Point2PartVolume: Human Body Volume Estimation from A Single Depth Image

Pengpeng Hu, Xinxin Dai, Ran Zhao, He Wang, Yingliang Ma, and Adrian Munteanu

***Abstract*—Human body volume is a useful biometric feature for human identification and an important medical indicator for monitoring body health. Traditional body volume estimation techniques such as underwater weighing and air displacement demand a lot of equipment, and are difficult to be performed under some circumstances, e.g. in clinical environments when dealing with bedridden patients. In this contribution, a novel vision-based method dubbed Point2PartVolume based on deep learning is proposed to rapidly and accurately predict the part-aware body volumes from a single depth image of the dressed body. Firstly, a novel multi-task neural network is proposed for jointly completing the partial body point clouds, predicting the body shape under clothing, and semantically segmenting the reconstructed body into parts. Next, the estimated body segments are fed into the proposed volume regression network to estimate the partial volumes. A simple yet efficient two-step training strategy is proposed for improving the accuracy of volume prediction regressed from point clouds. Compared to existing methods, the proposed method addresses several major challenges in vision-based human body volume estimation, including shape completion, pose estimation, body shape estimation under clothing, body segmentation, and volume regression from point clouds. Experimental results on both the synthetic data and public real-world data show our method achieved average 90% volume prediction accuracy and outperformed the relevant state-of-the-art.**

***Index Terms*—Volume estimation, Biometric data security, 3D Scanning, Deep learning, Human body shape reconstruction, Human body under clothing, Point cloud completion**

I. Introduction

HUMAN body volume data, including the volumes of the whole body and body parts, is necessary for many human-centered applications. Body volume data is valuable for determining the drug dosage for emergency patients [1], early detection of peripheral oedemas [2], fibrosis [3] and lymphedemas [4], measurements of muscle atrophy [5], supervision of recovery process after invasive surgeries [6]. Furthermore, body volume data is an important indicator to evaluate growth of muscle mass, which helps creating an optimal training schedule in sports [5]. Human body volume estimation is also of particular importance to derive total body composition, by measuring the resistivity of the whole body or only its segments and by determining their volumes. While resistivity measurements are very accurate, volume estimation remains the main uncertainty factor [7].

To estimate the body volume, traditional methods mainly consist of underwater weighing, air displacement, and medical imaging methods. Underwater weighing measures the volume change of water when a person is immersed into the water. Air displacement is similar to underwater weighing, but it uses air displacement rather than water immersion. However, both methods require a lot of non-portable equipment and can only measure the whole-body volume. Users have to remove clothes during volume acquisition. More important is the fact that such methods are difficult to use for disabled or bedridden people.

Medical imaging methods for body volume estimation make use of Magnetic Resonance Imaging (MRI) or Computed Tomography (CT). However, these kind of methods rely on the expensive systems and require expert knowledge for operation, being also inconvenient for the user. It is for these reasons that, some vision-based methods for human body volume estimation have been proposed. The vision-based approaches for volume estimation can be mainly classified into two categories according to the type of input: RGB image-based methods and depth image-based methods. RGB image-based methods [8] usually suffer from scale ambiguity due to the lack of accurate range information. The depth image-based approach [12] is more popular in the task of volume estimation. However, fast and accurate estimation of body volume is still under-researched.

With the advent of consumer depth cameras, depth images have been widely used in different works [9], [10], [11], [12], [13]. Given a precise 3D model of a subject, both the whole-body volume and partial volumes can be extracted [14]. However, 3D model-based body volume estimation methods have the following drawbacks: (1) existing methods require to manually extract the part volume values; (2) they require a clean watertight body mesh as the input. However, due to limited resolution of the scanner and acquisition conditions, noise and occlusions are inherent and strongly influence the quality of the resulting 3D scans. Even though many post-processing techniques such as denoising and hole-filling have been proposed to address these problems, they may introduce new errors during post-processing which influence the accuracy of the volume estimates; (3) the user has to wear minimal clothes and multiple depth images have to be captured from various views to obtain the complete shape of subjects, which

This work was supported in part by Innoviris under project AI43D and in part by Fonds Wetenschappelijk Onderzoek (FWO) under Project G094122N.

P. Hu is with the Centre for Computational Science and Mathematical Modelling, Coventry University, Coventry, UK, email: Pengpeng.Hu@coventry.ac.uk (corresponding author).

X. Dai is with the Department of Electronics and Informatics,Vrije Universiteit Brussel, Brussels, Belgium, email: xdai@etrovub.be.

R. Zhao is with the Department of Electronics and Informatics,Vrije Universiteit Brussel, Brussels, Belgium, email: rzhao@etrovub.be.

H. Wang is with the School of Computing, University of Leeds, Leeds, UK, email: h.e.wang@leeds.ac.uk.

YL. M is with the School of Computing Sciences, University of East Anglia, Norwich, UK, email: Yingliang.Ma@uea.ac.uk.

A. Munteanu is with the Department of Electronics and Informatics,Vrije Universiteit Brussel, Brussels, Belgium, email: acmuntea@etrovub.be.

makes these methods slow and inconvenient in practice, in particular to immobilized / bedridden patients; (4) 3D scanned body models contain personal information of scanned subjects (e.g. faces, gender, body measurements), which may result in biometric data leakage problems.

To address these issues, we propose to take a partial body scan as input and to predict the undressed complete body and its part volume values. A major advantage is that the method can be used for bedridden subjects. In addition, compared to complete 3D body models, the use of partial noisy body point clouds contributes by design to protecting the biometric data of subjects.

The main contributions of this paper can be summarized as follows:

- A novel vision-based approach is proposed for estimating human body volume from a single depth image. To the best of our knowledge, this is the first deep learning method for estimating human whole-body and part volumes.
- A novel **M**ulti-task network is proposed for **H**uman **B**ody **R**econstruction (MHBR) from a single depth image. It jointly completes partial point cloud of the subject, predicts the body shape under clothing, and segments the estimated complete body. A novel part-ware feature is presented to improve the performance of the MHBR.
- A novel human **B**ody **V**olume **N**etwork (BVN) is proposed for predicting volume values from point clouds. To improve the performance of BVN, we propose a simple yet efficient two-step training strategy to extend the learning of PointNet and its variants from sparse point sets to dense point sets.
- A novel 3D human dataset is constructed consisting of 400k models with actual volumes labeled and used to train and evaluate the proposed method.

The rest of the paper is organized as follows. First, we review the related works in Section II including 3D human body reconstruction, body volume estimation from a 3D model, and deep learning on point clouds. Second, we introduce the proposed method and dataset in Section III. Next, expensive experimental results are given in Section V. Finally, we conclude the article in Section VI.

II. Related Work

A. 3D Human Body Reconstruction

A great deal of works have proposed various methods of reconstructing a 3D human body model. Using structured light or a laser scan, high-quality human models can be created. However, these systems are generally very expensive and bulky, and usually require expert knowledge for operation. With the advent of Microsoft Kinect, many researchers have attempted its use for low-cost 3D scanning. In [15], the 3D data captured from varying viewing positions was fused based on non-rigid registration algorithms. The authors of [16] observed that a Kinect device should be put at around 3 meters away from the body in order to scan a complete human shape; this results in very low-resolution scan data. They proposed a scanning system based on three Kinect devices plus a turntable system to obtain better human models. However, these methods require that the subject should stand without moving for about 30 seconds or more, while ordinary people can keep a "frozen stand" for only several seconds (typically 3 seconds). [17] presented a multi-sensor Kinect system based on RGB-D devices to address this problem. Although these methods can reconstruct detailed human models, they assume the subject can *stand* still in a canonical pose, e.g. A-pose or T-pose. Old people or bedridden patients, however, are not able to stand still.

Another interesting approach has been proposed in [18], whereby body shapes are inferred from single RGB images. However, this method often has a large bias due to the scale ambiguity and occlusions, which is less accurate. Similarly, [19] predicts an opposite-view depth image from a single depth image using a convolutional neural network (CNN), then combines the 3D points from the depth images. Although they assume that the input depth image contains half of the whole-body and the predicted opposite-view depth image provides the other half, missing data on the boundary area still exists. Besides, this method has to be integrated with other techniques including denoising, surface completion, surface reconstruction and auto rigging to be able to perform body volume estimation. These post-processing steps, however, will introduce new errors.

B. Body Volume Extraction from a 3D Model

3D scanning has been successfully adopted by many healthcare applications. Earlier medical scientists make use of body scanners and manual post-processing, their accuracy having been validated [14]. Although the acquisition of scanned data is quick, the manual post-processing is time-consuming and highly depends on the technical expertise. To address this issue, several automated works have been proposed. [20] assessed the whole-body volume from 3D photogrammetric scanning by comparing with measurements from traditional underwater weighing and air displacement. Part volumes, however, are missing. In [14], a bespoke method for obtaining whole-body volumes and part volumes from 3D scanned data has been proposed. However, this method has to take an A-pose, clean and complete model as input. [7] developed a system consisting of sixteen stereo cameras, four projectors and a custom-built couch for human body volume estimation in a clinical environment. During testing, a parametric humanoid model from Makehuman open source project was applied to fit the frontal-view scan of the lying-down patient by minimizing the Euclidean distance between the parametric humanoid model and the partial scan of the subject. Once reconstructing a complete body, the whole-body volume is calculated using cross sections along the body. This work is a step forward for estimating body volume in the clinical environment. But the user has to manually select the joint positions and part volumes cannot be measured. Besides, this system needs precise calibration and is computationally expensive. [1] proposed a similar system by replacing the high number of cameras and projectors with a portable Kinect. However, this method adopted an inaccurate solution to recover the

back side of the patient by simply projecting the front surface along the rays emerging from the sensor to the stretcher plane. Moreover, the above methods will fail when the subject wear loose clothes.

Our study is mainly inspired by the work of [1][7]. Compared to the existing approaches, our method mainly has the following advantages: 1) we address the volume estimation problem from only a single depth image using a novel learning-based framework; 2) the proposed method can work for dressed bodies; 3) the method can accurately predict the body part volumes; 4) the proposed method is fully-automatic and fast; 5) our method is robust against the inherent pose variations and outliers.

C. Deep learning on point clouds

3D point clouds represent accurate shapes of subjects. Depth images are directly converted to point clouds given the intrinsic parameters of the depth camera. A pioneering work on deep learning for point cloud processing is PointNet proposed in [21]. It utilizes a pointwise multi-layer perceptron with a symmetric aggregation function to implement invariance to permutations, and shows competitive performance for extracting features from point clouds. PointNet-based methods have been employed in various applications [21]. Although PointNet has not been extended for the task of human body volume estimation, some works that address similar problems to those addressed in this study have been proposed. The authors of [22] proposed a method based on PointNet to complete point clouds. This method is not trained on human datasets, and outputs point clouds which cannot be accurately used to estimate volumes. In [23] a method to deform a predefined template to fit the input point cloud has been proposed. Its performance highly depends on an additional optimization refinement that minimizes the Euclidean distance between the prediction and the input. Such an approach would fail in our task, as one takes a partial scan as input point cloud. The closest to our work is [12], which trained two PointNet-based neural networks to implement point cloud completion and volume estimation of food for dietary assessment purposes. However, the models are trained on a synthesized food dataset and can only predict whole volumes. Compared to the work of [12], we aim to address a more challenging problem due to the following reasons: 1) human body is a articulated shape that is complex due to the large pose variation; 2) human body is usually covered by cloth while the food is exposed; 3) to estimate part volumes is more challenging than to predict the whole-body volume.

III. Problem Statement

Given a point set of the partial body scan $X = \{x_i \in \mathbb{R}^3, i = 1, ..., n\}$, the goal is to determine the volume vector $V = \{v_i \in \mathbb{R}, i = 1, ..., m\}$ which stores the body part volume values of the user. This problem can be conventionally resolved by factoring it into three sub-problems, namely (i) predicting a complete body mesh $B = \{(p_i \in \mathbb{R}^3, e_j \in \mathbb{Z}^2), i = 1, ..., n, j = 1, ..., m\}$ given X by fitting a template mesh to the partial scan, where p_i, e_j are the vertices and edges in B respectively, (ii) adjusting or normalizing the pose of B to obtain mesh $B^{'}$ in an suitable pose for volume extraction, and (iii) extracting V from $B^{'}$ by existing automatic algorithms or manual processing. Although this formulation is intuitive, such a method has the following disadvantages: 1) it requires the user to wear minimal clothes, and will fail for the dressed body; 2) it highly relies on the template-based fitting which is time-consuming and prone to the initialization and outliers; 3) when the posture of subject is not suitable, posture adjustment is required. However, an automatic rigging yields a poor pose normalization, leading to situations when the algorithm gets stuck; 4) human interaction is necessary for accurate volume prediction. In stark contrast, we do so using two key ideas based on deep learning. First, we propose a deep neural network taking X as input to reconstruct a complete body point cloud $Y = \{y_{part} \in \mathbb{R}^{N\times 3}, part = head, torso, left\_arm, right\_arm, left\_leg, right\_leg, N \in \mathbb{Z}\}$. Y is an assembly of semantic body parts. Such a formulation provides the foundation for jointly completing the partial body point clouds, predicting the body shape under clothing, and segmenting the reconstructed body into parts. Second, taking y_{part} as input, our proposed volume prediction network regresses the partial volumes from point clouds of reconstructed body parts. Our method avoids expensive template-fitting optimization and error-prone posture adjustment. More importantly, our method proves to work well for dressed bodies.

IV. Proposed Method

A. Overview

The overview of proposed method is illustrated in Figure 1. Given a depth image of the front-facing dressed body, it is firstly converted to a partial point cloud. Then, the partial point cloud is fed into the proposed HMBR for reconstructing a complete body shape under clothing with semantic segmentation. Each part of the reconstructed body is further fed into the proposed BVN for regressing corresponding part volume values. The architectures of HMBR and BVN are shown in Figure 2. The partial point cloud is fed into a multi-path encoder represented by a set of sub-encoders $\{HE, TE, LAE, RAE, LLE, RLE\}$, where HE is the Head Encoder, TE is the Torso Encoder, LAE is the Left Arm Encoder, RAE is the Right Arm Encoder, LLE is the Left Leg Encoder, and RLE is the Right Leg Encoder. Each sub-encoder focuses on learning features for different body parts. We, thus, call them part-aware features denoted by $\{f_{HE}, f_{TE}, f_{LAE}, f_{RAE}, f_{LLE}, f_{RLE}\}$. Each set of part-aware features aims at reconstructing the corresponding body part by means of a part-aware decoder. For instance, the head features f_{HE} are fed to the head decoder HD which reconstructs the head shape. All of the reconstructed body parts make up the complete body shape. It can be seen that our reconstructed body shape is complete and well-segmented. Besides, it is important to observe that the proposed method provides an estimate of the body shape under clothing. In this sense, the design in Figure 2(a) is a multi-task network. In the

subsequent steps, each of reconstructed parts is fed into the volume regression network to regress the volume values, as shown in Figure 2 (b).

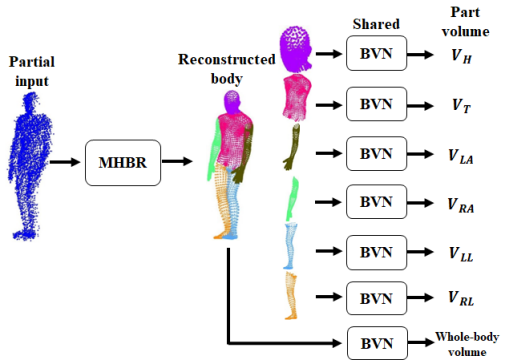

Fig. 1. Overview of the proposed method. First, the partial body scan is fed into the HMBR to obtain a complete body shape under clothing with semantic segmentation. Next, the part of reconstruct body and the whole reconstructed body are further fed into BVN for regressing the corresponding part volume and the whole-body volume.

The details of proposed HMBR and BVN are illustrated in Figure 2.

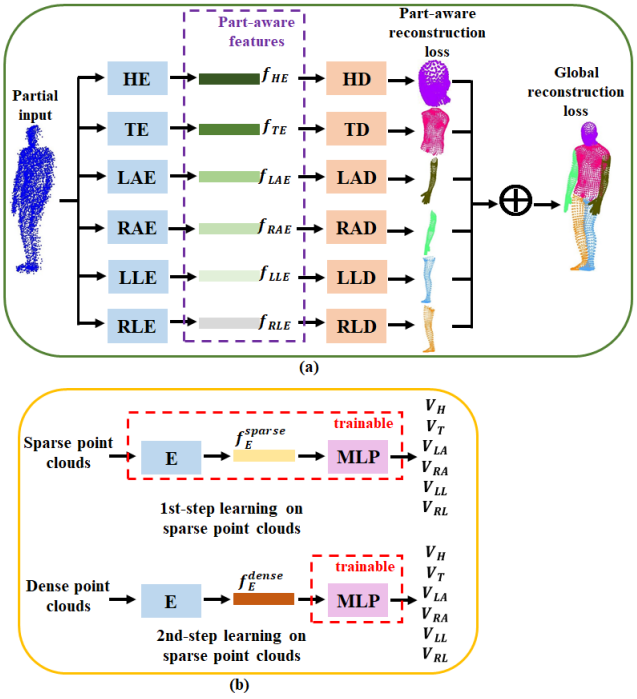

Fig. 2. Network architectures for body volume estimation. (a) Architecture of multi-task human body shape reconstruction network. $\{HE, TE, LAE, RAE, LLE, RLE\}$ represents the part-aware multi-path Encoder for $\{head, torso, left\_arm, right\_arm, left\_leg, right\_leg\}$, and $\{HD, TD, LAD, RAD, LLD, RLD\}$ represents the Multi-task Decoder for $\{head, torso, left\_arm, right\_arm, left\_leg, right\_leg\}$. (b) Architecture of human body volume network. It regresses the volume values for the reconstructed part point clouds of the subject. E denotes the point cloud encoder, MLP acts as the volume regressor, and $\{V_H, V_T, V_{LA}, V_{RA}, V_{LL}, V_{RL}\}$ denotes the volumes for $\{head, torso, left\_arm, right\_arm, left\_leg, right\_leg\}$.

B. Proposed Synthetic Dataset

To train our algorithm, a large-scale 3D dataset is required. More specifically, we require various arbitrarily-posed partial scans of dressed bodies as input and its paired complete undressed body labelled by volume values as the ground truth. There exists no such dataset in the literature. We, thus, propose a new synthetic dataset, termed Body Volume (BV) Dataset, which is needed in order to train our model. The proposed pipeline of synthesizing our dataset is summarized in Figure 3. It mainly consists of the following modules.

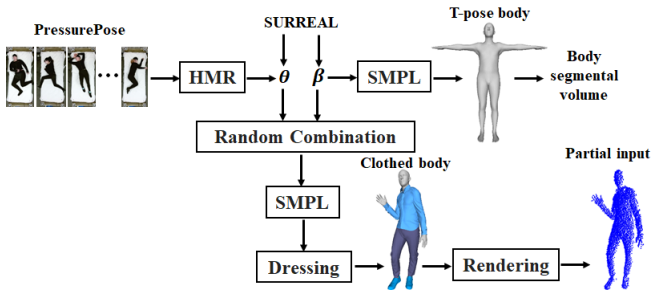

Fig. 3. Illustration of our pipeline for synthesizing the training dataset.

Realistic human shape and posture. To synthesize the realistic human bodies, we make use of SMPL [26], a state-of-the-art generative body model. SMPL is parameterized by a shape parameter denoted by β and a pose parameter denoted by θ. Both β and θ are represented by the vectors with the size of 10 and 72 respectively. Sets of β and θ values for the SMPL model are collected from the SURREAL dataset [27] in order to build realistically posed human bodies. One limitation of the SURREAL dataset [27] is it does not have in-bed poses. Without adapting these training poses, our algorithm generalized poorly to the in-bed patients. To overcome this limitation, we collected 1051 SMPL θ values from the PressurePose dataset [24] using HMR [25]. The PressurePose dataset has in-bed posed RGB images for 20 human subjects. HMR is a technique that fits the SMPL model to a single RGB image and outputs the β and θ values of SMPL. However, the predicted β is usually not reliable due to the ambiguity from 2D to 3D. Consequently, we only use the pose information extracted by HMR in our study. Finally, we sample $2{\cdot}10^5$ parameters by randomly combining our collected SMPL β and θ values for the male and female respectively. Our final dataset has $4{\cdot}10^5$ human meshes with a large variety of realistic poses and body shapes.

Clothing. Next, garments need to be put on the synthetic body meshes. As our dataset is very large-scale, to dress these synthetic body meshes is time-consuming and expensive using physically-based simulation. The method of [13] is adopted to dress our synthetic bodies due to its efficiency and simplicity. Moreover, the method of [13] can also dress shoes on the body. This method mainly consists of two procedures. Firstly, the fitted garments are manually designed based on a SMPL body in a canonical posture by a fashion expert. This manual preparation is only done once for each type of clothing. Next, the garments will be automatically transferred from the reference body onto other SMPL bodies in arbitrary postures.

In this study, we put the shoes and the common type of clothing, namely the long-sleeved shirt and long pants, onto our body shapes to validate our algorithm. More clothing types can be easily added to support the specific task of body shape estimation under clothing [13].

Rendering. We utilize the open-source Blender Sensor Simulation plugin Blensor to render realistic partial scans. In our experiments, we set the camera as Microsoft V2 sensor. During rendering, the position and orientation of camera are randomly selected at intervals from 1.5 to 2.5 meters in x, y, z directions and solid-angle orientations relative to the vertical axis from -10 degrees to 10 degrees. We also add noise with noise parameters set to $\mu = 0.0, \sigma \in (0.0, 0.025)$.

Volume Annotation. To avoid the effects of posture on the volume prediction, we perform the volume annotation on the canonical "T" pose. By setting the SMPL θ to be zero, a T-pose body can be obtained. The whole-body volume can then be easily computed [32]. To compute the part volumes, it is necessary to segment the body into parts. To address this problem, a simple yet efficient method is employed. We manually select five planes $\{s_1, s_2, s_3, s_4, s_5\}$ on one of the T-pose SMPL meshes to segment the body into five parts including head, left arm, right arm, left leg and right leg, as shown in 4. Each plane s_i is obtained by manually selecting three different points and by parametrizing the plane passing through them. It is possible that one of the selected points is not a vertex of the SMPL mesh (e.g., a point inside a triangle of the SMPL mesh). To address this, we represent the selected points via barycentric coordinates. It should be noted that this manual operation is only done once, and a more complicated or refined body segmentation can be obtained by selecting more planes.

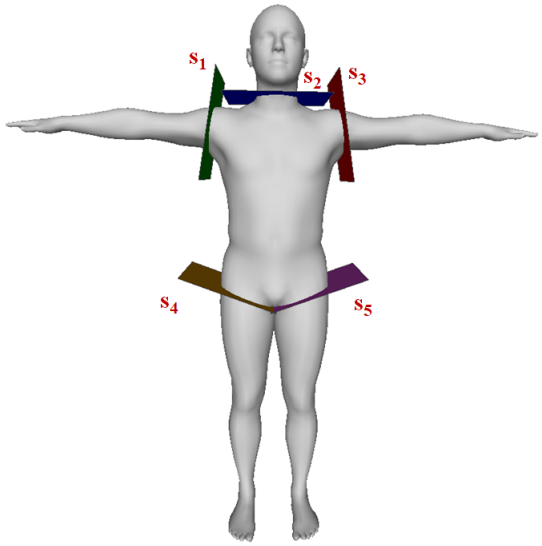

Fig. 4. The definition of the cutting planes for volume annotation.

C. Multi-task Human Body Shape Reconstruction Network

If using partial point clouds of dressed bodies to determine the whole-body and part volume values, predicted results will be largely overestimated. To address this problem, three key problems should be resolved before estimating the volume: completing the partial point cloud, estimating body shape under clothing, and semantically segmenting the complete body. To this end, a novel multi-task network (MHBR), shown in Figure 2(a), is tailored for jointly implementing these three tasks. The proposed multi-task network is built based on the proposed encoder–decoder architectures.

1) Part-aware Feature Learning: Following the paradigm of encoder-decoder architecture, we first attempt to extract features from partial point clouds. Point clouds are unstructured data, which is not easy for direct analysis. PointNet [21] has become a popular deep learning-based encoder that directly takes point clouds as input. The readers can refer to PointNet for detailed discussion. State-of-the-art encoder-decoder architectures usually aggregate a single feature represented by a vector f from the input signals, and then interpret f to the designed predicted signals. For instance, Point2Volume [12] used a PointNet-based encoder to learn a feature vector from a partial point cloud, and used a decoder to interpret the feature vector to a complete point cloud. Such a feature was learned by considering all of the input points. However, our insight is that the foot points of the input are significantly important for reconstructing the foot shape while they are nearly useless for reconstructing the head shape. To address this, we proposed multi-path encoders to extract part features. Specifically, each sub-encoder focuses on extracting corresponding part features. We represent the part-aware features by a matrix $F = [f_1; f_2; ...; f_M]$ consisting of M vectors f_i. M is a hyper parameter, and in this study we set $M = 6$ as we hope to predict the head, left arm, right arm, left leg, right leg, and the torso parts. We design the part-aware feature extractor e based on the unit of PointNet-based encoder g. In this study, similar to PCN [22], we use two stacked PointNet architectures with maxpooling operation to model g, which is expressed as:

$$f = g\left(X|w_g\right), g = g^1 \circ g^2 \tag{1}$$

where w_g denotes the weights of g^1 and g^2, where g^1 and g^2 represent the two PointNet-based sub-networks, respectively. e is a set of g, which is expressed as:

$$[f_1; f_2; ...; f_M] = e\left(X|w_{g_1}, w_{g_2}, ..., w_{g_M}\right) \tag{2}$$

where $w_{g_i}, i = 1, 2, ..., M$ denotes the weights of $g_1, g_2, ..., g_M$, respectively.

2) Multi-task Prediction: In this study, the multi-task prediction encompasses three components: to complete the partial point cloud, to estimate the body shape under clothing, and to segment the predicted complete body point cloud. Despite these tasks have been separately investigated, no existing methods can offer an all-in-one solution to jointly deal with the three problems. To this end, we propose a multi-path decoder architecture to take the $[f_1; f_2; ...; f_M]$ as input and output the complete segmented body point cloud under clothing $Y = \{y_{part} \in \mathbb{R}^{N\times 3}, part = head, torso, left\_arm, right\_arm, left\_leg, right\_leg, N \in \mathbb{Z}\}$. As shown in Figure2(b), our multi-task decoder consists of M units. These units act as point cloud predictors, which

are responsible for interpreting the feature vector f_i to the part-specific point cloud of the complete body from the proposed part-aware features $[f_1; f_2; ...; f_M]$. Similar to the coarse-to-fine prediction strategy [13], we use two stacked architectures to model the unit h, which is expressed as

$$y = h\left(f|w_h\right), h = h^1 \circ h^2 \tag{3}$$

where w_h denotes the weights of h^1 and h^2, h^1 and h^2 represent the two MLP-based sub-networks, respectively. Specifically, h^1 is a MLP with 1024, 1024, and $m \times 3$ neurons, where m is the number of predicted coarse points. In this study, we set $m = 1024$. h^2 is a stack of g^1 and h^1. Accordingly, our multi-task decoder d is expressed as:

$$\begin{aligned}\{y_{head}, y_{torso}, y_{left_arm}, y_{right_arm}, y_{left_leg}, y_{right_leg}\} = \\ d\left([f_1; f_2; ...; f_M]\,|w_{h_1}, w_{h_2}, ..., w_{g_M}\right)\end{aligned} \tag{4}$$

3) Multi-task Loss Function: The loss function of our multi-task human body shape reconstruction network consists of two parts: part-aware multi-stage reconstruction (PMR) loss and the global reconstruction (GR) loss. PMR loss measures the difference between the ground truth of body segments and the predicted body segments. GR loss tries to make the estimated body shape closer to the ground truth.

Part-aware Multi-stage Reconstruction Loss. Previous works [12], [13], [22] have introduced two permutation-invariant metrics to compare the similarity of two unstructured point clouds: the Earth Mover's Distance (EMD) and the Chamfer Distance (CD). We choose (CD) to design our loss as it is differential and more computationally efficient compared to EMD. Given two point clouds P_1 and P_2 and their cardinalities are denoted by $|P_1|$ and $|P_2|$, respectively, CD measures the average closest squared distance between them, which is defined by

$$\begin{aligned}CD\left(P_1, P_2\right) = \frac{1}{|P_1|}\sum_{x\in P_1}\min_{y\in P_2}||x-y||^2 \\ +\frac{1}{|P_2|}\sum_{x\in P_2}\min_{y\in P_1}||x-y||^2\end{aligned} \tag{5}$$

Since our multi-task decoder will output $M \times 2$ point clouds (M coarse point clouds and M fine point clouds) in different resolution, our part-aware multi-stage reconstruction loss consists of $M \times 2$ terms, as shown in Equation 6.

$$\begin{aligned}L_{PMR} = \sum_{part\in Y}\lambda_{part} \times CD\left(y_{part}^{coarse}, y_{part}^{GT}\right) \\ + \sum_{part\in Y}\alpha_{part} \times CD\left(y_{part}^{fine}, y_{part}^{GT}\right)\end{aligned} \tag{6}$$

where λ_{part} and α_{part} denote the weights that satisfy the following condition: $\alpha_{part} = 2 \times \lambda_{part} = 1$.

Global Reconstruction Loss. GR loss is designed to improve the performance of the proposed network. Similar to the PMR loss, but GR focuses on the global shape. Our global reconstruction loss is defined as:

$$L_{GR} = CD\left(Y, Y^{GT}\right) \tag{7}$$

Joint Reconstruction Loss. The joint loss of network is defined as:

$$L = \rho \times L_{PMR} + \xi \times L_{GR} \tag{8}$$

ρ and ξ are the weight of PMR loss and GR loss. In this study, we set $\rho = 1$, $\xi = 0.001$.

D. Human Body Volume Network

To estimate object volumes, the common approach is to employ the alpha-shape algorithm. [14] extended it to estimate body volume from 3D body model. However, this method requires a complete clean body model as input and manual interaction, and it is is prone to the selection of α. To address this, we propose a novel human body volume network (BVN) to predict volume values directly from unstructured point clouds. As Figure 2 (b) shows, BVN takes point clouds as input and outputs a scalar which represents the corresponding volume value.

1) Architecture: The architecture of BVN consists of two modules: feature encoder and volume regressor. The feature encoder of BVN follows the design principle of MHBR, which also takes point clouds as input and outputs a latent feature vector k. We use the first sub-network encoder architecture g^1 from MHBR to model the feature encoder, which is expressed as:

$$k = e\left(y_{part}|\psi_{g^1}\right), e = g^1 \tag{9}$$

where ψ_e is the weight of e in BVN. The volume regressor is built by means of a MLP with 1024, 1024, and 1 neurons, which is expressed as:

$$Volume^{estimated} = r\left(k|\tau_r\right) \tag{10}$$

τ_r denotes the weight of r. BVN is optimized by minimizing the following L1-norm loss:

$$Loss = |Volume^{estimated} - Volume^{GT}| \tag{11}$$

2) Performance Enhancement: We observed that deep neural network performs worse in volume regression task compared to the task of point cloud reconstruction. Our insight is to regress volume from point clouds is a global task, but the volume regression network is prone to the individual point. PointNet-based encoders have to be trained based on the sparse point clouds instead of dense point clouds due to the memory problem. In the shape reconstruction task, sparse point clouds are proved to well represent the global 3D shape for the shape prediction [13], [22]. However, for the regression task, since the neural network is trained based on the sparse point clouds, individual points will cause larger variations for volume prediction. To address this, we propose a two-step training method to enhance the performance of BVN.

Learning on Sparse Point Clouds. The first step is similar to the previous work [12] that takes sparse point clouds as input. We sampled a set of sparse points y_{part}^{sparse} from the dense points y_{part}, and feed y_{part}^{sparse} into BVN for training. We denote the trained volume prediction model by:

$$Volume^{estimated} = v\left(y_{part}\right), v = e \circ r \tag{12}$$

Learning on Dense Point Clouds. After the first-step training based on the sparse point clouds, we use the trained feature encoder e as a volume-specific feature extractor. By applying e onto the dense point clouds y_{part}, we obtain the features k^{dense}. Then, we take k^{dense} as input to retrain the volume regressor r. Note that the feature encoder is not trainable in the second-step training. We denote the updated weight of r as τ_r^{dense}. In the inference stage, the final volume prediction model can be expressed as:

$$Volume^{estimated} = v\left(y_{part}|\psi_e, \tau_r^{dense}\right), v = e \circ r \quad (13)$$

V. Experiments

A. Training Setup

We randomly split the dataset into training, validation, and testing using 97%, 2% and 1% of the samples in the dataset respectively. We note that, given the large size of the employed dataset (400k samples), the testing dataset has 4000 samples which is deemed to be large. Based on Tensorflow [29], Sanity checks based on the loss curves on the training and validation datasets indicate no overfitting on the training data. The resulting body point cloud is normalized in two steps: $1)$ it is centered to the origin, $2)$ and it is scaled by the Z-axis length of its bounding box. The training is carried out using the Adam optimizer [28] with an initial learning rate of 0.0001 for 50 epochs and a batch size of 16. The training is performed on a desktop PC (Intel(R) Xeon(R) Silver 4112 CPU @2.60GHz 64GB RAM GPU GeForce GTX 1080Ti) based on TensorFlow [29].

B. Evaluation Metrics

To evaluate the performance of our algorithm, we employ two metrics to analyze the shape reconstruction error (SRE) and relative volume error (RVE). The Chamfer Distance (CD) is used to define our SRE. Let the predicted points and volume be P and V. SRE is defined as:

$$SRE\left(P, P_{GT}\right) = CD\left(P, P_{GT}\right) \quad (14)$$

The measurement unit for SRE is the millimeter. We also calculate the average value μ and average standard deviation σ of SRE. The RVE is defined as:

$$RVE(V, V_{GT}) = |\frac{V - V_{GT}}{V_{GT}}| \times 100\% \quad (15)$$

Intuitively, the accuracy of volume prediction is defined as $(1 - RVE(V, V_{GT})) \times 100\%$.

C. Real-world Results

Our method, solely trained based on our synthetic data, is able to generalize well to the real-world data. To demonstrate its generalization, we, thus, perform more experiments based on two real-world datasets: the PDT13 dataset [30], and the BUFF dataset [31]. They provide an ID for each subject, such as M1 and 00005. In the PDT13 dataset, the front- and back-facing partial scans of the standing subjects are captured by a Microsoft Kinect V1 sensor, while the "ground-truth shape" are obtained by fitting a statistic human body model to a full-body laser-scan. BUFF is a scanned dressed body dataset consisting of 5 subjects wearing 2 clothing styles (T-shirt and long pants, and soccer outfit) in three motions by a custom-built multi-camera active stereo system. Its "ground-truth shape" is obtained by fitting the SMPL to the sequences in "A-T-U-Squat" motion and in "minimal clothing". To obtain the accurate "ground-truth volume", we manually segment the "ground-truth shape" in PDT13 referring to the definition in Figure 4. Since the "ground-truth shape" in BUFF has the same topology with SMPL, our volume annotation technique is directly applied for obtaining "ground-truth volume" for BUFF data. Figure 5 depicts our reconstruction results based on the PDT13 data. By overlapping our reconstructed bodies and the partial inputs, it can be seen that our results are visually correct while the "ground-truth shape" from [30] is visually incorrect. For instance, it can be observed the height of the "ground-truth shape" for M1 subject is smaller compared with its real-world partial scan. Similar phenomena can be observed in the M3 subject. Accordingly, as shown in Table I, consistent results are obtained. Focusing on whole volume prediction accuracy, it is seen that "M2>M3>M1". Figure 6 illustrates our reconstruction results on the BUFF data, which is visually correct by observing the overlap between our reconstructed bodies and the partial inputs. The volume prediction results is shown in Table II, which demonstrates the average accuracy reaches about 90%.

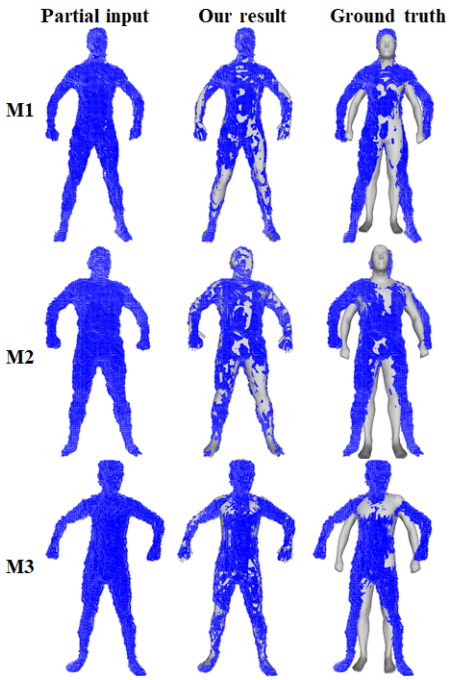

Fig. 5. Our results on PDT13 data.

D. Results on Unseen Synthetic Data

To further study the measurement accuracy of the proposed method, we randomly selected 450 unseen samples from the synthetic testing dataset and fed them into the proposed

TABLE I
VOLUME PREDICTION ACCURACY ON PDT13 DATA (UNIT: %).

Subject ID	Head	Torso	Left arm	Right arm	Left leg	Right leg	Whole body
M1	95.13	98.14	18.33	13.44	72.68	52.83	75.34
M2	96.85	78.29	44.27	29.32	62.56	68.38	91.28
M3	88.13	85.81	21.66	47.17	72.69	61.83	86.55

TABLE II
VOLUME PREDICTION ACCURACY ON BUFF DATA (UNIT: %).

Subject ID	Head	Torso	Left arm	Right arm	Left leg	Right leg	Whole body
00005	96.42	98.81	99.99	99.28	98.47	94.17	96.3
00032	89.74	78.48	93.95	97.48	88.59	85.60	82.23

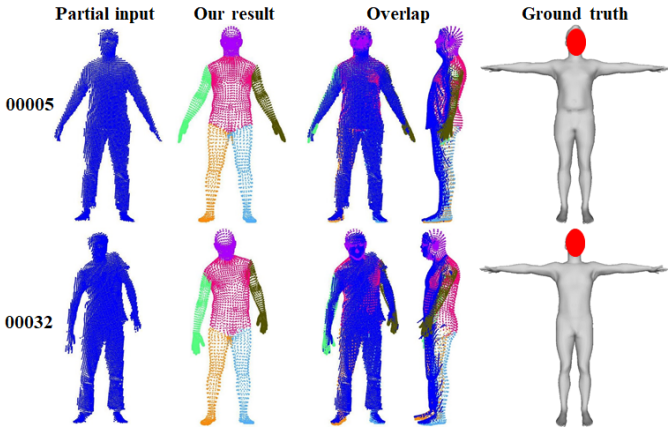

Fig. 6. Our results on BUFF data. Different color indicates different reconstructed body parts.

method. We define the accuracy for a given threshold (confidence level) as:

$$Accuracy^T = \frac{N}{M} \tag{16}$$

where T is a threshold of the volume prediction accuracy, M is the total number of tested samples and N is the number of samples above the threshold. In this study, $M = 450$, and we set T to be 75%, 80%, 85% and 90%. Table III depicts the results. It can be seen that almost 90% of the results achieved 80% accuracy and 80% of the results achieved 85% accuracy in terms of part-volume estimation. The proposed method worked best for predicting the torso volume with 94.4% of results achieving 90% accuracy.

TABLE III
VOLUME PREDICTION ACCURACY ON UNSEEN SYNTHETIC DATA.

Threshold	Head	Torso	Left arm	Right arm	Left leg	Right leg
75%	99.1%	99.6%	91.1%	93.6%	98.2%	96.4%
80%	97.3%	99.1%	84.9%	88.9%	94.9%	90.4%
85%	93.8%	98.0%	71.6%	81.8%	84.4%	75.3%
90%	79.1%	94.4%	53.6%	64.4%	54.7%	42.2%

E. Comparisons with Related Works

We proposed two deep neural networks in this study. To quantitatively compare our approach with related works, we firstly compare MHBR with state-of-the-art methods for single view-based body shape reconstruction. Next, we compare the performances of volume estimation using different methods. For reliable comparisons, the data used for comparisons should have partial dressed body scans, accurate ground-truth body shapes, accurate ground-truth part volume values. We, thus, perform the following experiments based on the male testing data (450 samples) that is not included in the training data.

1) Performance of MHBR: We compare our algorithm against state-of-the-art single-view based methods including DecoMR [33] takes a RGB image as input, Point2Volume [12] takes a partial point cloud as input and outputs the complete shape, and the method proposed in [19] that takes the front-facing body depth image as input and outputs back-facing body depth image. Table V shows the average reconstruction error comparisons, it can be seen [33] obtained the worst results due to the scale ambiguity from 2D to 3D, and our method achieve the best performance. Figure 7 illustrates some randomly-selected per-vertex error comparisons.

2) Performance of BVN: To compare the effectiveness of the proposed BVN, we compare our volume estimation model with VolumeNet proposed in Point2Volume [12] that is the state-of-the-art deep learning-based volume estimation method from partial point clouds. As shown in Table VI, our method significantly increases the volume prediction accuracy for the whole body, and we have better results in terms of the torso, left arm and right arm. We obtain comparable results in terms of left leg, right leg and head. Therefore, it can be concluded that our method outperforms [12]. It should be noted that the whole-body volume is not obtained by adding up all body-part volumes since the errors of each body part volume prediction will be accumulated. We take the reconstructed complete body point clouds as input and output the whole-body volume.

F. Selectivity and Sensitivity

In this section, we evaluated the Selectivity and Sensitivity of the proposed method.

TABLE IV
COMPARISONS WITH DIFFERENT SINGLE VIEW-BASED BODY RECONSTRUCTION METHODS.

Method	Input	Body shape reconstruction	Body shape completion	Body estimation under clothing	Body segmentation
DecoMR [33]	1 RGB image	✓	×	×	×
Point2Volume [12]	1 depth image	×	×	×	×
Point2Volume+Our dataset	1 depth image	✓	✓	✓	×
[19]	1 depth image	✓	✓	×	×
Ours	1 depth image	✓	✓	✓	✓

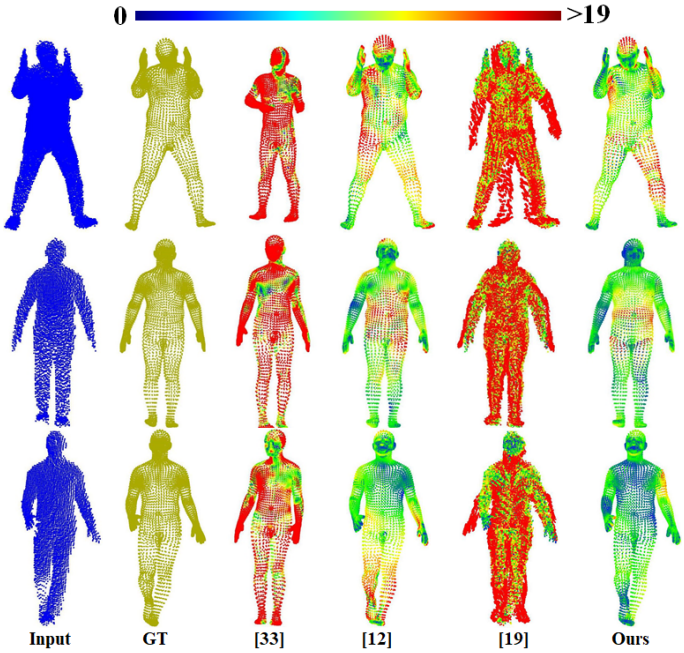

Fig. 7. Comparison of shape reconstruction error with related works. Each point is colorized based on the per-vertex error in millimeters.

TABLE V
COMPARISONS OF RECONSTRUCTION ERRORS WITH STATE-OF-THE-ART SINGLE VIEW-BASED BODY RECONSTRUCTION METHODS (UNIT: mm).

Method	DecoMR [33]	Point2Volume [12]	[19]	Ours
μ	13.94	0.18	0.83	0.02
σ	13.08	0.42	0.59	0.01
max	137.52	8.03	3.05	0.13

1) Selectivity evaluation: Our method takes a single front-facing partial scan of the body as input and produces estimates of body volume parts. To study the selectivity of input, we also rendered partial scans from the back and the side of subjects. As shown in Table VII, the results obtained from the front-facing partial scans are better than the results obtained from the side-view and back-facing partial scans. Note that the results from the back-facing are worst because the arms may be not visible from the back-facing view.

2) Sensitivity evaluation: Sensitivity is how sensitive is our method to a change in the input body volume for a given body part. We randomly select 100 samples from unseen synthetic dataset and progressively increase the input scale or decrease the input point cardinality which corresponds to decreasing the number of points at the input for the same body volume. The first set of experiments assesses the sensitivity of the method against scale changes. The second set of experiments illustrates the sensitivity of the method with the number of points at the input. The sensitivity K is defined as:

$$K = \frac{\overline{\Delta V}}{\Delta X} \tag{17}$$

where $\overline{\Delta V}$ denotes the average change of the predicted volumes and ΔX represents the change of the input.

In this study, we investigated the input scale change and input point size change. As Figure 8 and Figure 9 show, the K value of torso changes more significantly than K values of other parts remain stable. When the input scale increases, a larger volume change can be observed. Figure 10 and Figure 11 illustrate the sensitivity against input point cardinality. It can be noted that the volume change decreases when the sample ratio increases. The K values reach the minimal at the 60% sample ratio.

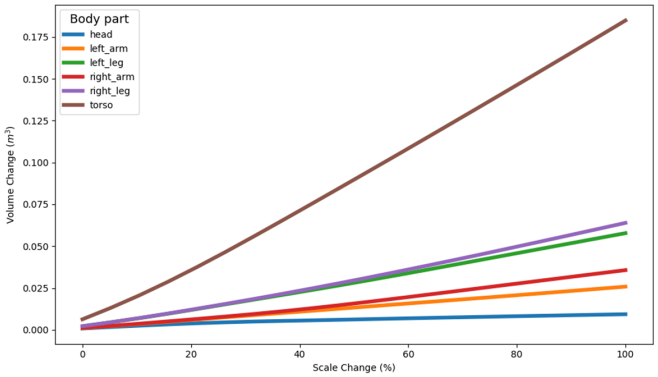

Fig. 8. Measured volume change against input scale changes.

G. Measurement Uncertainty

We compute the Type A measurement uncertainty for volume prediction based on both synthetic data (unseen male samples from the synthetic dataset) and real-world data (PDT13 and BUFF). We define the volume measurement uncertainty as:

$$\mu_A = \frac{s}{\sqrt{N}} \tag{18}$$

where s is the standard deviation of measurements and N is the total number of measurements. The obtained type A measurement uncertainties are depicted in Table VIII.

TABLE VI
COMPARISONS OF VOLUME PREDICTION ACCURACY WITH STATE-OF-THE-ART METHODS

Method	Whole body	Head	Torso	Left arm	Right arm	Left leg	Right leg
Point2Volume[12]	90.21%	95.04%	90.76%	84.30%	86.16%	89.43%	89.12%
Ours	99.37%	92.68%	95.65%	88.71%	90.49%	90.19%	88.31%

TABLE VII
COMPARISON OF THE RECONSTRUCTION RESULTS FROM FRONT-, BACK- AND SIDE-VIEW PARTIAL SCANS (UNIT: mm).

	front-facing	side-view	back-facing
μ	0.0182	31.158	38.3256
σ	0.013	15.715	17.7564
max	0.1339	155.6918	143.8

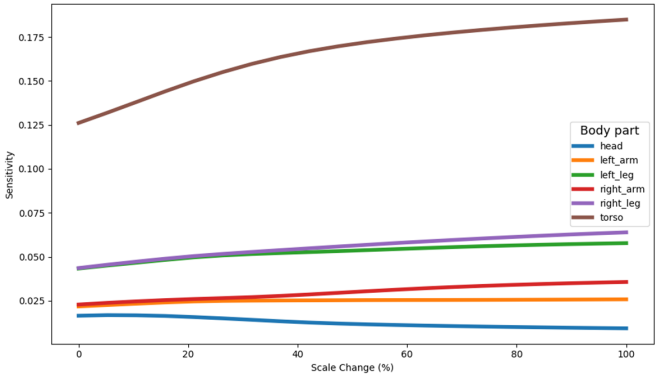

Fig. 9. Graph of $\frac{\Delta V}{\Delta X}$ (sensitivity) against input scale.

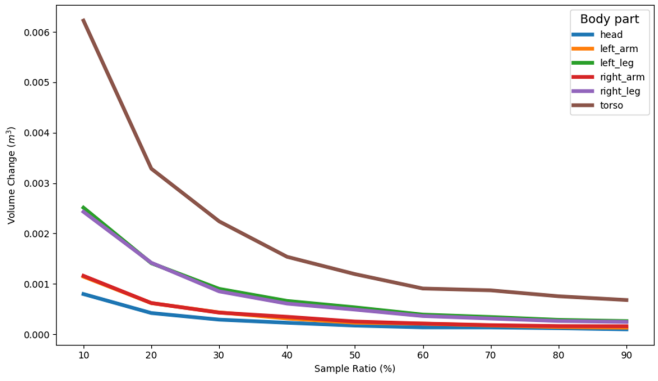

Fig. 10. Measured volume change against input scale changes.

TABLE VIII
TYPE A UNCERTAINTIES FOR BODY VOLUME MEASUREMENTS BASED ON SYNTHETIC DATA AND REAL-WORLD DATA (UNIT: cm^3)

Data type	Head	Torso	Left arm	Right arm	Left leg	Right leg
Synthetic	0.32	4.21	0.53	0.56	0.98	0.99
Real-world	4.37	27.39	4.68	5.73	9.98	10.11

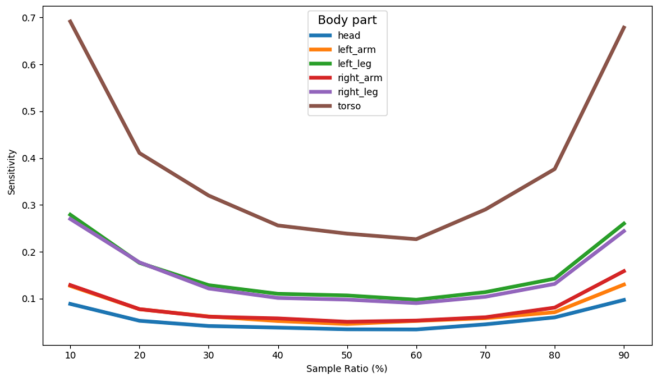

Fig. 11. Graph of $\frac{\Delta V}{\Delta X}$ (sensitivity) against input point cardinality.

H. Ablation Study

We conduct ablation experiments to understand the value of our network design and the influence of the different terms in our loss function. The Chamfer Distance is calculated to show the average reconstruction error, and RVE error is calculated to compare the volume regression results.

1) Part-aware features VS Global features: Firstly, we compare the performance of part-aware features and global features in the proposed multi-task network. We preserve one sub-encoder and remove the rest of five sub-encoders for learning the global features. As shown in Table IX, the proposed part-aware features can reduce the average reconstruction errors compared to the popular global features.

TABLE IX
ABLATION STUDY ON FEATURES (UNIT: mm).

Feature	Part-aware	Global
μ	0.0182	0.023
σ	0.013	0.0181
max	0.1339	0.1761

2) Loss selection: In the multi-task network, our loss consists of two terms: L_{PMR} and L_{GR}. L_{PMR} aims to minimize the body part shape reconstruction. L_{GR} is designed as a constraint to better stitch different reconstructed body parts into a complete body shape by minimizing the global body shape reconstruction. To validate the contribution of the proposed constraint, we compared L_{PMR} and $L_{PMR} + 0.001 \times L_{GR}$. Table X shows that $L_{PMR} + 0.001 \times L_{GR}$ indeed performed better than L_{PMR} since lower average reconstruction errors can be observed.

TABLE X
ABLATION STUDY ON THE LOSS (UNIT: mm).

SRE	L_{PMR}	$L_{PMR} + 0.001 \times L_{GR}$
μ	0.0186	0.0182
σ	0.0142	0.013
max	0.1675	0.1339

3) Volume regression from sparse point clouds VS Volume regression from dense point clouds: Regressing volume values from point clouds is a challenging problem. We observed that the regression result is prone to the position of each point. PointNet-based models have to be trained based on the sparse/sub-sampled point clouds. As Table XI shows, our proposed two-step training strategy can significantly improve the accuracy of volume regression from point clouds.

TABLE XI
ABLATION STUDY ON BODY VOLUME REGRESSION

RVE	downsampled 2048 points	whole points
μ	96.82%	99.37%
max	99.94%	100.00%
min	86.88%	93.89%

VI. CONCLUSION

In this article, a novel vision-based method was proposed to estimate human part volumes from a single depth image. It was built based on deep learning, and consisted of two networks. To the best of our knowledge, this is the first deep learning method for estimating human whole-body and part volumes. Firstly, the dressed body partial point cloud was converted to a complete body shape under clothing with semantic segmentation via the proposed multi-task human body shape reconstruction network. Next, each part of the reconstructed body was further fed into the developed body volume network for regressing the corresponding part volume. We observed that the volume regression was prone to each point since it was a global problem. We, thus, proposed a two-step training strategy to improve the performance of body volume network. Extensive experiments based on real-world and synthetic datasets showed the feasibility and efficiency of the proposed method, and showed our method outperformed the relevant approaches. It is attractive to extend the proposed method to various applications such as chest volume estimation for bra customization and breast cancer diagnosis, edema diagnosis by comparing the volume changes, and body weight estimation, to name a few. They are the interests of our works in the future. Besides, studying the effects of the ambient environment conditions on volume extraction and assessing the contribution of each point to the reconstruction accuracy are interesting research aspects. These are left as topics of further investigation.

REFERENCES

[1] C. Pfitzner, S. May, C. Merkl, L. Breuer, M. Köhrmann, J. Braun, F. Dirauf, and A. Nüchter, "Libra3d: Body weight estimation for emergency patients in clinical environments with a 3d structured light sensor," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 2888–2893.

[2] I. Haponiuk, M. Chojnicki, M. Steffens, R. Jaworski, A. Szofer-Sendrowska, J. Juscinski, E. Kwasniak, K. Paczkowski, J. Zielinski, and K. Gierat-Haponiuk, "Miniinvasive interventional bridge to major surgical repair of critical aortic coarctation in a newborn with severe multiorgan failure," *Videosurgery and Other Miniinvasive Techniques*, vol. 8, no. 3, p. 244, 2013.

[3] R. C. B. Ribeiro, S. M. P. F. Lima, A. C. G. Carreira, D. Masiero, and T. R. Chamlian, "Inter-tester reliability assessment of the volumetric measurement of the hand in subjects without any changes in their upper extremities," *Acta Fisiatrica*, vol. 17, no. 1, pp. 3–7, 2010.

[4] S. H. Ridner, L. Montgomery, J. Hepworth, B. Stewart, and J. Armer, "Comparison of upper limb volume measurement techniques and arm symptoms between healthy volunteers and individuals with known lymphedema," *Lymphology*, vol. 40, no. 1, pp. 35–46, 2007.

[5] M. d. A. Silva-Couto, C. L. Prado-Medeiros, A. B. Oliveira, C. C. Alcântara, A. T. Guimaraes, T. d. F. Salvini, R. Mattioli, and T. L. d. Russo, "Muscle atrophy, voluntary activation disturbances, and low serum concentrations of igf-1 and igfbp-3 are associated with weakness in people with chronic stroke," *Physical therapy*, vol. 94, no. 7, pp. 957–967, 2014.

[6] K. Wachal, K. Szmyt, and G. Oszkinis, "Diagnosis and treatment of a patient with type iv endoleak as a late complication after endovascular aneurysm repair," *Videosurgery and Other Miniinvasive Techniques*, vol. 9, no. 4, p. 667, 2014.

[7] K. Pirker, M. Rüther, H. Bischof, F. Skrabal, and G. Pichler, "Human body volume estimation in a clinical environment," 2010.

[8] C. Xu, Y. He, N. Khannan, A. Parra, C. Boushey, and E. Delp, "Image-based food volume estimation," in *Proceedings of the 5th international workshop on Multimedia for cooking & eating activities*, 2013, pp. 75–80.

[9] N. N. Kaashki, P. Hu, and A. Munteanu, "Deep learning-based automated extraction of anthropometric measurements from a single 3-d scan," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–14, 2021.

[10] Y. Su, W. Gao, Z. Liu, S. Sun, and Y. Fu, "Hybrid marker-based object tracking using kinect v2," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 9, pp. 6436–6445, 2020.

[11] W. Xie, P. X. Liu, and M. Zheng, "Moving object segmentation and detection for robust rgbd-slam in dynamic environments," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–8, 2020.

[12] P. W. Lo, Y. Sun, J. Qiu, and B. Lo, "Point2volume: A vision-based dietary assessment approach using view synthesis," *IEEE Transactions on Industrial Informatics*, 2019.

[13] P. Hu, N. N. Kaashki, V. Dadarlat, and A. Munteanu, "Learning to estimate the body shape under clothing from a single 3d scan," *IEEE Transactions on Industrial Informatics*, 2020.

[14] C.-Y. Chiu, D. L. Pease, S. Fawkner, and R. H. Sanders, "Automated body volume acquisitions from 3d structured-light scanning," *Computers in biology and medicine*, vol. 101, pp. 112–119, 2018.

[15] Y. Cui, W. Chang, T. Nöll, and D. Stricker, "Kinectavatar: fully automatic body capture using a single kinect," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 133–147.

[16] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan, "Scanning 3d full human bodies using kinects," *IEEE transactions on visualization and computer graphics*, vol. 18, no. 4, pp. 643–650, 2012.

[17] G. Wu, D. Li, Y. Zhong, and P. Hu, "A study on improving the calibration of body scanner built on multiple rgb-depth cameras," *International Journal of Clothing Science and Technology*, 2017.

[18] 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 *European Conference on Computer Vision*. Springer, 2016, pp. 561–578.

[19] N. Lunscher and J. Zelek, "Deep learning whole body point cloud scans from a single depth map," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1095–1102.

[20] J. Wells, I. Douros, N. Fuller, M. Elia, and L. Dekker, "Assessment of body volume using three-dimensional photonic scanning," *Annals of the New York Academy of Sciences*, vol. 904, no. 1, pp. 247–254, 2000.

[21] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[22] W. Yuan, T. Khot, D. Held, C. Mertz, and M. Hebert, "Pcn: Point completion network," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 728–737.

[23] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry, "3d-coded: 3d correspondences by deep deformation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 230–246.

[24] H. M. Clever, Z. Erickson, A. Kapusta, G. Turk, K. Liu, and C. C. Kemp, "Bodies at rest: 3d human pose and shape estimation from a pressure image using synthetic data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6215–6224.

[25] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7122–7131.

[26] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM transactions on graphics (TOG)*, vol. 34, no. 6, pp. 1–16, 2015.

[27] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 109–117.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.

[30] T. Helten, A. Baak, G. Bharaj, M. Müller, H.-P. Seidel, and C. Theobalt, "Personalization and evaluation of a real-time depth-based full body tracker," in *2013 International Conference on 3D Vision-3DV 2013*. IEEE, 2013, pp. 279–286.

[31] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll, "Detailed, accurate, human shape estimation from clothed 3d scan sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4191–4200.

[32] C. Zhang and T. Chen, "Efficient feature extraction for 2d/3d objects in mesh representation," in *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, vol. 3. IEEE, 2001, pp. 935–938.

[33] W. Zeng, W. Ouyang, P. Luo, W. Liu, and X. Wang, "3d human mesh regression with dense correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7054–7063.

Pengpeng Hu is currently an Assistant Professor at the Centre for Computational Science and Mathematical Modelling, Coventry University, Coventry, UK. Before joining Coventry University, he was a Senior Researcher at the Electronics and Informatics Department, Vrije Universiteit Brussel (VUB), Brussels, Belgium. In 2016, he was a Visiting Scholar with the School of Informatics of Edinburgh University, Edinburgh, U.K. In 2017, he was a Postdoctoral Fellow with the Computer and Information Sciences Department, Northumbria University, Newcastle upon Tyne, U.K. Since 2018, he had been at VUB. His research interests include biometrics, geometric deep learning, 3D human body reconstruction, point cloud processing, and measurement. He is the Early Career Advisory Board Member for the journals MEASUREMENT, MEASUREMENT: SENSORS, JOURNAL OF TEXTILE RESEARCH, and JOURNAL OF SILK. He is also the Editorial Member for the JOURNAL OF MODERN INDUSTRY AND MANUFACTURING and is the Topical Advisory Panel Member for the journals MDPI SENSORS and MDPI DESIGNS. He is the Guest Editor of the MDPI SENSORS, the Technical Support Chair of BMVC 2018, and a member of the Program Committee in SKIMA 2017, SKIMA 2018, and SKIMA 2019. He is the outstanding paper winner of the Emerald Literati Award 2019.

Xinxin Dai the M.E. degree from Guangdong University of Technology, Guangzhou, China, in 2019. She is currently a Ph.D. with the department of Electronics and Informatics Department, Vrije Universiteit Brussel (VUB), Ixelles, Belgium. Her research interests include point cloud processing, 3D hand reconstruction, biometrics, and person identification.

Ran Zhao received the M.E. degree in applied computer science from Vrije Universiteit Brussel(VUB) Ixelles, Belgium, in 2021, and the M.S. degree in mathematics from East China Normal University, Shanghai, China, in 2018. She is currently a Ph.D. with the Electronics and Informatics Department, Vrije Universiteit Brussel (VUB), Ixelles, Belgium. Her research focuses on point cloud processing, 3D hand scanning, and biometrics.

He Wang is an Associate Professor at the School of Computing, University of Leeds, UK. He is a Turing Fellow, the Director of High-Performance Graphics and Game Engineering and Academic Lead at the Centre for Immersive Technology at Leeds. His research interest is mainly in computer graphics, computer vision and machine learning. Previously he was a Senior Research Associate at Disney Research Los Angeles after receiving his PhD from the University of Edinburgh, UK. He is an Associate Editor of Computer Graphics Forum and has served in programme committees in more than 20 international conferences in computer graphics and computer vision.

Yingliang Ma is an Associate Professor at the School of Computing Science, University of East Anglia, UK. He has a PhD in Computer Science from University of Manchester. Ma's research areas mainly lie in the development of novel computer vision and computational geometric algorithms as well as applications in medical image analysis, image-guided surgery and procedure risk assessment. As a Principal Investigator (PI) or Co-PI, Dr Ma has been conducting research in relation to Computer Vision/Machine Learning/AI and parallel computing funded by EPSRC (UK) and NIHR. He has been invited to give a number of talks in different national and international institutions.

Adrian Munteanu is professor at the Electronics and Informatics (ETRO) department of the Vrije Universiteit Brussel (VUB), Belgium.He received the MSc degree in Electronics and Telecommunications from Politehnica University of Bucharest, Romania, in 1994, the MSc degree in Biomedical Engineering from University of Patras, Greece, in 1996, and the Doctorate degree in Applied Sciences from Vrije Universiteit Brussel, Belgium, in 2003. In the period 2004-2010 he was post-doctoral fellow with the Fund for Scientific Research-Flanders (FWO), Belgium, and since 2007, he is professor at VUB. Adrian Munteanu contributed to more than 400 publications and holds 7 patents. He is the recipient of the 2004 BARCO-FWO prize for his PhD work, the (co-)recipient of the Most Cited Paper Award from Elsevier for 2007. Adrian Munteanu served as Associate Editor for IEEE Transactions on Multimedia and currently serves as Associate Editor for IEEE Transactions on Image Processing.