] performs reasonable on the visible parts, it fails on unseen part of the human body. Unlike the others, our method is able to recover clothed 3D human body and predict the textured on both visible and unseen parts.
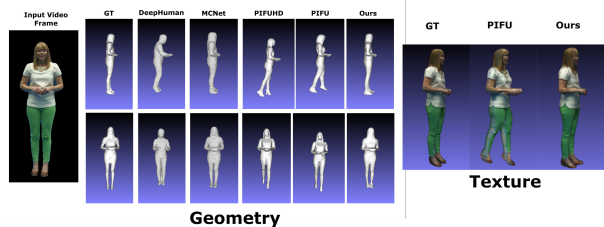

Figure 9. Qualitative results on real image from TV presenter dataset [1].

Limitations: Although the proposed method demonstrates significant improvement in the reconstruction quality over state-of-the-art methods, it suffers from the same limitations as previous methods. The approach assumes complete visibility of the person in the scene and can not handle partial occlusions with objects, as with previous approaches the method also requires silhouettes of the person along with the monocular video for 3D reconstruction.

5. Conclusion and Future Work

This paper introduces a novel method for temporally consistent textured 3D human reconstruction from a monocular video. The approach is trained with a temporal consistency loss and uses a hybrid volumetric-implicit decoder to learn both overall body shape and surface detail. A novel synthetic *3DVH Video* dataset for training is introduced comprising realistic video of 400 people with a wide variation in clothing, hair, body shape, pose and viewpoint. The proposed method demonstrates significant improvement in the reconstruction accuracy, completeness and temporal consistency and improving over state-of-the-art single image methods. Temporal consistency loss together with hybrid implicit decoder are demonstrated to significantly improve the geometry and appearance reconstruction and achieve reliable texture reconstruction of human shape from a monocular video. Future work will exploit self-supervised learning approach for temporally consistent 3D textured human reconstruction from in-the-wild video.

References

[1] Multiview video repository,. http://cvssp.org/data/cvssp3d/. In Centre for Vision Speech and Signal Processing, University of Surrey, UK. 8

[2] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 1

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. 2

[4] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5419–5429, 2019. 1

[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 *European Conference on Computer Vision*, pages 561–578. Springer, 2016. 2

[6] Akin Caliskan, Armin Mustafa, Evren Imre, and Adrian Hilton. Learning dense wide baseline stereo matching for people. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[7] Akin Caliskan, Armin Mustafa, Evren Imre, and Adrian Hilton. Multi-view consistency loss for improved single-image 3d reconstruction of clothed people. *arXiv preprint arXiv:2009.14162*, 2020. 1, 2, 4, 6, 7, 8

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 1

[9] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6970–6981, 2020. 3, 4, 5, 6

[10] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9026–9035, 2019. 1

[11] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2232–2241, 2019. 1, 2

[12] Andrew Gilbert, Marco Volino, John Collomosse, and Adrian Hilton. Volumetric performance capture from minimal camera viewpoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 566–581, 2018. 1

[13] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (TOG)*, 38(6):1–19, 2019. 1

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[15] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *arXiv preprint arXiv:2006.08072*, 2020. 2, 3

[16] Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and Wei Liu. Real-time neural style transfer for videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 783–791, 2017. 2

[17] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 2, 4, 5

[18] Krishna Murthy J., Edward Smith, Jean-Francois Lafleche, Clement Fuji Tsang, Artem Rozantsev, Wenzheng Chen, Tommy Xiang, Rev Lebaredian, and Sanja Fidler. Kaolin: A pytorch library for accelerating 3d deep learning research. *arXiv:1911.05063*, 2019. 8

[19] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1031–1039, 2017. 4

[20] Aaron S Jackson, Chris Manafas, and Georgios Tzimiropoulos. 3d human body reconstruction from a single image via volumetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 1, 4

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5, 6

[22] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 2

[23] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019. 1, 3

[24] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 1, 3

[25] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1086, 2019. 1

[26] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose

and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2252–2261, 2019. 2

[27] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Multi-view dynamic shape refinement using local temporal integration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3094–3103, 2017. 1, 3

[28] Vincent Leroy, Jean-Sebastien Franco, and Edmond Boyer. Shape reconstruction using volume sweeping and learned photoconsistency. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1

[29] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. *arXiv preprint arXiv:2007.13988*, 2020. 2, 5

[30] Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. Neural rendering and reenactment of human actor videos, 2018. 1

[31] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. In *Advances in Neural Information Processing Systems*, pages 965–975, 2019. 3

[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. 2

[33] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3d people in generative clothing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[34] Armin Mustafa, Hansung Kim, Jean-Yves Guillemaut, and Adrian Hilton. General dynamic scene reconstruction from multiple view video. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1

[35] Armin Mustafa, Hansung Kim, Jean-Yves Guillemaut, and Adrian Hilton. Temporally coherent 4d reconstruction of complex dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4660–4669, 2016. 1, 3

[36] A. Mustafa, C. Russell, and A. Hilton. U4d: Unsupervised 4d dynamic scene understanding. In *ICCV*, 2019. 1

[37] Armin Mustafa, Marco Volino, Jean-Yves Guillemaut, and Adrian Hilton. 4d temporally coherent light-field video. In *2017 International Conference on 3D Vision (3DV)*, pages 29–37. IEEE, 2017. 1, 3

[38] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4480–4490, 2019. 1, 2

[39] Hayato Onizuka, Zehra Hayirci, Diego Thomas, Akihiro Sugimoto, Hideaki Uchiyama, and Rin-ichiro Taniguchi. Tetratsdf: 3d human reconstruction from a single image with a tetrahedral outer shell. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6011–6020, 2020. 1

[40] Vaishakh Patil, Wouter Van Gansbeke, Dengxin Dai, and Luc Van Gool. Don't forget the past: Recurrent depth estimation from monocular video. *arXiv preprint arXiv:2001.02613*, 2020. 2

[41] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018. 1, 2

[42] Albert Pumarola, Jordi Sanchez, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3DPeople: Modeling the Geometry of Dressed Humans. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[43] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2304–2314, 2019. 1, 2, 4, 5, 6, 7, 8

[44] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 1, 2, 4, 7, 8

[45] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2500–2509, 2017. 1

[46] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36, 2018. 1, 2

[47] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018. 2

[48] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10965–10974, 2019. 1

[49] Lu Yang, Qing Song, Zhihui Wang, and Ming Jiang. Parsing r-cnn for instance-level human analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 364–373, 2019. 1

[50] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7287–7296, 2018. 1

[51] Haokui Zhang, Chunhua Shen, Ying Li, Yuanzhouhan Cao, Yu Liu, and Youliang Yan. Exploiting temporal consistency for real-time video depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1725–1734, 2019. 2

[52] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a

single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7739–7749, 2019. 1, 2, 4, 7, 8