Image-Guided Human Reconstruction via Multi-Scale Graph Transformation Networks

Kun Li, *Member, IEEE,* Hao Wen†, Qiao Feng†, Yuxiang Zhang, Xiongzheng Li, Jing Huang, Cunkuan Yuan, Yu-Kun Lai, *Member, IEEE,* and Yebin Liu, *Member, IEEE*

***Abstract*—3D human reconstruction from a single image is a challenging problem. Existing methods have difficulties to infer 3D clothed human models with consistent topologies for various poses. In this paper, we propose an efficient and effective method using a hierarchical graph transformation network. To deal with large deformations and avoid distorted geometries, rather than using Euclidean coordinates directly, 3D human shapes are represented by a vertex-based deformation representation that effectively encodes the deformation and copes well with large deformations. To infer a 3D human mesh consistent with the input real image, we also use a perspective projection layer to incorporate perceptual image features into the deformation representation. Our model is easy to train and fast to converge with short test time. Besides, we present the D^2Human (Dynamic Detailed Human) dataset, including variously posed 3D human meshes with consistent topologies and rich geometry details, together with the captured color images and SMPL models, which is useful for training and evaluation of deep frameworks, particularly for graph neural networks. Experimental results demonstrate that our method achieves more plausible and complete 3D human reconstruction from a single image, compared with several state-of-the-art methods. The code and dataset are available for research purposes at *http://cic.tju.edu.cn/faculty/likun/projects/MGTnet*.**

***Index Terms*—3D Human Reconstruction, Single Image, Deformation Representation, Graph Neural Networks, Dataset**

I. Introduction

Recovering a 3D human model from a single image is an important and challenging problem, which has a wide range of applications in VR/AR content creation [3], motion analysis [4], and virtual try-on [5]. Topology-consistent 3D reconstruction is very important to reduce the storage and jitters, which is also helpful to learn parametric models and generate high-quality textures. Considering the sensitivity to initialization of traditional optimization-based methods [6]–[8], learning-based methods have drawn much attention from both academia and industry. Many methods [9], [10] predict the parameters of a statistical body model, *e.g.*, SMPL (Skinned Multi-Person Linear model) [11] to achieve 3D human shape recovery, but this parametric shape cannot represent clothing details. Non-parametric volumetric approaches [2], [12] can estimate complete shapes, but they are limited by the resolution of the output grid and large memory requirement. Therefore, these methods fail to recover the details of face and hands. Alldieck *et al.* [13] achieve detailed inference of body shapes including face, hair and clothing, but they require a frontal photo as input and the recovered pose is restricted to A-pose for images not included in their training dataset. PIFU [14] generates detailed human recovery results using an implicit representation, but this method requires an extra mask image and has difficulties for complex poses. PIFuHD [15] proposes an end-to-end trainable coarse-to-fine framework for high-resolution 3D clothed human reconstruction at 1k image resolution. All the above non-parametric methods use weak-projection and cannot obtain topology-consistent 3D models, which is important for editing and reducing jitters and storage for videos. To obtain topology-consistent detailed 3D models, Zhu *et al.* [16] use four stages constrained by joints, silhouettes and shading, due to lack of detailed 3D human datasets. However, this method sometimes generates geometry details irrelevant to the input image. The Euclidean convolutional operation assumes self-similarity under rigid transformations which is not suitable for non-rigid deformations. It is more effective to use *graph convolutions* directly on the mesh. Kolotouros *et al.* [17] recover human body meshes using a graph convolutional neural network, but the regressed mesh may have artifacts for complex poses because Euclidean coordinates have obvious limitations on rotations for mesh deformation. Gao *et al.* [18] develop a simple but effective representation that addresses ambiguities of local rotation axes and rotation angles through as-consistent-as-possible optimization, which has been widely used in localized deformation component analysis [19] and unpaired shape deformation transfer [20].

To obtain topology-consistent deformed human models, in this paper, we propose a novel deep learning framework with cascaded multi-scale graph transformation networks. The 3D clothed human shapes are represented by a vertex-based deformation representation rather than Euclidean coordinates which well deals with large deformations and avoids distorted geometries. To generate shapes that well correspond to given real images, we also design a perspective projection layer to incorporate appearance features into the deformation representation. Our model is easy to train and fast to converge with short test time. Besides, we present the D^2Human

† Equal contribution.

This work was supported in part by Tianjin Research Program of Application Foundation and Advanced Technology (18JCYBJC19200).

Kun Li, Hao Wen, Qiao Feng, Xiongzheng Li and Jing Huang are with the College of Intelligence and Computing, Tianjin University, Tianjin 300350, China.

Yu-Kun Lai is with the School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, United Kingdom.

Cunkun Yuan is with Beijing Kuaishou Technology Co., Ltd, 6 Shangdi West Road, Haidian District, Beijing, 100092, China. This work was done when he was with Tianjin University.

Yuxiang Zhang and Yebin Liu are with the Department of Automation, Tsinghua University, Beijing 10084, China.

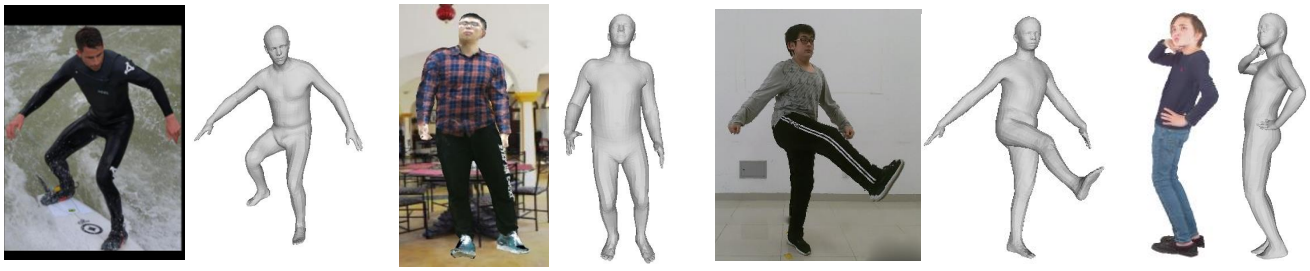

Figure 1. Given a single color image containing a clothed person, our method can reconstruct a topology-consistent 3D human mesh for various poses on different datasets. From left to right shows results on LSP [1], THuman [2], D^2Human, and scanned datasets.

(Dynamic Detailed Human) dataset, which contains thousands of topology-consistent dynamic human meshes with geometry details together with the real captured color images and SMPL models. Experimental results demonstrate that our method achieves more complete and plausible 3D human recovery from a single image, compared with several state-of-the-art methods. Our method aims to reconstruct topology-consistent human models with large poses and the recovered human shape is consistent with the input image. Detailed geometries can be reconstructed by optimization using an estimated normal map as demonstrated in Section V-D. Figure 1 shows some reconstructed examples of our method on different datasets. The code and dataset are available for research purposes at *http://cic.tju.edu.cn/faculty/likun/projects/MGTnet*.

The main contributions of this work are summarized as:

- We design cascaded multi-scale graph transformation networks for clothed 3D human reconstruction from a single color image, which avoids fake geometry details inconsistent with the input image by using a perspective projection layer to incorporate appearance features into the deformation representation. This makes our model applicable to real images.
- Instead of directly using 3D coordinates, we encode a 3D mesh with flexible deformations in a compact latent space, which well deals with large deformations and avoids distorted geometries at extreme pose or viewpoint. This makes our model robust to various poses.
- We present the D^2Human (Dynamic Detailed Human) dataset, including variously posed 3D detailed human meshes with consistent topologies (19019 vertices) and continuous motions for 20 subjects (15 male and 5 female) together with real captured color images and SMPL models. It is particularly useful for training and evaluation of deep neural networks.

II. Related Work

A. 3D Human Recovery from a Single Image

Human shape reconstruction from images is a popular research area, which has a wide range of applications in various fields. Despite the great progress in human reconstruction from multi-view images [21]–[25] or a video [26]–[28], we focus on more relevant work that recovers the full body shape from a single image, which can be categorized into parametric methods and non-parametric methods.

1) Parametric Methods: Parametric methods rely on a pre-trained generative human model, *e.g.*, SCAPE [29], SMPL [11] and SMPL-X [30]. Given the manually annotated 2D landmarks and the smooth shading, Guan *et al.* [31] recover the human shape and pose by optimizing the parameters of the SCAPE model. Dibra *et al.* [32] achieve automatic estimation of the SCAPE model using a convolutional neural network (CNN). Recently, the SMPL model [11] has drawn much attention from academia and industry due to its efficiency and flexibility. Bogo *et al.* [6] propose an optimization-based method called SMPLify to fit the SMPL model. Lassner *et al.* [7] build a dataset of 3D bodies each annotated with 91 landmarks and 31 segments and use a random forest to recover the SMPL model. Zanfir *et al.* [8] address the monocular inference problem for multiple interacting people by providing a model for 2D and 3D pose and shape reconstruction over time. However, these traditional optimization-based methods are sensitive to initialization and easy to fall into local minima. Tan *et al.* [33] design a two-phase indirect learning procedure to regress the parameters of body shape and pose from real images without requiring any ground-truth parameters. Pavlakos *et al.* [7] propose an efficient and effective direct prediction method based on two sub-networks by first estimating the silhouette and 2D joints and then predicting the parameters of the SMPL model, which can be fine-tuned end-to-end without requiring images with 3D shape ground truth. Kanazawa *et al.* [9] design an end-to-end framework to regress the parameters of SMPL using a weakly supervised approach relying on 2D keypoint reprojection and a pose prior learned in an adversarial manner. Omran *et al.* [34] achieve more effective shape and pose estimation by using body part segmentation. Kolotouros *et al.* [10] present a self-improving approach for training a neural network through the tight collaboration of a regression method and an optimization method (SMPLify [6]). Considering the difficulty of obtaining natural images with 3D ground truth, Rüegg *et al.* [35] propose a new deep learning architecture that facilitates unsupervised or lightly supervised learning. SMPL-X [30] can express finger motions and facial expressions compared to SMPL [11]. Rong *et al.* [36] build a fast motion capture system based on SMPL-X to estimate the compatible output of 3D hands and 3D body from a single RGB image by designing the body and hand expert modules and integration strategies. Choutas *et al.* [37] present a fast and accurate model for holistic expressive body reconstruction by estimating SMPL-X parameters directly. Although directly estimating the parameters of the SMPL or SMPL-X model

is simple and efficient, the results generated by the above methods are over-smooth and without geometry details.

2) Non-parametric Methods: Recently, non-parametric approaches have also been proposed for human body estimation, which directly predict the 3D representation from the image. Some methods [2], [12], [38] regress the volumetric representation using convolutional neural networks, but these methods require intensive memory and have limited resolution of the model. In order to overcome the limitation for the resolution of the output grid, Gabeur *et al.* [39] employ a double depth map to represent the 3D shape of a person, which allows a higher resolution output with a much lower dimension. Natsume *et al.* [40] introduce a new silhouette-based representation for modeling clothed human bodies using deep generative models, inspired by the visual hull algorithm. However, this method needs a segmented 2D silhouette and inferred 3D joints as inputs. I2L-MeshNet [41] is an image-to-lixel (line+pixel) prediction network that converts the output of the network to the lixel-based 1D heatmap, which preserves the spatial relationship in the input image and models the uncertainty of prediction. All the above methods are difficult to generate fine details, especially for the face and hands. In order to obtain detailed shapes, Zhu *et al.* [16] use four stages constrained by joints, silhouettes and shading, due to lack of detailed 3D human datasets. However, this method sometimes generates image-irrelevant geometry details, especially for the head. Alldieck *et al.* [13] turn the shape regression problem into an image-to-image conversion problem by using UV mapping. However, this method can only obtain the result of A-pose for test images not included in their training dataset. Saito *et al.* [14] propose an efficient implicit function representation method, PIFu, and the corresponding end-to-end deep learning method. This method can reconstruct fine geometry details from a single image or multiple images. However, it takes a long time and is difficult to reconstruct detailed face. PIFuHD [15] proposes an end-to-end trainable coarse-to-fine framework for high-resolution 3D clothed human reconstruction at 1k image resolution, but it is difficult to deal with complex poses. For more efficient operations, Kolotouros *et al.* [17] directly regress the human body mesh using a graph convolutional neural network, but the recovered mesh tends to be smooth and lacks fine details. Non-parametric methods can recover more detailed shapes than parametric methods, but they may have large distortion or unstable results due to the uncertainty in high dimensions.

In this paper, we propose an efficient and effective method by taking advantages of both parametric methods and non-parametric methods. Specifically, we use the estimated SMPL as an initialization to regress the clothed human model through cascaded hierarchical graph transformation networks featured by an effective vertex-based deformation representation.

B. 3D Human Datasets

Most of 3D human datasets are available to evaluate 2D or 3D pose estimation. For example, HumanEva [42] and Human3.6M [43] datasets contain multi-view video sequences with ground-truth 3D skeletons and motions captured using a marker-based motion capture system. However, the captured people need to wear markers or special suits, and hence natural clothes are not captured. MPI-INF-3DHP dataset [44] has the clothing appearance by using a markerless motion capture method. However, all the above datasets have no 3D human models for each temporal frame. In order to provide a dynamic 3D model dataset, Varol *et al.* [45] present SURREAL (Synthetic hUmans foR REAL tasks) dataset, which is generated by rendering SMPL models with different clothing textures using motion capture data. This dataset also contains ground-truth pose, depth maps and segmentation masks. 3D People dataset [46] is also a synthetic human dataset with photo-realistic images of 80 subjects performing 70 activities and wearing diverse outfits, but it does not share the 3D meshes for copyright reasons. The "Unite the People" dataset [7] and the THuman dataset [2] provide real-world human images and 3D SMPL models which lack geometry details. The BUFF dataset [47] provides high-resolution 3D scan sequences of 3 males and 2 females wearing 2 clothing styles but lacks real images. All the above methods do not provide topology-consistent dynamic human meshes with detailed surface geometry, which are very useful for training and evaluating graph-based deep neural networks. The CAPE [48] dataset is a dynamic 3D clothed human model dataset with consistent SMPL mesh topology (6890 vertices), but the geometry details are limited due to using a small number of vertices. Moreover, the color images are re-rendered, not real captured.

In this paper, we present the D^2Human dataset, which includes variously posed 3D human meshes with consistent topologies (19019 vertices) and continuous motions for 20 subjects (15 male and 5 female) together with real captured color images and SMPL models.

III. Method

Define a 3D human mesh as a set of vertices and edges, $\mathcal{M} = (\mathbf{V}, \mathcal{N})$, with $|\mathbf{V}| = n$ vertices that lie in 3D Euclidean space, $\mathbf{V} \in \mathbb{R}^{n\times 3}$. The adjacency matrix $\mathcal{N}$ is a collection of edge sets that represents the neighborhood for each vertex. We calculate a 9-dimensional deformation representation vector [18] for each vertex, and obtain a new graph $\mathbf{G} \in \mathbb{R}^{n\times 9}$ as the input to our network.

A. Effective Deformation Representation

Euclidean coordinates are the most straightforward way to represent each deformed shape, but with obvious limitations on rotations. To avoid distortion during deformation, we use an effective deformation representation as the input to train our graph neural network. Assume $\mathcal{S}$ is a dataset of $N(N > 1)$ shapes with the same connectivity. S_m is the m-th shape in $\mathcal{S}$, and $p_{m,i} \in \mathbb{R}^3$ is the Euclidean coordinates of the i-th vertex on S_m. Specifically, we use the canonical model with the same topology as the reference model. The deformation gradient of the i-th vertex on S_m, $T_{m,i} \in \mathbb{R}^{3\times 3}$, can be obtained by minimizing the energy:

$$E(T_{m,i}) = \sum_{j\in\mathcal{N}_i} c_{ij} \|(p_{m,i} - p_{m,j}) - T_{m,i}(p_{1,i} - p_{1,j})\|_2^2, \quad (1)$$

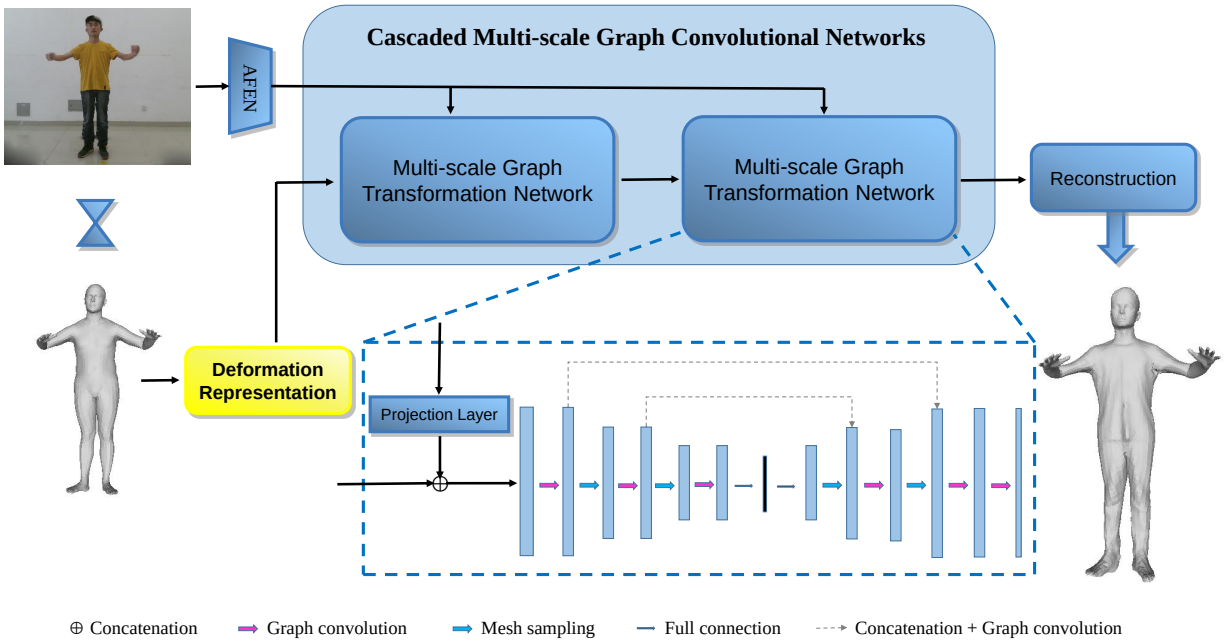

Figure 2. Overview of our framework. Given an input color image, the appearance feature extraction network (AFEN) extracts appearance features, which guide the graph convolution using perspective projection, and the cascaded multi-scale graph transformation networks regress the detailed mesh from the deformation representation of the estimated SMPL model.

where c_{ij} is the cotangent weight and $\mathcal{N}_i$ represents the 1-ring neighbors of the i-th vertex. For the rank-deficient case, we add the normal of the plane to the 1-ring edges for computing the deformation gradient to ensure a unique solution. However, the deformation gradient representation cannot handle large-scale rotations. To address this problem, Gao *et al.* [18] decompose the deformation gradient $T_{m,i}$ into a rotation matrix and a scaling/shear matrix as $T_{m,i} = R_{m,i}S_{m,i}$. By axis-angle representation, $R_{m,i}$ can be represented using a rotation axis $\omega_{k,i}$ and rotation angle $\theta_{k,i}$. The logarithm of $R_{m,i}$ is a skew-symmetry matrix, and hence we extract a 3-dimensional vector for $R_{m,i}$. Because $S_{m,i}$ is a symmetry matrix, we extract a 6-dimensional vector for $S_{m,i}$ using non-duplicated entries. Finally, we concatenate these vectors to a 9-dimensional deformation representation vector $q_{m,i} \in \mathbb{R}^9$ for each vertex.

To reconstruct the 3D mesh from deformation representation, we solve a linear system like [18] and speed up the algorithm by precomputation and union operation.

B. Network

As shown in Figure 2, our network includes two sub-networks: appearance feature extraction network and multi-scale graph transformation network. The appearance feature extraction network uses six residual blocks to extract appearance features and guide the graph convolution using perspective projection. The multi-scale graph transformation network is a graph-based encoder-decoder network that contains topologies of three different scales and two skip connections.

1) Appearance Feature Extraction Network: We use ResNet [49] to extract the detailed appearance features of human 3D mesh from images, because ResNet has good performance for detailed feature extraction. Our sub-network has 6 residual blocks with two convolution operations, and the output of each residual block is denoted as $layer_i, i = 1, 2, ..., 6$. We concatenate the $layer_2, layer_4, layer_6$ as the output of appearance feature extraction network: $\mathbf{I}$. All the convolution operations use a 3×3 filter and preserve the size and channel of the feature maps. Therefore, the size of the network output $\mathbf{I}$ is the same as the input image, and the number of channels is 9.

2) Multi-scale Graph Transformation Network: Our multi-scale graph transformation network consists of three parts: *an image projection layer*, *an encoder* and *a decoder*. In the image projection layer, we first reconstruct the original mesh representation $\mathbf{S}$ from the current deformation representation $\mathbf{G}$ as explained in Section III-A. Then, given the vertex coordinates of the mesh, we calculate the image coordinates of 2D perspective projections of mesh vertices $\mathbf{V}$ through the camera parameters. Bidirectional linear interpolation is used to extract the 4 pixels closest to the vertex image coordinates from $\mathbf{I}$. The weighted average of image features at these pixels gives the image feature for the corresponding vertex, and image features of all vertices $\mathbf{V}$ are denoted as $\mathbf{U} \in \mathbb{R}^{\mathbf{n}\times 9}$. The image features $\mathbf{U}$ are finally concatenated with the deformation representation $\mathbf{G} \in \mathbb{R}^{n\times 9}$ as the input of multi-scale graph transformation network: $\mathbf{F} \in \mathbb{R}^{n\times 18}$.

The encoder E encodes the graph $\mathbf{F}$ with mixed features into a latent vector $\mathbf{z} = E(\mathbf{F})$, and the decoder D decodes the latent vector into a recovered graph $\mathbf{G}' = D(\mathbf{z})$. The final mesh can be calculated by reconstructing the mesh coordinate representation from the deformation representation $\mathbf{G}'$ as explained in Section III-A. The skip connections contain cascading and convolution operations, which are used to prevent the loss of the original details during the up-sampling.

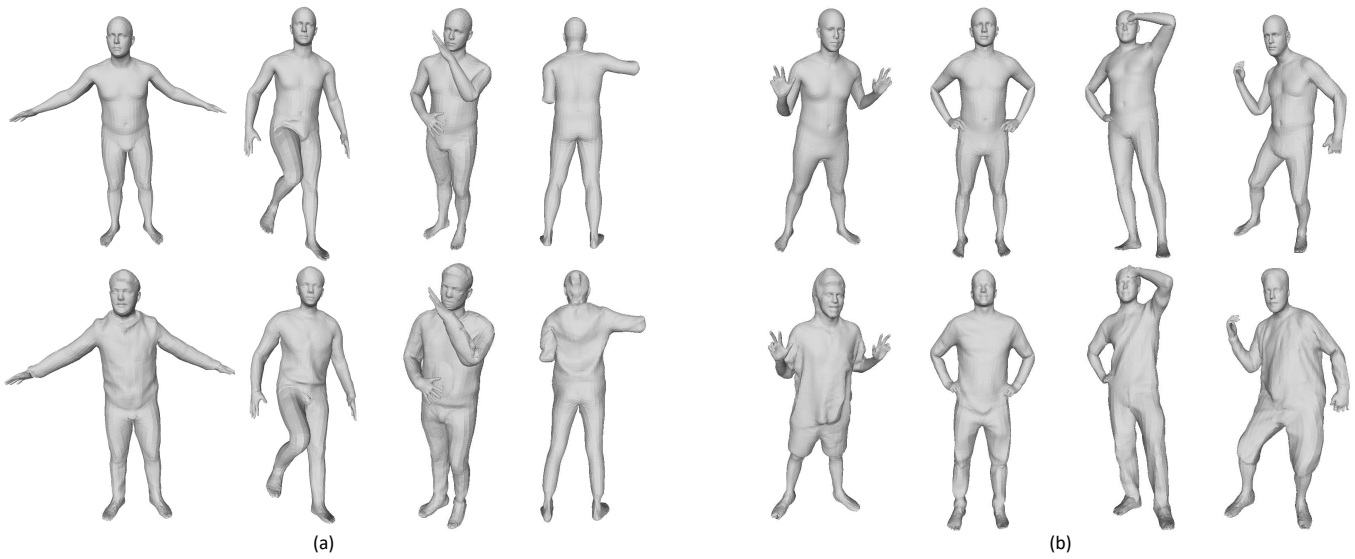

Figure 3. Some examples of our two datasets: (a) D^2Human Dataset, and (b) Scanned Dataset. The top row shows the up-sampled SMPL or SMPL-X models, and the bottom row shows our registered results.

Traditional CNNs cannot deal with such irregular data graphs, and thus we use dynamic filtering convolutional layers [50] to process the graph. It can learn the mapping from the neighborhood patch to filter weights, which considers the intrinsic characteristic of the mesh. Specifically, the input to a layer is a feature vector $\mathbf{x}_i$ associated with a vertex $i \in \{1, ..., n\}$, and the output is also a vector $\mathbf{y}_i$:

$$\mathbf{y}_i = \mathbf{b} + \sum_{m=1}^{M} \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} e_m(\mathbf{x}_i, \mathbf{x}_j) \mathbf{W}_m \mathbf{x}_j, \tag{2}$$

where $\mathcal{N}_i$ is the set of neighbors of vertex i, and $\{\mathbf{W}_m \in \mathbb{R}^{N_x \times N_y}\}$ is a set of M weight matrices for the filters. N_x and N_y are the dimensions of $\mathbf{x}_i$ and $\mathbf{y}_i$ respectively. $e_m(\mathbf{x}_i, \mathbf{x}_j) \propto \exp(\mathbf{t}_m^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j) + c_m)$ are positive edge weights in the patch normalized to sum to one over m, which leads to translation invariance of the weights in the feature space. $\mathbf{b}$, $\mathbf{W}_m$, $\mathbf{t}_m$ and c_m are trainable weights, and M is a fixed design parameter.

To capture global and local features, we also achieve multi-scale convolution on meshes by using mesh sampling [51] to get a new topology and connection relationship for the mesh. Specifically, we down-sample a mesh with n vertices to k vertices ($n > k$) using permutation matrix $P_d \in \{0, 1\}^{k \times n}$ by iteratively contracting vertex pairs, which uses a quadratic matrix [52] to maintain surface error approximations. $P_d(p, q)$ denotes whether the q-th vertex is kept during down-sampling: $P_d(p, q) = 1$ if the vertex is kept, and 0 if it is discarded. The down-sampled vertex set $\mathbf{V}_d$ is a subset of the original mesh. The process of up-sampling is to re-add the vertices v_q discarded during the down-sampling process into the down-sampled mesh, *i.e.* mapping v_q into the closest triangle (h, i, j) in the down-sampled mesh and computing the barycentric coordinates by $\widetilde{v} = w_h v_h + w_i v_i + w_j v_j$ where $v_h, v_i, v_j \in \mathbf{V}_d$ and $w_h + w_i + w_j = 1$. The weights in up-sampling matrix $P_u \in \mathbb{R}^{n \times k}$ are then updated as $P_u(q, h) = w_h$, $P_u(q, i) = w_i$, $P_u(q, j) = w_j$, and $P_u(q, l) = 0$ otherwise. The up-sampled vertices $\mathbf{V}_u = P_u \mathbf{V}_d$.

3) Implementation Details: In our multi-scale graph transformation network, the encoder consists of 3 graph convolutions with filter dimensions of 16, accompanied by a batch normalization [53] and a ReLU activation [54]. The ratios for two down-samplings are $[4, 4]$. The last layer of encoder is a full connection used to convert the feature graph into the latent vector $\mathbf{z} \in \mathbb{R}^{128}$. The decoder starts from a full connection, recovering feature graph structure by up-sampling and convolutions with symmetrical parameters. The last convolution converts feature graph into $\mathbf{G}' \in \mathbb{R}^{\mathbf{n} \times 9}$. From left to right, the feature sizes are 19019×18, 19019×16, 6890×16, 6890×16, 1723×16, 1723×16, 128, 1723×16, 6890×16, 6890×16, 19019×16, 19019×16, and 19019×9. For the training samples, we remove the outliers to keep reasonable data distribution.

4) Loss Function: In our model, each multi-scale graph transformation network block uses the same loss on the deformation representation defined as

$$Loss = \sum_{k=1}^{K} \sum_{m=1}^{N} ||\mathbf{G}'_{\mathbf{m}} - \hat{\mathbf{G}}_{\mathbf{m}}||^2, \tag{3}$$

where K and N are the number of multi-scale graph transformation network blocks and the number of shapes, respectively. $\hat{\mathbf{G}}_{\mathbf{m}}$ is the ground-truth deformation representation corresponding to the detailed mesh.

IV. Datasets

In order to train and evaluate our graph neural network, we collect two datasets, a real-captured dynamic detailed human dataset and a synthesized dataset using varied static scanned human models. Due to copyright limitation, we will only publish the first dataset for research purposes.

A. D^2Human Dataset

This is a dynamic detailed human dataset, including variously posed 3D detailed human meshes with consistent topol-

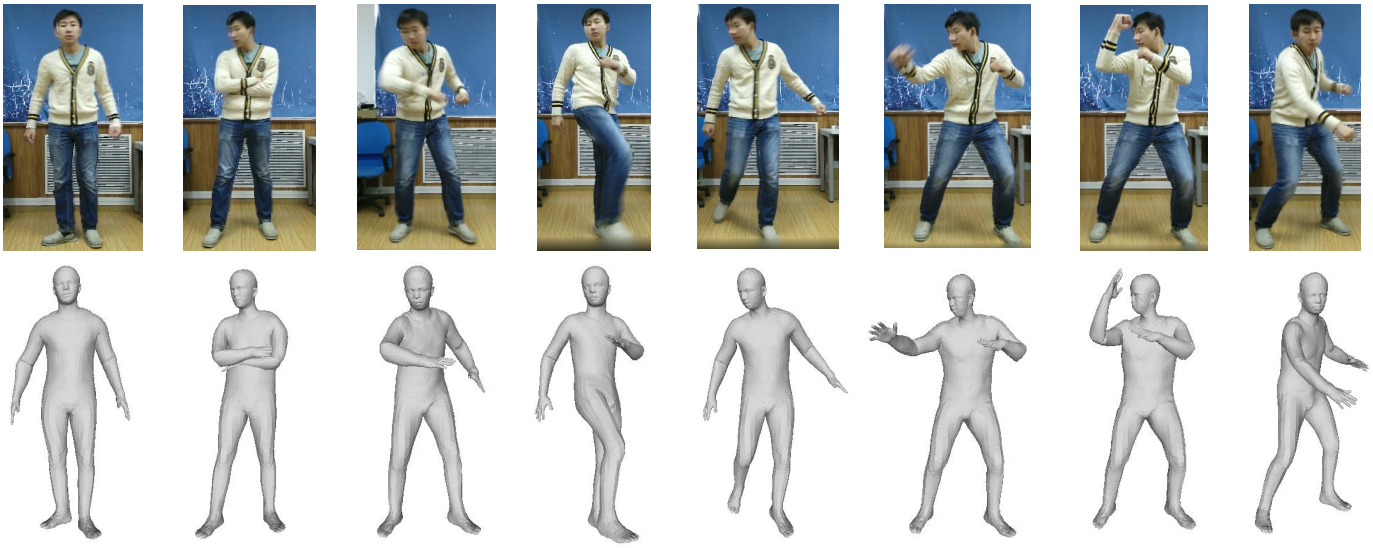

Figure 4. Our reconstruction results on a video input.

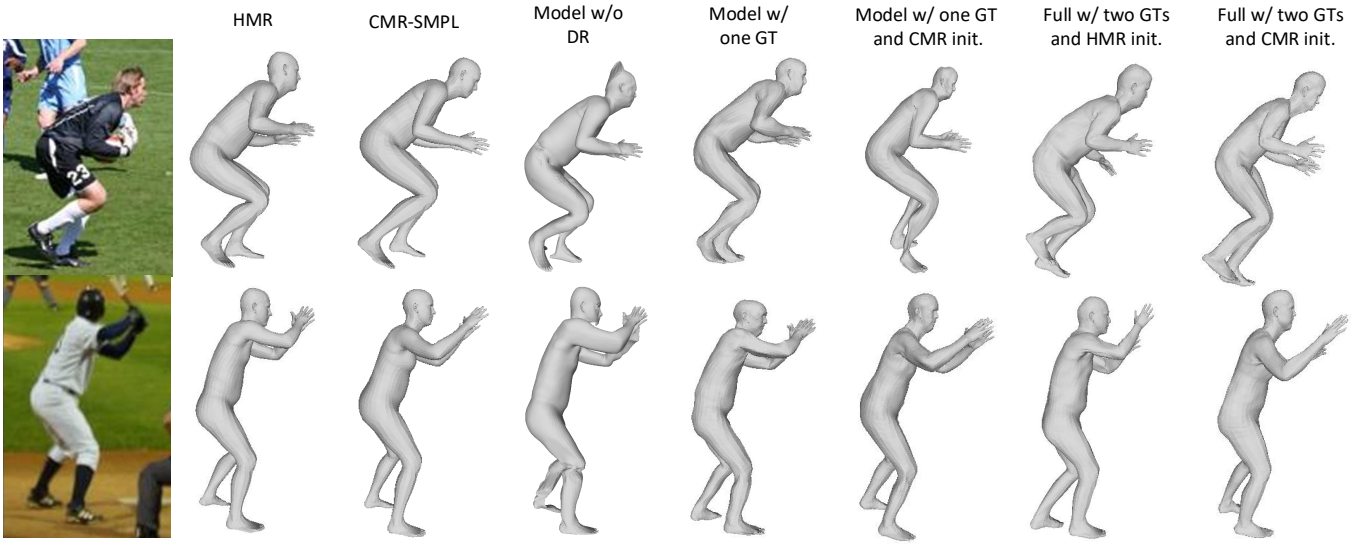

Figure 5. Reconstruction results using different variants of our method on LSP dataset [1]. From left to right are the input image, SMPL results of HMR [9] and CMR [17], and the results of five variants.

ogy and corresponding color images and SMPL models. We capture 20 subjects performing different motions with a Kinect v2.0 camera and obtain a 3D detailed human mesh, a color image and a SMPL model for each frame in real time by the DoubleFusion technique [55]. However, the detailed meshes at different time instances are topologically inconsistent. Therefore, we non-rigidly register the SMPL model against each detailed mesh by ray tracing along the normal of each vertex because the SMPL is inside the detailed mesh. The original SMPL model with 6890 vertices and 13776 faces has limited description ability for fine details, and hence we up-sample the body part of the SMPL model into a new mesh with 19019 vertices and 38034 faces, except for face, ears, hands and feet to ensure the density of main body. Figure 3(a) gives two examples of our registered meshes.

B. Scanned Dataset

We also synthesize a dataset from hundreds of 3D scanned human models[1]. In order to get more details for the face and hands, we use a newest parametric human model, SMPL-X [30], with 10475 vertices to fit the scanned models. To better capture the fine details, we up-sample the original SMPL-X model into a new mesh with 22228 vertices, and use this to non-rigidly register the scanned models. Specifically, we first render each scanned model from 32 viewpoints and find 2D keypoints of body, face and hands using OpenPose [56]. Then, we find 3D landmarks by minimizing the 2D re-projection error to the detected 2D keypoints, and optimize the pose and shape parameters of the new SMPL-X model. Finally, we adopt ED graph-based non-rigid deformation and per-vertex

[1]https://web.twindom.com

Figure 6. Reconstruction results on D^2Human dataset shown at two different viewpoints by DeepHuman [2], PIFu [14], PIFuHD [15], CMR [17], HMD [16], and our method.

refinement [57] to obtain a fitted mesh with fine details. We obtain the fitted mesh with 19019 vertices by finding the mapping between the two T-pose meshes using barycentric coordinate transformation. Figure 3(b) shows two examples of our registered meshes.

V. Experimental Results

In this section, we first demonstrate that our method is able to reconstruct human models with various poses in different scenes in Section V-A, and perform an ablation study to

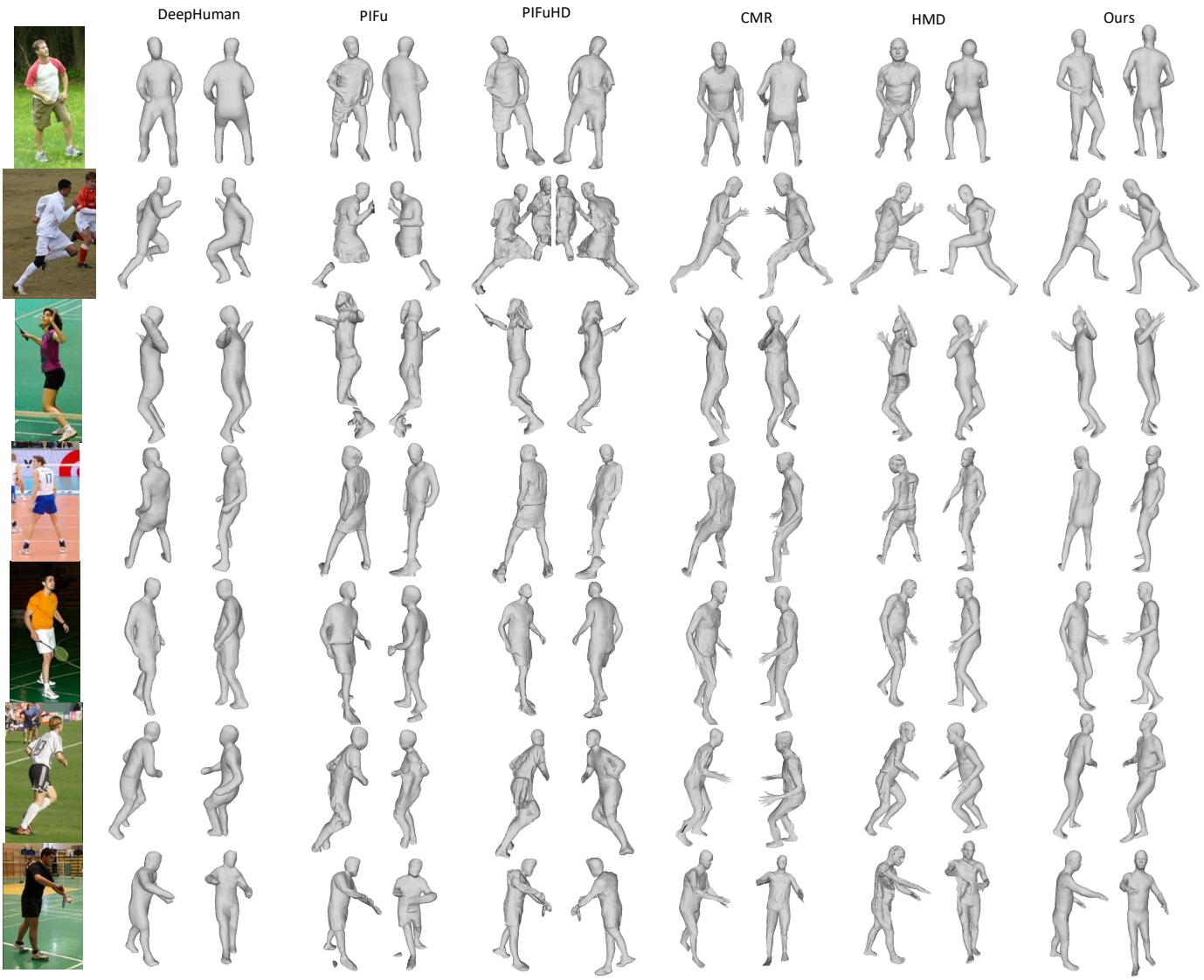

Figure 7. Reconstruction results on LSP dataset [1] shown at two different viewpoints by DeepHuman [2], PIFU [14], PIFuHD [15], CMR [17], HMD [16], and our method.

analyze the effects of different components of our approach in Section V-B. Then, we compare our method with several state-of-the-art methods quantitatively and qualitatively in Section V-C. Finally, we present detail enhancement results by a normal map based optimization and discuss the failure cases of our method in Section V-D.

A. Results

We demonstrate the robustness and generalization capability of our approach in Figure 1, using our model trained on our D^2Human dataset. Note that the subjects in the test images are not included in the training set. It can be seen that our method achieves realistic 3D reconstruction from a single image with various poses. We also show our reconstruction results for a video input of an unseen subject in Figure 4. The results are generated by using our method on each frame without any temporal smoothing. Thanks to the topology-consistent output, the dynamic reconstructed models are prevented from motion jitters.

B. Ablation Study

We train several variant models to prove our hypotheses and validate the effect of our improvements.

The Model without Deformation Representation (Model w/o DR). The model is trained with a vertex coordinate based representation, to assess the importance of deformation representation.

The Model with One Graph Transformation Block (Model w/ one GT). The model is trained using one graph transformation block with deformation representation, which is designed to assess the effectiveness of progressive graph transformation.

The Model with One Graph Transformation Block and CMR Initialization (Model w/ one GT and CMR init.). The model is trained using one graph transformation block with deformation representation but initialized by CMR [17], which aims to evaluate the effectiveness of progressive graph transformation and the influence of initialization.

Full Model with HMR Initialization (Full w/ two GTs and HMR init.). Our full model uses two graph transformation

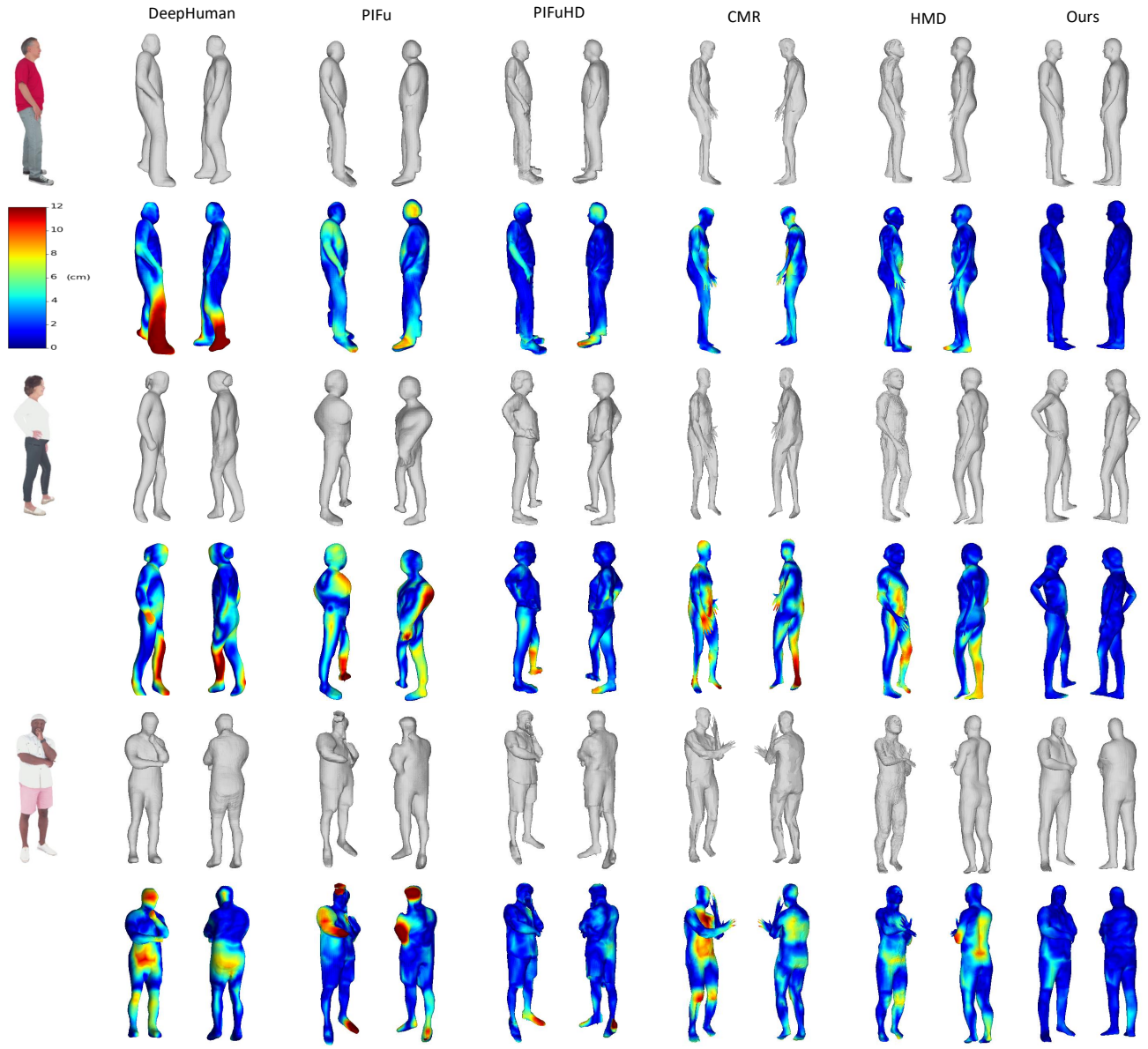

Figure 8. Reconstruction results on the scanned dataset shown at two different viewpoints by DeepHuman [2], PIFU [14], PIFuHD [15], CMR [17], HMD [16], and our method.

blocks with deformation representation, and this model is initialized by HMR [9].

Full Model with CMR Initialization (Full w/ two GTs and CMR init.). Our full model that uses two graph transformation blocks with deformation representation, and this model is initialized by CMR [17], which is designed to assess the influence of initialization.

1) Deformation Representation: To demonstrate the effectiveness of deformation representation, we compare our method with a baseline method that directly uses 3D coordinates of vertices as input. Table I gives quantitative evaluation results. We test these two methods on the test set of our D^2Human dataset, and calculate the mean distance with standard deviation between the estimated models and its corresponding ground truths. The models are first aligned by iterative closest point (ICP) and then the errors are computed using the standard Metro tool [58]. Metro is a popular tool designed to evaluate the difference between two triangular meshes. It adopts an approximated approach based on surface sampling and point-to-surface distance computation. Because our source and target are two surface meshes instead of point sets, it is more suitable to use this tool for quantitative evaluation. As shown in the table, our method with deformation representation achieves more accurate reconstruction. Figure 5 shows some visual results on the LSP dataset [1]. It can be seen that the results without deformation representation (Model w/o DR) are unstable and prone to artifacts, such as the head in the first row and the legs in the second row. Our model (the seventh column) with deformation representation achieves better results.

2) Progressive Graph Transformations: To demonstrate the effectiveness of our progressive graph transformation blocks, we compare our method using two graph transformation blocks with a variant using one graph transformation block.

Table II
QUANTITATIVE EVALUATION AND RUNNING TIMES OF DIFFERENT METHODS.

Mean±Std. (*cm*) Method / Dataset	DeepHuman [2]	PIFu [14]	PIFuHD [15]	CMR [17]	HMD [16]	Ours
D^2Human	3.844±0.366	5.113±1.214	3.391±0.405	4.990±1.717	4.973±1.742	1.372±0.464
Scanned Dataset	3.434±0.446	3.308±0.641	1.916±0.347	3.844±0.818	2.897±0.593	1.071±0.166
BUFF	3.500±0.648	4.440±2.252	1.425±0.602	3.416±0.724	3.194±0.993	2.627±0.515
Average Time (*s*)	196.496	57.671	16.500	7.563	24.288	1.300

Table I
QUANTITATIVE EVALUATION FOR DIFFERENT VARIANTS OF OUR METHOD (CM).

Method	Mean	Std.
Model w/o DR	1.841	0.444
Model w/ one GT	1.751	0.362
Full w/ two GTs and HMR init.)	**1.441**	**0.253**

Mean and Std.: The mean and standard deviation on the test set.

Quantitative evaluation and qualitative evaluation results are shown in Table I and Figure 5, respectively. Our progressive design with two multi-scale graph transformation blocks has smaller mean and standard deviation, compared with using one graph transformation block, which demonstrates that multi-scale design with two graph transformation blocks is more stable and is able to approach the optimal solution in a progressive manner. The last four columns of Figure 5 show that the models with two graph transformation blocks estimate more accurate poses and avoid some artifacts. This verifies the effectiveness of our network architecture.

3) Initialization: We also evaluate the influence of different inputs initialized by different methods. As shown in the last four columns of Figure 5, the performance of our method is not very sensitive to the initialization. Therefore, we use the SMPL result of HMR [9] as our initialization for comparison with other methods.

C. Comparisons

We compare our method with five state-of-the-art methods: DeepHuman [2], PIFu [14], PIFuHD [15], CMR [17] and HMD [16]. We use $750 \times 750 \times 750$ resolution for the PIFu method. We do not compare our method with Tex2Shape [13] because it requires a frontal image and recovers only an A-pose model for the images not included in their training set. For simplicity, we also ignore comparisons with some work that have already been compared like BodyNet [38] and SiCloPe [40]. Figure 6 shows the visual comparison results on our D^2Human dataset. The models are first aligned by iterative closest point (ICP) and then the errors are computed using the standard Metro tool [58]. The errors between the reconstructed models and the ground-truths are color-coded on the reconstructed models for visual inspection. Our results have the smallest reconstruction errors. Figure 7 shows the results on LSP dataset [1] that includes many challenging poses. Due to the lack of the ground-truth 3D meshes, our model is trained on our D^2Human dataset. As shown in the figure, DeepHuman [2] reconstructs human bodies with reasonable poses even for tough cases with occlusions and/or large deformations. However, the reconstructed models lack fine geometry details, *e.g.*, with faces and hands over-smoothed. PIFu [14], PIFuHD [15] and HMD [16] provide rich details at the image viewpoint, but their results at unseen viewpoints are still over-smooth or sometimes wrong. As shown in the last four rows of Figure 6, PIFu and PIFuHD cannot handle complex poses. CMR [17] is able to recover complete 3D models, but lacks realistic geometry details and has some artifacts, *e.g.*, the legs in the second and sixth rows of Figure 7. Our method is able to reconstruct complete topology-consistent 3D bodies with reasonable poses and geometries for all the test cases, which demonstrates the effectiveness of our proposed method. Figure 8 shows the results on our scanned dataset. We use 492 subjects for training and 7 subjects for testing. Training images are obtained by rendering each scanned model from 32 viewpoints. For quantitative evaluation, the models are also first aligned by iterative closest point (ICP) and then the errors are computed using the standard Metro tool [58]. The errors between the reconstructed models and the ground truths are color-coded on the reconstructed models for visual inspection. Our method also achieves the most accurate reconstruction. Table II gives the quantitative evaluation results on three datasets together with the average running time comparison. Our results have the smallest average reconstruction errors on D^2Human and scanned datasets. On BUFF dataset [59], our method achieves the second smallest mean error and the smallest standard deviation. In the running time comparison, all the experiments are run on a desktop with a GTX Titan X GPU and an i7-6800k CPU, and we remove the startup time of the model for fair comparison. Our method is the fastest.

D. Discussion

Our method aims to deal with large poses and generate the human shape consistent with the input image. For the scenarios that require fine geometry with the same topology, our estimated meshes can be optimized using an estimated normal map to obtain more geometry details. Define the estimated 3D human mesh as a set of vertices and edges $\mathcal{M}' = (\mathbf{V}', \mathcal{N})$, where $\mathbf{V}' \in \mathbb{R}^{n\times 3}$. The normal map $\hat{N}_m$ is estimated from the input image by PIFuHD [15] as the supervision to optimize $\mathbf{V}'$. Due to the complexity of the hands and face, we ignore the vertices near the hands and face in the optimization, which are recorded as $\mathbf{V}'_{\mathbf{F}}$. The overall function to be minimized is defined as

$$L_{op} = L_{normal} + \lambda L_{smooth} + \beta L_{position}, \qquad (4)$$

where L_{normal} is the normal term to measure the difference between the rendered normal map and the normal map estimated from the input image, L_{smooth} is the smoothness term to measure the smoothness of local deformation, and $L_{position}$ is the position term to measure the vertex displacement before and after optimization. λ and β are the weights to balance these terms. The normal term is defined as

$$L_{normal} = ||M_F \cdot (\pi(\mathbf{V}', \mathcal{N}) - \hat{N}_m)||_2^2, \tag{5}$$

where π is a differentiable renderer for normal mapping and M_F is a binary mask generated by rendering $\mathbf{V}' \backslash \mathbf{V}'_\mathbf{F}$. The smoothness term L_{smooth} is defined as

$$L_{smooth} = \sum_{p'_i \in \mathbf{V}' \backslash \mathbf{V}'_\mathbf{F}} || \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} (p'_i - p'_j) ||_2^2, \tag{6}$$

where p'_i represents the i-th vertex after optimization, $\mathcal{N}_i$ represents the 1-ring neighbors of p'_i. The position term $L_{position}$ is defined as

$$L_{position} = \sum_{p'_i \in \mathbf{V}' \backslash \mathbf{V}'_\mathbf{F}} ||p'_i - \hat{p}'_i||_2^2, \tag{7}$$

where $\hat{p}'_i$ is the p'_i before optimization. The optimization problem is solved by L-BFGS [60]. Figure 9 gives an example of geometry optimization which helps recover better surface details.

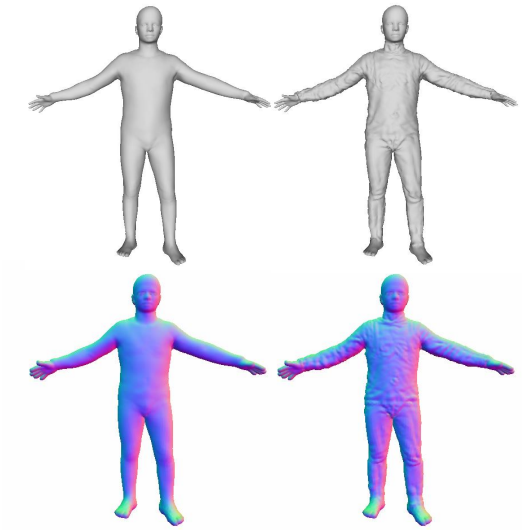

Figure 9. Geometry optimization result. Left: original mesh and normal map; Right: optimized mesh and normal map.

Figure 10 shows some failure cases using our method. For the pose with hands on the waist, our method fails to reconstruct correct poses of hands. When the hand or arm is close to the body, self-intersection sometimes happens. In future work, we will add other losses, *e.g.*, collision loss, to avoid self-intersection. Besides, we will expand and refine the flexibility and expressiveness of deformation representation.

VI. Conclusion

In this paper, we propose cascaded multi-scale graph transformation networks for detailed 3D human reconstruction from a single color image, which can reasonably infer the unseen body regions for a wide range of poses. Instead of directly using 3D coordinates, we encode a 3D mesh with effective deformation representation, which well deals with large deformations and avoids distorted geometries at extreme pose or viewpoint. We also design a perspective projection layer to incorporate appearance features into 9D deformation representation, which avoids fake geometries and make our model applicable to real images. Besides, we present the D^2Human (Dynamic Detailed Human) dataset, useful for training and evaluation of deep learning frameworks. Experimental results demonstrate that our method achieves more reasonable 3D human reconstruction from a single image, compared with several state-of-the-art methods.

References

[1] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation." in *Proc. British Machine Vision Conference*, 2010.

[2] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, "Deephuman: 3D human reconstruction from a single image," in *Proc. IEEE International Conference on Computer Vision*, 2019.

[3] W. Huang, X. Cao, K. Lu, Q. Dai, and A. C. Bovik, "Toward naturalistic 2D-to-3D conversion," *IEEE Transactions on Image Processing*, vol. 24, no. 2, pp. 724–733, 2015.

[4] F. Xu and Q. Dai, "Occlusion-aware motion layer extraction under large interframe motions," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2615–2626, 2011.

[5] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black, "ClothCap: Seamless 4D clothing capture and retargeting," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–15, 2017.

[6] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proc. European Conference on Computer Vision*, 2016.

[7] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3D and 2D human representations," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[8] A. Zanfir, E. Marinoiu, and C. Sminchisescu, "Monocular 3D pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[9] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[10] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3D human pose and shape via model-fitting in the loop," in *Proc. IEEE International Conference on Computer Vision*, 2019.

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

[12] A. S. Jackson, C. Manafas, and G. Tzimiropoulos, "3D human body reconstruction from a single image via volumetric regression," in *Proc. European Conference on Computer Vision*, 2018.

[13] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor, "Tex2shape: Detailed full human body geometry from a single image," in *Proc. IEEE International Conference on Computer Vision*, 2019.

[14] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "PIFU: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proc. IEEE International Conference on Computer Vision*, 2019.

[15] S. Saito, T. Simon, J. Saragih, and H. Joo, "Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[16] H. Zhu, X. Zuo, S. Wang, X. Cao, and R. Yang, "Detailed human shape estimation from a single image by hierarchical mesh deformation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

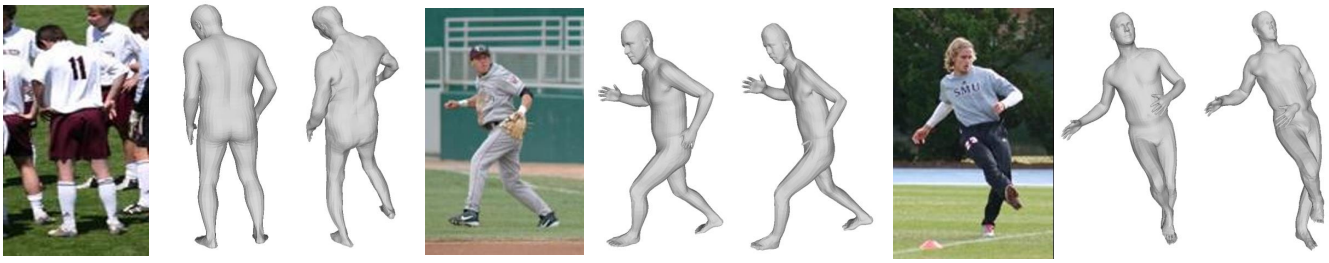

Figure 10. Examples of failure cases using our method.

[17] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[18] L. Gao, Y.-K. Lai, J. Yang, Z. Ling-Xiao, S. Xia, and L. Kobbelt, "Sparse data driven mesh deformation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 3, pp. 2085–2100, 2019.

[19] Q. Tan, L. Gao, Y.-K. Lai, J. Yang, and S. Xia, "Mesh-based autoencoders for localized deformation component analysis," in *Proc. AAAI Conference on Artificial Intelligence*, 2018.

[20] L. Gao, J. Yang, Y.-L. Qiao, Y.-K. Lai, P. L. Rosin, W. Xu, and S. Xia, "Automatic unpaired shape deformation transfer," *ACM Transactions on Graphics*, vol. 37, no. 6, pp. 1–15, 2018.

[21] A. Gilbert, M. Volino, J. Collomosse, and A. Hilton, "Volumetric performance capture from minimal camera viewpoints," in *Proc. European Conference on Computer Vision*, 2018.

[22] J. Liang and M. C. Lin, "Shape-aware human pose and shape reconstruction using multi-view images," in *Proc. IEEE International Conference on Computer Vision*, 2019.

[23] Z. Huang, T. Li, W. Chen, Y. Zhao, J. Xing, C. LeGendre, L. Luo, C. Ma, and H. Li, "Deep volumetric video from very sparse multi-view performance capture," in *Proc. European Conference on Computer Vision*, 2018.

[24] M. Dou, P. Davidson, S. R. Fanello, S. Khamis, A. Kowdle, C. Rhemann, V. Tankovich, and S. Izadi, "Motion2fusion: Real-time volumetric performance capture," *ACM Transactions on Graphics*, vol. 36, no. 6, p. 246, 2017.

[25] Y. Liu, J. Gall, C. Stoll, Q. Dai, H.-P. Seidel, and C. Theobalt, "Markerless motion capture of multiple characters using multiview image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2720–2735, 2013.

[26] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video based reconstruction of 3D people models," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[27] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt, "Livecap: Real-time human performance capture from monocular video," *ACM Transactions on Graphics*, vol. 38, no. 2, pp. 1–17, 2019.

[28] Y. Sun, Y. Ye, W. Liu, W. Gao, Y. Fu, and T. Mei, "Human mesh recovery from monocular images via a skeleton-disentangled representation," in *Proc. IEEE International Conference on Computer Vision*, 2019.

[29] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "SCAPE: shape completion and animation of people," in *Proc. ACM Special Interest Group for Computer Graphics and Interactive Techniques*, 2005.

[30] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[31] P. Guan, A. Weiss, A. O. Balan, and M. J. Black, "Estimating human shape and pose from a single image," in *Proc. IEEE International Conference on Computer Vision*, 2009.

[32] E. Dibra, H. Jain, C. Öztireli, R. Ziegler, and M. Gross, "HS-Nets: Estimating human body shape from silhouettes with convolutional neural networks," in *Proc. IEEE International Conference on 3D Vision*, 2016.

[33] V. Tan, I. Budvytis, and R. Cipolla, "Indirect deep structured learning for 3D human body shape and pose prediction," in *Proc. British Machine Vision Conference*, 2017.

[34] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, "Neural body fitting: Unifying deep learning and model based human pose and shape estimation," in *Proc. IEEE International Conference on 3D Vision*, 2018.

[35] N. Rüegg, C. Lassner, M. J. Black, and K. Schindler, "Chained representation cycling: Learning to estimate 3D human pose and shape by cycling between representations," in *Proc. AAAI Conference on Artificial Intelligence*, 2020.

[36] Y. Rong, T. Shiratori, and H. Joo, "Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration," *arXiv preprint arXiv:2008.08324*, 2020.

[37] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black, "Monocular expressive body regression through body-driven attention," in *Proc. European Conference on Computer Vision*, 2020.

[38] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid, "Bodynet: Volumetric inference of 3D human body shapes," in *Proc. European Conference on Computer Vision*, 2018.

[39] V. Gabeur, J.-S. Franco, X. Martin, C. Schmid, and G. Rogez, "Moulding humans: Non-parametric 3D human shape estimation from single images," in *Proc. IEEE International Conference on Computer Vision*, 2019.

[40] R. Natsume, S. Saito, Z. Huang, W. Chen, C. Ma, H. Li, and S. Morishima, "Siclope: Silhouette-based clothed people," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[41] G. Moon and K. M. Lee, "I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image," *arXiv preprint arXiv:2008.03713*, 2020.

[42] L. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International Journal of Computer Vision*, vol. 87, no. 1-2, p. 4, 2010.

[43] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[44] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3D human pose estimation in the wild using improved cnn supervision," in *Proc. IEEE International Conference on 3D Vision*, 2017.

[45] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[46] A. Pumarola, J. Sanchez, G. Choi, A. Sanfeliu, and F. Moreno-Noguer, "3DPeople: Modeling the Geometry of Dressed Humans," in *Proc. IEEE International Conference on Computer Vision*, 2019.

[47] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll, "Detailed, accurate, human shape estimation from clothed 3D scan sequences," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[48] Q. Ma, J. Yang, A. Ranjan, S. Pujades, G. Pons-Moll, S. Tang, and M. J. Black, "Learning to dress 3d people in generative clothing," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[50] N. Verma, E. Boyer, and J. Verbeek, "FeaStNet: Feature-Steered graph convolutions for 3D shape analysis," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[51] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3D faces using convolutional mesh autoencoders," in *Proc. European Conference on Computer Vision*, 2018.

[52] M. Garland and P. S. Heckbert, "Surface simplification using quadric

error metrics," in *Proc. Annual Conference on Computer Graphics and Interactive Techniques*, 1997.
[53] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. International Conference on Machine Learning*, 2015.
[54] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. International Conference on Machine Learning*, 2013.
[55] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu, "Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
[56] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
[57] X. Chen, A. Pang, Y. Zhu, Y. Li, X. Luo, G. Zhang, P. Wang, Y. Zhang, S. Li, and J. Yu, "Towards 3D human shape recovery under clothing," *arXiv preprint arXiv:1904.02601*, 2019.
[58] P. Cignoni, C. Rocchini, and R. Scopigno, "Metro: Measuring error on simplified surfaces," in *Comput. Graph. Forum*, 1998.
[59] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll, "Detailed, accurate, human shape estimation from clothed 3D scan sequences," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
[60] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, no. 1, pp. 503–528, 1989.

Kun Li received the B.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2006, and the master and Ph.D. degrees from Tsinghua University, Beijing, in 2011. She visited École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, in 2012 and from 2014 to 2015. She is currently an Associate Professor with the College of Intelligence and Computing, Tianjin University, Tianjin, China. Her research interests include dynamic scene 3D reconstruction and image/video processing.

Hao Wen received the B.E. degree from the Tianjin University of Mathematics, Tianjin, China in 2017. He is currently pursuing the Ph.D. degree with the College of Intelligence and Computing in Tianjin University, Tianjin, China. His research interests include computer vision and human 3D reconstruction.

Qiao Feng is currently an undergraduate and will pursue the master degree with the College of Intelligence and Computing, Tianjin University, Tianjin, China. His research interests include machine learning and computer graphics.

Yuxiang Zhang is currently pursuing the Ph.D. degree with Department of Automation, Tsinghua University, Beijing, China. His research interests include computer vision and computer graphics.

Xiongzheng Li received the B.E. degree from East China Jiaotong University, Jiangxi Province, China, in 2019. He is currently pursuing the Ph.D. degree with the College of Intelligence and Computing in Tianjin University, Tianjin, China. His research interests include 3D vision and computer graphics.

Jing Huang is currently an undergraduate and will pursue the master degree with the College of Intelligence and Computing, Tianjin University, Tianjin, China. His research interests include computer vision and computer graphics.

Cunkuan Yuan receiveed the B.E degree from Jiangsu University of Science and Technology, Zhenjiang, Chinain 2017, and the master degree from Tianjin University, Tianjin, China, in 2020. He is currently a graphics engineer in Beijing Kuaishou Technology Co., Ltd, Beijing, China. His research interests include high-performance deep learning inference, computer graphics and image processing.

Yu-Kun Lai received his bachelor and Ph.D. degrees in computer science from Tsinghua University in 2003 and 2008, respectively. He is currently a professor in the School of Computer Science & Informatics, Cardiff University. His research interests include computer graphics, geometry processing, image processing and computer vision. He is on the editorial boards of Computer Graphics Forum and The Visual Computer.

Yebin Liu received the BE degree from Beijing University of Posts and Telecommunications, China, in 2002, and the PhD degree from the Department of Automation, Tsinghua University, Beijing, China, in 2009. He has been working as a research fellow at the computer graphics group of the Max Planck Institute for Informatik, Germany, in 2010. He is currently an associate professor in the Department of Automation, Tsinghua University. His research areas include computer vision, computer graphics and computational photography.