HEI-Human: A Hybrid Explicit and Implicit Method for Single-View 3D Clothed Human Reconstruction

Leyuan Liu[1,2], Jianchi Sun[1], Yunqi Gao[1], and Jingying Chen[1,2(✉)]

[1] National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China
{lyliu,chenjy}@ccnu.edu.cn, {sunjc0306,gaoyunqi}@mails.ccnu.edu.cn
[2] National Engineering Laboratory for Educational Big Data, Central China Normal University, Wuhan 430079, China

Abstract. Single-view 3D clothed human reconstruction is a challenging task, not only because of the need to infer the complex global topology of human body but also due to the requirement to recover delicate surface details. In this paper, a method named HEI-Human is proposed to hybridize an explicit model and an implicit model for 3D clothed human reconstruction. In the explicit model, the SMPL model is voxelized and then integrated into a 3D hourglass network to supervise the global geometric aligned features extraction. In the implicit model, 2D aligned features are first extracted by a 2D hourglass network, and then an implicit surface function is employed to construct the occupancy field of human body using the hybrid 2D and 3D aligned features. As the explicit model and implicit model are mutually beneficial, our HEI-Human method not only generates reconstructions with plausible global topology but also recovers rich and accurate surface details. The HEI-Human is evaluated on the current largest publicly available dataset, and the experimental results demonstrate that our method outperforms the state-of-the-art methods including DeepHuman, PIFu, and GeoPIFu.

Keywords: 3D human reconstruction · Hybrid model · Single-view

1 Introduction

3D human reconstruction has attracted more and more attention in the field of computer vision and computer graphics, since it has a wide range of potential applications such as virtual dressing [1] and game design [2]. To obtain accurate reconstructed 3D human models, conventional methods usually employ multi-view images [3] or stereo imaging sensors [4]. Recently, benefiting from the

This work was supported by the National Natural Science Foundation of China (62077026, 61937001), the Fundamental Research Funds for the Central Universities (CCNU19ZN004), and the Research Project of Graduate Teaching Reform of CCNU (2019JG01).

H. Ma et al. (Eds.): PRCV 2021, LNCS 13020, pp. 251–262, 2021.
https://doi.org/10.1007/978-3-030-88007-1_21

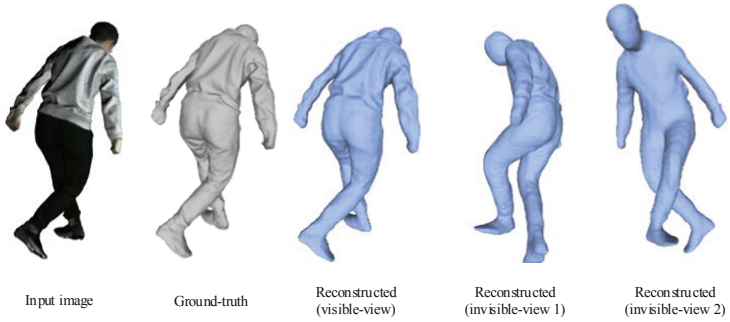

Fig. 1. 3D clothed human reconstruction results by our HEI-human method. Our method not only generates 3D human models with plausible global topology but also recovers rich and accurate surface details.

development of deep neural networks, promising progress [5–7] has been made in reconstructing 3D human models from a single RGB image.

Technically, single-view 3D human reconstruction is a very challenging task. To reconstruct satisfactory 3D human models, algorithms not only need to infer the plausible 3D global topology of the articulated human body from a monocular 2D image that lacks depth information but also are required to recover delicate surface details such as hairs and clothes wrinkles. Recent researches have tried to address these two challenges using different methods, which can be roughly divided into two categories: explicit methods and implicit methods. Explicit methods such as DeepHuman [5] and Tex2Shape [15] usually represent human body by a parametric model (e.g. SMPL [8]) and infer the 3D reconstructions explicitly. Due to the geometrical prior provided by the parametric model, explicit methods can infer the plausible 3D global topology of human body. However, constrained by the low resolution of the parametric model, such explicit methods often have difficulty in recovering surface details. Implicit methods like PIFu [6] and Geo-PIFu [7] employ implicit surface functions to estimate dense occupancy fields for reconstructing 3D human meshes. Benefiting from the dense sampling and interpolation strategies in the feature spaces, implicit methods are able to recover rather richer surface details than explicit methods. However, lacking of the guidance of global information, implicit methods tend to produce unreasonable artifacts on the reconstructed models.

To infer the reasonable global topology of 3D human body from a single RGB image while recovering rich surface details, in this paper, a method named HEI-Human is proposed to hybridize an explicit model and an implicit model for 3D clothed human reconstruction. In our explicit model, the SMPL model [8] is voxelized and then integrated into a 3D hourglass network [9] to supervise the 3D geometric aligned features extraction. In our implicit model, the 2D aligned features are first extracted by a 2D hourglass network and then an implicit surface function is employed to construct the occupancy field of the reconstructed

human body using the hybrid 2D and 3D aligned features queried from interpolation feature spaces. The explicit model and implicit model in our HEI-human method are mutually beneficial. On one hand, as the implicit model is supervised by the topology prior provided by the explicit model, our method rarely produces unreasonable artifacts; On the other hand, the implicit model not only guarantees the recovery of surface details but also helps the explicit model to more accurate geometric aligned features by producing detailed reconstructions. As a result, our HEI-Human method not only generates reconstructions with plausible global topology but also recovers abundant and accurate surface details (As illustrated in Fig. 1). Experimental results on the DeepHuman dataset [5] demonstrates that our HEI-Human method achieves the state-of-the-art performance.

In summary, the main contributions of this paper are two-fold:

(1) We propose a framework to hybridize an explicit model and an implicit model for reconstructing 3D clothed human from a single RGB image. By squeezing the advantages of both the explicit model and the implicit model, this hybrid framework has the ability to generate 3D human reconstructions with plausible global topology and rich surface details.

(2) We design deep neural networks that mainly consist of 2D/3D hourglass structures to implement the hybrid framework for 3D clothed human reconstruction. Although very simple loss functions are used, our HEI-human method outperforms the current state-of-the-art methods including DeepHuman [5], PIFu [6], and GeoPIFu [7], which employ much more delicate and complex losses.

2 Related Work

Parametric Model Based 3D Human Reconstruction. A parametric model is an explicit model with three main representations: voxel, point cloud or mesh [10]. A parametric human model is a statistic template trained from many human models, which can be used to drive arbitrary human bodies with a limited number of parameters. The parametric human model allows for supervision and normalization during the 3D human reconstruction process, preventing the reconstruction results from varying significantly in comparison with the human model. Most parametric-based human reconstruction methods including HMR [11], SPIN [12], DaNet [13], and GCMR [14] focus on human shape and pose estimation. DeepHuman [5] uses the SMPL model to constrain the degrees of freedom in the output space. After obtaining the voxel occupation field, the surface normals and the computed depth values are employed to refine the details of the model surface, but this method has little effect due to the storage limitation. Tex2Shape [15] considers that the small resolution of the SMPL model can affect the details of the reconstruction, and adds normal maps and vector displacement maps to the SMPL model to enhance the details of the reconstruction. HMR uses the pose and shape parameters of the SMPL model to transform the 3D reconstruction problem into a parametric regression problem of the model but without surface reconstruction. What's more, model-based reconstruction

can also lead to 3D reconstruction failure when the SMPL model is registering incorrectly [7].

Non-parametric Based 3D Human Reconstruction. Non-parametric human reconstruction does not require the 3D model as a priori hypothesis to obtain the 3D human model directly from a single RGB image. After the emergence of the 3D reconstruction task, some methods (such as Bodynet [16], TetraTSDF [17] and VRN [18]) reconstruct the 3D geometry of the human body via volumetric regression. Volumetric regression is limited by the resolution, which is a huge challenge for both network and memory of high-resolution volumetric regression. Secondly, volumetric regression ignores the details of the human surface. For solving these problems, the implicit function is taken into account for the 3D reconstruction task. In the field of 3D human reconstruction, the main work on implicit functions are PIFu [6], PIFuHD [19], ARCH [20], Geo-PIFu [7], SiCloPe [21] etc. The sampling strategy of the implicit surface function is an issue worth investigating. Dense and sparse sampling have a significant impact on the reconstruction results. The optimal sampling parameters by comparison experiments are given in PIFu. PIFuHD feeds higher resolution color images to obtain high quality reconstructions. ARCH proposes an opacity-aware distinguishable rendering in generating datasets to improve the implicit function representations in arbitrary poses. Geo-PIFu converts image 2D features to 3D latent features in feature extraction to constrain spatial degrees of freedom. SiCloPe uses an implicit representation of 2D silhouettes to describe complex human body. However, the implicit surface function lacks constraints on the global features of the human body model, which can cause some errors.

3 Methodology

3.1 Overview

Our HEI-human method is implemented by deep neural networks. As illustrated in Fig. 2, the networks are divided into two main parts: the explicit part (upper) and the implicit part (bottom). Given an input image ($\mathcal{I}$), our method feeds it to both the explicit part and the implicit part. The explicit part starts from the parametric SMPL model estimation and voxelization, then extracts latent 3D aligned features from the voxelized SMPL model ($\tilde{\mathcal{S}}$) by a 3D Hourglass network based encoder ($\mathcal{V}$), and finally reconstructs a coarse result ($\hat{\mathcal{S}}$) by a 3D convolutional network based decoder ($\mathcal{D}$). The implicit part first extracts latent a 2D aligned features using a 2D Hourglass network ($\mathcal{G}$), then feeds both the latent 2D and 3D aligned features into an implicit surface function (f) implemented by a multi-layer perception (MLP) for constructing an occupancy field. The occupancy field is constructed in a voxel-by-voxel manner. For each voxel in the occupancy field, the latent 2D and 3D aligned features with respect to the voxel are queried and hybridised to compute the probability of occupancy. Finally, the 3D human mesh is reconstructed from the refined occupancy fields ($\widehat{\mathcal{S}}$) via Marching Cubes [22].

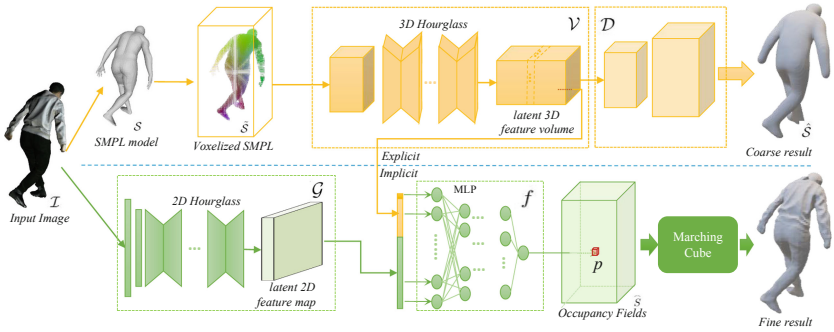

Fig. 2. Overview of our method. Our networks are divided into the explicit part (upper) and the implicit part (bottom). The explicit part extracts latent 3D aligned features from the voxelized SMPL model ($\tilde{\mathcal{S}}$) by a 3D Hourglass based encoder ($\mathcal{V}$). The implicit part first extracts the latent 2D aligned features using a 2D Hourglass network ($\mathcal{G}$), then hybridises both the latent 2D and 3D aligned features into an implicit surface function (f) to construct an occupancy field. Finally, the 3D human mesh is reconstructed from the occupancy filed via Marching Cubes.

3.2 Explicit Model

In our explicit model, a 3D human body is represented by a voxelized SMPL (Skinned Multi-Person Linear) [8]. The SMPL is a parametric model that represents a specific 3D human body by a shape vector α and a pose vector β:

$$\mathcal{T}(\alpha, \beta) = \overline{\mathcal{T}} + \mathcal{B}_s(\alpha) + \mathcal{B}_p(\beta) \tag{1}$$

$$\mathcal{S}(\alpha, \beta) = W(\mathcal{T}(\alpha, \beta), J(\alpha), \beta, \delta) \tag{2}$$

where $\overline{\mathcal{T}}$ is the mean model, $\mathcal{B}_s(\cdot)$ is a blend shape function, $\mathcal{B}_p(\cdot)$ is a pose-dependent blend shape function, $J(\cdot)$ is a joint prediction function, and $W(\cdot, \delta)$ is a skinning function with blend weights δ. In the public available SMPL model [8], $\overline{\mathcal{T}}$, $\mathcal{B}_s(\cdot)$, $\mathcal{B}_p(\cdot)$, $J(\cdot)$, and δ are given. The shape vector α and the pose vector β are estimated using HMR [11] and SMPLify [23]. For the sake of integrating the SMPL model into the deep learning network, the vertex-based SMPL model ($\mathcal{S}$) is voxelized into a voxel volume ($\tilde{\mathcal{S}}$):

$$\tilde{\mathcal{S}} = \mathcal{H}(\mathcal{S}) \tag{3}$$

where $\mathcal{H}(\cdot)$ denotes the voxelization algorithm [5]. As illustrated in Fig. 2, the voxelized SMPL model only generates a naked-like 3D human body with plausible pose and body shape to the input image.

A deep neural network with an "encoder-decoder" reconstruct is designed to reconstruct a clothed 3D human body ($\hat{\mathcal{S}}$) from the voxelized SMPL model ($\tilde{\mathcal{S}}$) explicitly:

$$\hat{\mathcal{S}} = \mathcal{D}_\varphi(\mathcal{V}_\mu(\tilde{\mathcal{S}})) \tag{4}$$

where $\mathcal{V}(\cdot)$ and $\mathcal{D}(\cdot)$ are respectively the encoder and decoder. The encoder $\mathcal{V}(\cdot)$, which is implemented by a 4-layer tandem 3D-Hourglass network with trainable weights μ, is utilized to extract latent 3D aligned features. The decoder $\mathcal{D}(\cdot)$ consists of two 3D convolutions with weights φ. It should note that only the latent 3D aligned features extracted by the encoder $\mathcal{V}$ are employed to infer the finally fine reconstruction by the implicit model, which will described in Subsect. 3.3. Although the decoder $\mathcal{D}$ also can generate a coarse result (i.e., $\hat{\mathcal{S}}$), it is only used for training the encoder.

3.3 Implicit Model

In our implicit model, the surface of a 3D human mesh is represented implicitly by the occupied/unoccupied decision boundary of a continuous occupancy field. An occupancy mapping function $f(\cdot)$ is employed to map each 3D point p in the occupancy field into an occupancy value o ($o \in [0, 1]$):

$$f(p) = o \tag{5}$$

An occupancy value $o > 0.5$ indicates point p is inside the mesh, while $o < 0.5$ means point p is outside the mesh. Thereby, the surface of a 3D human mesh is defined as a 0.5 level set of the continuous occupancy field.

Besides the input RGB image ($\mathcal{I}$), the latent feature volume ($\mathcal{V}$) produced by the explicit model is also utilized to learn the occupancy mapping function $f(\cdot)$. Thereupon, $f(\cdot)$ is formulated as:

$$f_\theta(\{\mathcal{G}_\omega(\mathcal{I}, \ddot{\mathcal{X}}(\pi(p))_m)_k\}_{k=1,\cdots K}^{m=1,\cdots M}, \{\mathcal{V}_\mu(\dddot{\mathcal{X}}(p)_n)_d\}_{d=1,\cdots D}^{n=1,\cdots N}, p_z) = o \tag{6}$$

where $\mathcal{G}(\mathcal{I}, \cdot)$ is a feature extraction function that generates latent feature maps from the input image $\mathcal{I}$, K and k are respectively the number of channels and the channel index of the feature map extracted by g, $\pi(p)$ represents the weak perspective transformation that projects the 3D query point p into the 2D feature map plane, D and d are the number of channels and the channel index of the latent feature volume ($\mathcal{V}$), $\ddot{\mathcal{X}}$ and $\dddot{\mathcal{X}}$ respectively denote the bi-linear and tri-linear interpolation functions, (M, m) and (N, n) are respectively the numbers and indexes of interpolations in the 2D and 3D feature spaces, and p_z represents the depth of p. Hence, our implicit occupancy mapping function with respect to the query point p totally has $(K \times M) + (D \times N) + 1$ input parameters, which fuse the 2D latent features extracted from the input image, the 3D latent features extracted by the explicit model, and the depth information of the query point.

As illustrated in Fig. 2, the functions $\mathcal{G}_\omega(\mathcal{I})$ and $f_\theta(\cdot)$ in Eq. (6) are respectively implemented by a 2D hourglass network with weights ω and a MLP with weight θ. Unlike the Geo-PIFu [7] that directly learns latent 2D and 3D aligned features from the input image, our method integrates the latent 3D aligned features learnt from the parametric SMPL model to regularize the implicit model. As a consequence, our method not only recovers rich surface details but also rarely produces unreasonable artifacts.

3.4 Loss Functions

The explicit part of our network is trained using an extended cross-entropy loss between the estimated voxel volume and ground-truth [24]:

$$\mathcal{L}_v(\ddot{\mathcal{S}}, \dot{\mathcal{S}}) = -\frac{1}{|\dot{\mathcal{S}}|} \sum_{x,y,z} \gamma \ddot{\mathcal{S}}_{x,y,z} \log \dot{\mathcal{S}}_{x,y,z} + (1-\gamma)(1-\ddot{\mathcal{S}}_{x,y,z})(1-\log \dot{\mathcal{S}}_{x,y,z}) \quad (7)$$

where $\ddot{\mathcal{S}}$ is the voxel volume voxelized from the ground-truth 3D human mesh, $\dot{\mathcal{S}} \in \{\hat{\mathcal{S}}, \widehat{\mathcal{S}}\}$ is the estimated (coarse or fine) voxel volume, (x, y, z) are the voxel indices for the width, height and depth axes, and γ is the weight to balance the losses of occupied/unoccupied voxels. The implicit part of our network is trained based on a set of query points randomly sampled from the occupancy field. The mean square error loss is adopted to measure the errors between the ground-truth and the predicted occupancy values:

$$\mathcal{L}_p = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} (o_p - \hat{o}_p)^2 \quad (8)$$

where $\mathcal{P}$ is the set of sampled query points for a training sample, $|\mathcal{P}|$ denotes the number of sampled points, and o_p and $\hat{o}_p$ are respectively the ground-truth and the predicted occupancy values of the query point p.

4 Experiments

4.1 Dataset and Protocol

Dataset. The proposed 3D human reconstruction method is extensively evaluated on the DeepHuman dataset [5], which contains 6,795 data items of 202 subjects with various body shapes, poses, and clothes. The raw data of DeepHuman is captured by consumer-grade RGB-D cameras, and then processed into 3D human data items by an improved DoubleFusion algorithm [4]. Consequently, each data item in DeepHuman consists of a 3D textured surface mesh, a RGB image, and an aliened SMPL model. In our experiments, the aliened SMPL models are used.

Protocols. In our experiments, our method and the other competing 3D human reconstruction methods are all trained on the 5,436 training data items in the DeepHuman dataset, and evaluated on the rest 1,359 testing data items.

4.2 Training Details

A two-stage scheme is used to train our HEI-human model. At the first stage, the explicit model is trained only with the $\mathcal{L}_v(\ddot{\mathcal{S}}, \hat{\mathcal{S}})$ loss for 10 epochs. At the second stage, the whole network (explicit model + implicit model) is trained for another

10 epochs using $\lambda\mathcal{L}_v(\ddot{\mathcal{S}}, \widehat{\mathcal{S}}) + (1-\lambda)\mathcal{L}_q$. The weight parameter γ in Eq. (7) is set to 0.7, the number of sampled query points for each training sample in Eq. (8) is set as 5,000, and the hyper-parameter λ is set to 0.5. For both the two training stages, the RMSprop is adopted as the optimizer, the batch size is fixed to 8, and the learning rate is set as 0.0001 at the beginning and decayed by a factor of 10 after every 4 epochs. All the networks in our method are implemented by PyTorch [25], and the whole training task takes about 7 days on a computer with two NVIDIA GeForce 1080Ti GPUs.

4.3 Quantitative Results

Metrics. The Chamfer distance (ε_{cd}), point-to-surface distance (ε_{psd}), cosine distance (ε_{cos}), and L2-norm (ε_{l2}) are adopted as the metrics for evaluating different 3D human reconstruction methods. The Chamfer distance and point-to-surface distance focus more on the overall quality of model topology, while the cosine distance and L2-norm tend to evaluate local surface details. For all these four metrics, smaller values indicate better performance.

Table 1. Quantitative comparisons on the DeepHuman dataset.

Methods	ε_{cd}	ε_{psd}	ε_{cos}	ε_{l2}
DeepHuman [5]	11.928	11.246	0.2088	0.4647
PIFu [6]	2.6004	4.0174	0.0949	0.3048
Geo-PIFu (coarse) [7]	2.2907	2.6260	0.0874	0.3175
Geo-PIFu [7]	1.7794	1.9548	0.0717	0.2649
HEI-Human (Ours)	**0.1742**	**0.2297**	**0.0661**	**0.2540**

Results and Comparisons. The proposed HEI-Human is compared against three code available state-of-the-art 3D human reconstruction methods: the DeepHuman [5], PIFu [6], and Geo-PIFu [7]. The quantitative results achieved by these methods are presented in Table 1. In terms of global topology quality, our method outperforms the second-best method Geo-PIFu [7] (an implicit method) by a Chamfer distance (ε_{cd}) of 1.6052 and by a point-to-surface distance (ε_{psd}) of 1.4251; In terms of the local details, our method also surpasses Geo-PIFu by a cosine distance of 0.0056 and by a L2-norm of 0.0109. Although our method outperforms all the competing methods on the four metrics, it should be note that it is even not a completely fair comparison. Our method is only trained for 20 (10+10) epochs with a small batch-size of 8 due to the restriction in experimental condition, while all the three competing methods are trained for more than 40 epochs with a larger batch-size (Geo-PIFu [7] is trained for 45 epochs with a batch-size of 36).

4.4 Qualitative Results

Figure 3 shows the qualitative reconstruction results produced by DeepHuman [5], PIFu [6], Geo-PIFu [7], and our HEI-Human on the DeepHuman dataset. It can been seen that DeepHuman recovers rather less local surface details than the other three methods and generates "fattened" and "naked-like" human bodies. Although PIFu and Geo-PIFu can generate surface details of clothes and plausible global topology from the visible view, we can find many unreasonable artifacts (marked in red circles) on their results from the invisible view. Benefiting from the combination of the explicit voxelized SMPL model and the implicit surface function representation, our method not only generates better global topology from the views of both the visible and invisible sides but also recovers richer surface details (such as hair and clothes winkles) than the competing methods. However, the resolution of the 3D human model in the THuman dataset is low, and the details of the face are blurred. In Fig. 3, the facial details of all methods are not be recovered. 3D face reconstruction methods [26,27] can be employed to solve this problem.

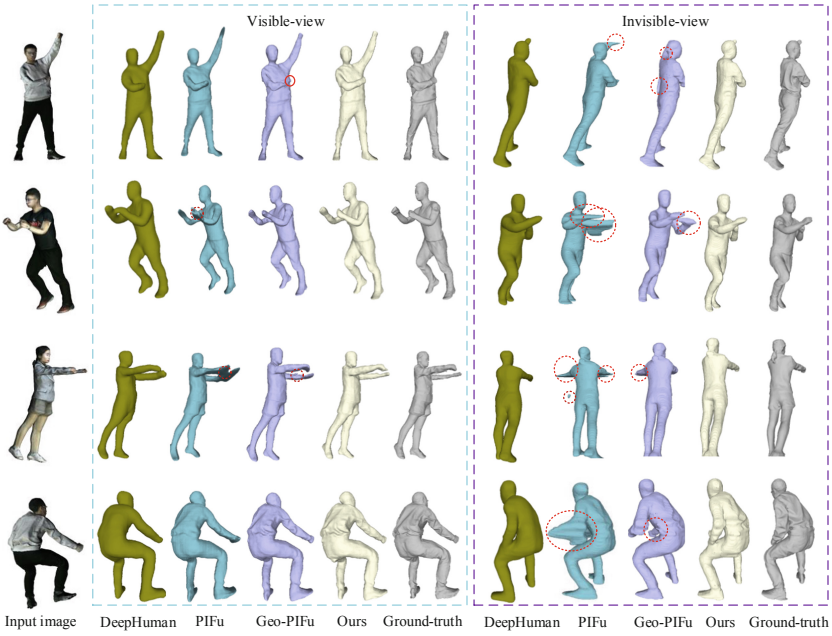

Fig. 3. Qualitative 3D human reconstruction results on the DeepHuman dataset.

4.5 Ablation Studies

To further explain the effect of the explicit model and the implicit model, we conduct ablation studies. The implicit module (the bottom part in Fig. 2) and

the explicit module (the upper part in Fig. 2) are respectively removed from our method and the remained networks are trained and tested with the same data and protocol described in Subsect. 4.1. Table 2 shows the results achieved by our method with different module configurations. Obviously, the Chamfer distance (ε_{cd}) and point-to-surface distance (ε_{psd}) produced by our method significantly increase after removing the explicit module, while the cosine distance (ε_{cos}) and L2-norm (ε_{l2}) grow larger after removing the implicit module. These quantitative changes have also been visually proved by the qualitative results illustrated in Fig. 4. On one hand, some topology artifacts and distortions (such as the arms and feet circled in red) can be found on the reconstructed human bodies from the invisible view after removing the explicit module. On the other hand, most of the local details on the surface of reconstructed 3D human bodies disappear after removing the implicit module. It can be inferred from these ablation experimental results that the explicit module is beneficial to improve the global regularities and the implicit module helps capture fine-scale surface details from the input image. By combining the explicit and implicit modules, our hybrid method (i.e., HEI-human) not only generates reconstructions with plausible global topology but also recovers abundant and accurate surface details.

Table 2. Quantitative results achieved by our method with different module configurations on the DeepHuman dataset.

Modules	ε_{cd}	ε_{psd}	ε_{cos}	ε_{l2}
Explicit + Implicit	**0.1742**	**0.2297**	**0.0661**	**0.2540**
Only implicit	2.6004	4.0174	0.0949	0.3048
Only explicit	0.6134	0.4997	0.0968	0.3211

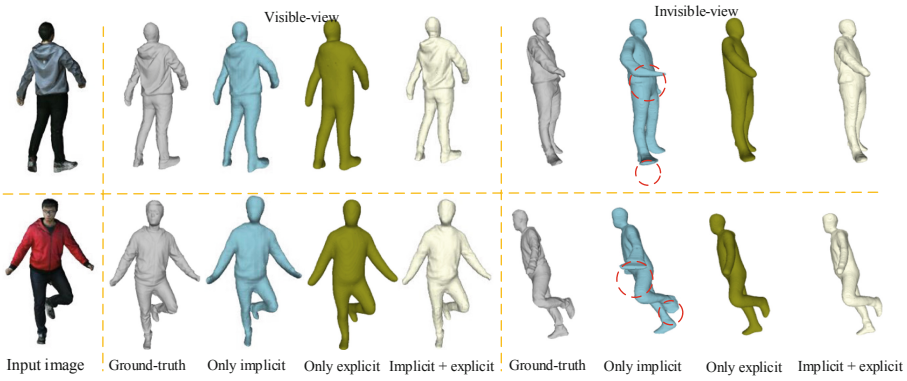

Fig. 4. Qualitative results achieved by our method with different module configurations.

5 Conclusions

As an inherently ill-pose problem, 3D human reconstruction is challenging not only because of the requirement to infer the complex global topology of human body but also due to the need to recover surface details. In order to outcome these challenges, a method named HEI-Human has been proposed to hybridize an explicit model and an implicit model for 3D clothed human reconstruction. Ablation studies have shown that the explicit model is beneficial for global topology while the implicit model mainly takes charge of recovering the surface details. As a consequence, our hybrid method recovers rich surface details and rarely produces unreasonable artifacts. Experimental results on the DeepHuman dataset demonstrate that our HEI-human outperforms the current state-of-the-art methods including DeepHuman [5], PIFu [6], and GeoPIFu [7].

References

1. Pons-Moll, G., Pujades, S., Hu, S., et al.: ClothCap: Seamless 4D clothing capture and retargetting. ACM Trans. Graph. (TOG) **36**(4), 1–15 (2017)
2. Cha, W., Price, T., Wei, Z., et al.: Towards fully mobile 3D face, body, and environment capture using only head-worn cameras. IEEE Trans. Visual. Comput. Graph. **24**(11), 2993–3004 (2018)
3. Ji, M., Gall, J., Zheng, H., et al.: Surfacenet: an end-to-end 3D neural network for multiview stereopsis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2307–2315. IEEE, Italy (2017)
4. Yu, T., Zheng, Z., Guo, K., et al.: 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, pp. 7287–7296. IEEE, Salt Lake City (2018)
5. Zheng, Z., Yu, T., Wei, Y., et al.: Deephuman: 3D human reconstruction from a single image. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 7739–7749. IEEE, Korea (2019)
6. Saito, S., Huang, Z., Natsume, R., et al.: PIFu: pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2304–2314. IEEE, Korea (2019)
7. He, T., Collomosse, J., Jin, H., et al.: Geo-PIFu: geometry and pixel aligned implicit functions for single-view human reconstruction. In: Advances in Neural Information Processing Systems, pp. 9276–9287, NIPS (2020)
8. Loper, M., Mahmood, N., Romero, J., et al.: SMPL: a skinned multi-person linear model. ACM Trans. Graph. (TOG) **34**(6), 1–16 (2015)
9. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Proceedings of the IEEE Conference on European Conference on Computer Vision, pp. 483–499. IEEE, Holland (2016)
10. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3D object reconstruction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 605–613. IEEE, U.S.A (2017)
11. Kanazawa, A., Black, J., Jacobs, W., et al.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7122–7131. IEEE, U.S.A (2018)

12. Kolotouros, N., Pavlakos, G., Black, M.J., et al.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2252–2261. IEEE, Korea (2019)
13. Zhang, H., Cao, J., Lu, G., et al.: Learning 3D human shape and pose from dense body parts. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
14. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4501–4510. IEEE, U.S.A (2019)
15. Alldieck, T., Pons-Moll, G., Theobalt, C., et al.: Tex2shape: detailed full human body geometry from a single image. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2293–2303. IEEE, Korea (2019)
16. Varol, G., Ceylan, D., Russell, B., et al.: Bodynet: volumetric inference of 3D human body shapes. In: Proceedings of the European Conference on Computer Vision, pp. 20–36. IEEE, Germany (2018)
17. Onizuka, H., Hayirci, Z., Thomas, D., et al.: TetraTSDF: 3D human reconstruction from a single image with a tetrahedral outer shell. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pp. 6011–6020. IEEE(2020)
18. Jackson, S., Manafas, C., Tzimiropoulos, G.: 3D human body reconstruction from a single image via volumetric regression. In: Proceedings of the European Conference on Computer Vision Workshops. IEEE, Germany (2018)
19. Saito, S., Simon, T., Saragih, J., et al.: PIFuHD: multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 84–93. IEEE (2020)
20. Huang, Z., Xu, Y., Lassner, C., et al.: Arch: animatable reconstruction of clothed humans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3093–3102. IEEE (2020)
21. Natsume, R., Saito, S., Huang, Z., et al.: Siclope: silhouette-based clothed people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4480–4490. IEEE, U.S.A (2019)
22. Lorensen, E., Cline, E.: Marching cubes: a high resolution 3D surface construction algorithm. ACM SIGGRAPH Comput. Graph. **21**(4), 163–169 (1987)
23. Bogo, F., Kanazawa, A., Lassner, C., et al.: Keep it SMPL: automatic estimation of 3D human pose and shape from a single image. In: Proceedings of the IEEE International Conference on European Conference on Computer Vision, pp. 561–578. IEEE, Holland (2016)
24. Jackson, S., Bulat, A., Argyriou, V., et al.: Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1031–1039. IEEE, Italy (2017)
25. Paszke, A., Gross, S., Massa, F., et al.: Pytorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, pp. 8024–8035. NIPS, Canada (2019)
26. Liu, L., Ke, Z., Huo, J., et al.: Head pose estimation through keypoints matching between reconstructed 3D face model and 2D image. Sensors **21**(5), 1841 (2021)
27. Liu, L., Zhang, L., Chen, J.: Progressive pose normalization generative adversarial network for frontal face synthesis and face recognition under large pose. In: Proceedings of the IEEE International Conference on Image Processing, pp. 4434–4438. IEEE, Taibei, China (2019)