AlignBodyNet: Deep Learning-based Alignment of Non-overlapping Partial Body Point Clouds from a Single Depth Camera

Pengpeng Hu, Edmond S.L. Ho, and Adrian Munteanu

***Abstract*—This paper proposes a novel deep learning framework to generate omnidirectional 3D point clouds of human bodies by registering the front- and back-facing partial scans captured by a single depth camera. Our approach does not require calibration-assisting devices, canonical postures, nor does it make assumptions concerning an initial alignment or correspondences between the partial scans. This is achieved by factoring this challenging problem into (i) building virtual correspondences for partial scans, and (ii) implicitly predicting the rigid transformation between the two partial scans via the predicted virtual correspondences. In this study, we regress the SMPL vertices from the two partial scans for building the virtual correspondences. The main challenges are (i) estimating the body shape and pose under clothing from single partial dressed body point clouds, and (ii) the predicted bodies from front- and back-facing inputs required to be the same. We, thus, propose a novel deep neural network dubbed AlignBodyNet that introduces shape-interrelated features and a shape-constraint loss for resolving this problem. We also provide a simple yet efficient method for generating real-world partial scans from complete models, which fills the gap in the lack of quantitative comparisons based on the real-world data for various studies including partial registration, shape completion, and view synthesis. Experiments based on synthetic and real-world data show that our method achieves state-of-the-art performance in both objective and subjective terms.**

***Index Terms*—Non-overlapping registration, ICP, Virtual correspondence, 3D scanning, Partial registration, Deep learning on point clouds**

I. Introduction

3D models of human bodies are key components for human-centric applications such as body measurement, healthcare, computer animation, virtual try-on, and virtual reality. 3D scanning technologies are popular means for acquiring accurate and realistic 3D human models. Traditional 3D scanning systems (e.g. laser scan or structured light) are very expensive and usually require expert knowledge for operation and data acquisition. With the advent of commodity depth cameras such as Microsoft Kinect and Intel Realsense, low-cost 3D body scanning becomes a new trend with potential for wide-scale deployments in numerous applications [1], [2].

According to the number of employed depth cameras, existing depth-based 3D scanning systems can be mainly classified into two main categories: multi-camera [3], [4], [5] and single-camera scanning systems [6], [7], [11] respectively. Multi-camera systems set several depth cameras at various predefined positions around the subject. Partial scans from each depth camera are aligned together to obtain omnidirectional body point clouds. Such systems are not convenient for home usage, and heavily rely on the quality of extrinsic calibration. In contrast, single-camera scanning brings stronger operability due to its good flexibility and ease-of-use characteristics. Unlike multi-camera scanning methods that only deal with multiple partial scans, single-camera scanning methods make use of single [8], dual [11], or more partial scans [7]. For single / dual partial scan-based methods, parametric body models are usually deformed to fit the acquired partial body point clouds [11]. However, detailed characteristics of the subject such as facial features and hairs cannot be preserved due to the limited subspace of the employed parametric body model. For multiple partial scan-based systems, a depth camera moves around the subject yielding geometry information of different body parts. Omnidirectional body point clouds are then generated by rigidly/non-rigidly aligning the resulting set of partial scans. However, such systems require low-speed and stable camera motion around the subject to avoid jittering effects. This cannot be always guaranteed for handheld scanning devices and the fusion process of the resulting partial point clouds has to be aborted when the camera tracking fails. Moreover, a global optimization [9], [10] is required and many redundant depth images are employed in data fusion, which increases the computational burden and is prone to registration inaccuracies.

To solve the above problems, we propose a novel deep learning approach to reconstruct omnidirectional 3D body point clouds from two partial body scans captured by a single depth camera. The main contributions of this work can be summarized as follows:

- A novel deep learning method for omnidirectional 3D point cloud reconstruction of human bodies is proposed that makes use of a two partial body scans obtained with a single depth camera.
- Novel shape-interrelated features and a novel shape-constraint loss are proposed, enhancing the performance of the proposed method.
- A simple yet efficient method is proposed to generate real-world partial point clouds from complete models, filling the gap in the lack of quantitative comparisons based on real-world data for various studies including

P. Hu (corresponding author) is with the Department of Electronics and Informatics,Vrije Universiteit Brussel, Brussels, Belgium, email: phu@etrovub.be.

E.S.L Ho is with School of Computing Science, University of Glasgow, Scotland, United Kingdom.

A. Munteanu is with the Department of Electronics and Informatics, Vrije Universiteit Brussel, Brussels, Belgium.

partial registration, shape completion, and view synthesis. It also has potential for generating large-scale datasets to train single- and multi-view algorithms when labeled partial scans are difficult to be collected.

II. Related works

A. Parametric body fitting

Automatic fitting of 3D human body models to point clouds is a classic task in computer vision and graphics. 3D scanning methods can yield high-quality body models. However, noises and holes cannot be avoided in scanned bodies due to device limitation and self-occlusion. To address these problems, researchers proposed to build parametric body models by fitting a template mesh to a range of scanned bodies [12], [13], [14]. The invention of parametric body model provides the foundation for solving many challenging tasks. For instance, [15] proposed an automatic rigging method by matching the pre-rigged parametric body to the 3D scan and then transferring the skeleton and skin weights from the fitted paramtric body to the scan. [16] proposed a deep learning-based method for estimating body shape and pose under clothing by fitting the SMPL model to a dressed body scan. [11] proposed to estimate the SMPL shape parameters from two partial dressed/undressed scans. However, these methods cannot preserve detailed information about the subjects such as facial features and hair due to inherent use of a parametric body model. [17] proposed to combine implicit function learning and parametric models for 3D human reconstruction. This method can capture garment, face and hair details only when complete body point clouds are available. Unlike these works that use the parametric body to represent the reconstructed body shapes, we aim at registering two raw partial scans of subjects. In our proposed method, we fit the SMPL model to the front- and back-facing partial point clouds simultaneously for the purpose of building virtual correspondences between the two partial inputs.

B. 3D Shape Rigid Registration

In 3D scanning, the key challenge is the registration of partial point clouds [18], [19]. Its goal is to find the optimal transformations that align the input partial point clouds. Existing registration methods can be mainly classified into two categories: ICP-based and deep learning-based.

The pioneering iterative closest point algorithm (ICP) [20] and many ICP-based variants (see e.g. [21]) have been proposed in the literature, all aiming to improve accuracy and robustness in the registration process. However, the performance of ICP-based methods highly depends on two assumptions: good initial alignment and sufficient overlap area between source and target shapes. Such assumptions enable a proper initialization for the iterative optimization by using the nearest neighbors as correspondences. If accurate correspondences are known, the registration can be well performed without the need of good initial alignment. To this end, a lot of researchers pay an attention to finding correspondences. [22] trained a deep convolutional neural network to find dense correspondences between partial human body scans by predicting the segmentation label for each point in the depth images. [23] proposed a two-step method. First, the authors trained a neural network to deform a body template to fit two complete body scans. Next, by searching nearest neighbors between the two inputs and the deformed template, dense correspondences between two inputs can be implicitly obtained. However, these methods fail for our task as they assume that correspondences actually exist between inputs. It should be noted that none of correspondences exist for non-overlapping front- and back-facing partial scans. What we aim at resolving is a more challenging problem: non-overlapping partial registration. Unlike but inspired by existing correspondence-finding works, we propose to estimate virtual correspondences for registration.

Recently, deep learning has been introduced to deal with registration problems. For example, [24] proposed Deep Closet Point (DCP) that uses a point cloud feature encoder, an attention-based module and a differentiable singular value decomposition (SVD) to predict rigid transformations for point clouds. However, an additional iteration is required to refine the results of DCP. The authors of [25] proposed Deep Global Registration (DGR) consiting of three modules: correspondence confidence prediction, pose estimation, and pose refinement. [26] proposed a deep learning-based approach to register multi-view 3D point clouds by firstly estimating the initial pairwise registration and then performing a globally consistent refinement. However, these methods are not designed for non-overlapping partial point cloud registration.

The closest work to ours is the recently published method in [27]. It predicted two completed bodies for registration from two partial point clouds, respectively. However, this method was not designed for dressed bodies, and it was only tested on the synthetic data, which cannot show its effectiveness in practice. Furthermore, this method ignored the fact that the partial scans were actually captured from the same subject. To resolve these problems, we propose a novel approach for registering two non-overlapping dressed body scans, and validate the effectiveness of proposed method on the real-world data.

III. Proposed Method

A. Problem Statement

As Figure 1 shows, the rigid registration consists of three scenarios: high-, low- and non-overlapping scenarios. The majority of existing methods focus on the point cloud registration with high overlap [20] while a few of methods attempted to deal with the registration with low overlap [28]. Unlike these methods that depends on the overlap, in this article, we focus on a more challenging problem, namely rigidly registering two non-overlapping partial point clouds of human bodies.

We use X and Y to denote the front- and back-facing partial body point clouds, respectively. Note that X and Y can be noisy or clean, and they are allowed to have the same or different number of points. The goal is to find an optimal rigid transformation to align X with Y. The rigid transformation is denoted as $[R_{xy}, t_{xy}]$, where $R_{xy} \in SO(3)$ and $t_{xy} \in \mathbb{R}^3$ represent the rotation matrix and the translation

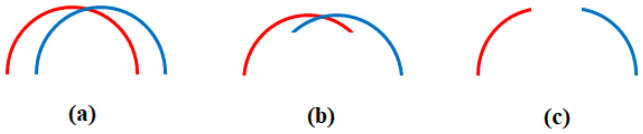

Fig. 1. Schematic diagrams of different overlapping scenarios:(a) high-overlapping, (b) low-overlapping, (c) non-overlapping. Red and blue indicates the source and target, respectively.

vector, respectively. In ICP-based methods, the registration problem can be solved by minimizing the following loss:

$$Loss = \frac{1}{Q}\sum_{i=0}^{Q-1} ||R_{xy} \cdot x_i + t_{xy} - y_{c(x_i)}||^2 \quad (1)$$

where Q is the number of points in X and $c(.)$ is a mapping function that establishes the correspondences in Y for each point from X. Intuitively, this approach fails for our task because no correspondences between the front- and back-facing partial body point clouds exist due to the lack of an overlap area between the scans. Inspired by the work of [27], we address this task by converting it to a problem of finding virtual correspondences, and rewrite equation (1):

$$Loss = ||R_{xy} \cdot \tau(X) + t_{xy} - \zeta(Y)||^2 \quad (2)$$

where $\tau()$ and $\zeta()$ are two mappings interpreting partial body point clouds X, Y to complete body point clouds $\tau(X)$, $\zeta(Y)$, respectively. Note that $\tau(X)$ and $\zeta(Y)$ represent the same body shape, and more importantly, they have the save point order. Therefore, intuitively, virtual correspondences $(\tau(X), \zeta(Y))$ can be obtained. Next, the transformation $[R_{xy}, t_{xy}]$ is directly computed using the normal equation. Our method is summarized in Figure 2. Two non-overlapping partial scans are fed into a deep neural network (DNN) to predict virtual correspondences. Rotation R and translation t are computed based on the virtual correspondences. Compared to ICP-based methods, such an approach does not require assumptions regarding initial alignments or necessary overlap areas, and avoids an expensive iterative refinement procedure. Its key step is to learn $\tau()$. Specifically, $\tau(X)$ and $\tau(Y)$ should represent absolutely the same body shape. Our main insights are: (i) $\tau()$ and $\zeta()$ should be learned in an joint manner, (ii) the features of X and Y should have a communication, and (iii) there must be a constraint to force $(\tau(X)$ to be close to $\zeta(Y)$. Therefore, we design a novel two-stream encoder-decoder network architecture depicted by Figure 3.

B. Shape-interrelated features

Motivated by our first insight, we design a two-stream encoder-decoder architecture as the backbone network. Given the front-facing partial body point cloud X and the back-facing partial body point cloud Y, the first task is to extract features. We used a simplified PointNet [29] encoder to act as our feature extractor due to its effectiveness and simplicity. Note that our two feature extractors have the same architecture but with different weights. Specifically, shared MLPs are employed to learn per-point feature matrices G_X and G_Y

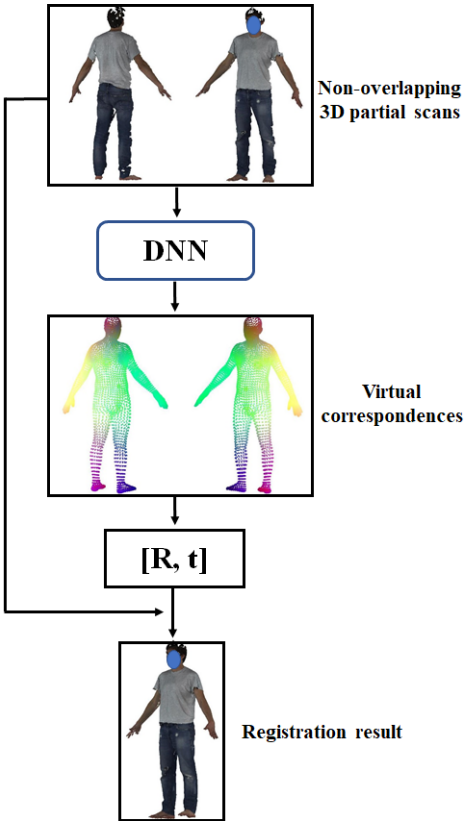

Fig. 2. Method overview. Given non-overlapping front- and back-facing partial body point clouds, we leverage the proposed novel network to predict virtual correspondences for calculating the transformation between the two partial inputs.

(each row represents the feature of each point). Next, two global features v_X and v_Y are obtained by passing G_X and G_Y to the point-wise maxpooling layers. Despite there are many more advanced feature extractor candidates [30], [31], [32], experimental results show the simple encoder can work well in our study, and to compare the performances of different encoders is not the target of this study.

Revisiting our second insight, v_X and v_Y require a communication. To this end, v_X and v_Y are jointly concatenated to G_X and G_Y to obtain two augmented per-point feature matrices M_X and M_Y. By such a simple yet efficient feature-fusion strategy, M_X contains the information from Y while M_Y contains the information from X. Finally, M_X and M_Y are processed by two more shared MLPs and point-wise maxpooling layers to generate the shape-interrelated features f_X and f_Y.

C. Virtual Correspondence Prediction

None of correspondences exist between X and Y due to the lack of overlapping area. We, thus, define the virtual correspondences $(\tau(X), \zeta(Y))$ based on the parametric body vertices. Therefore, shape-interrelated features require to be interpreted to parametric body vertices. MLP is used to this end in [27]. However, the two outputs can have large shape and pose variations. We argue there are two main reasons for this phenomenon: (i) in [27] that authors predicted complete

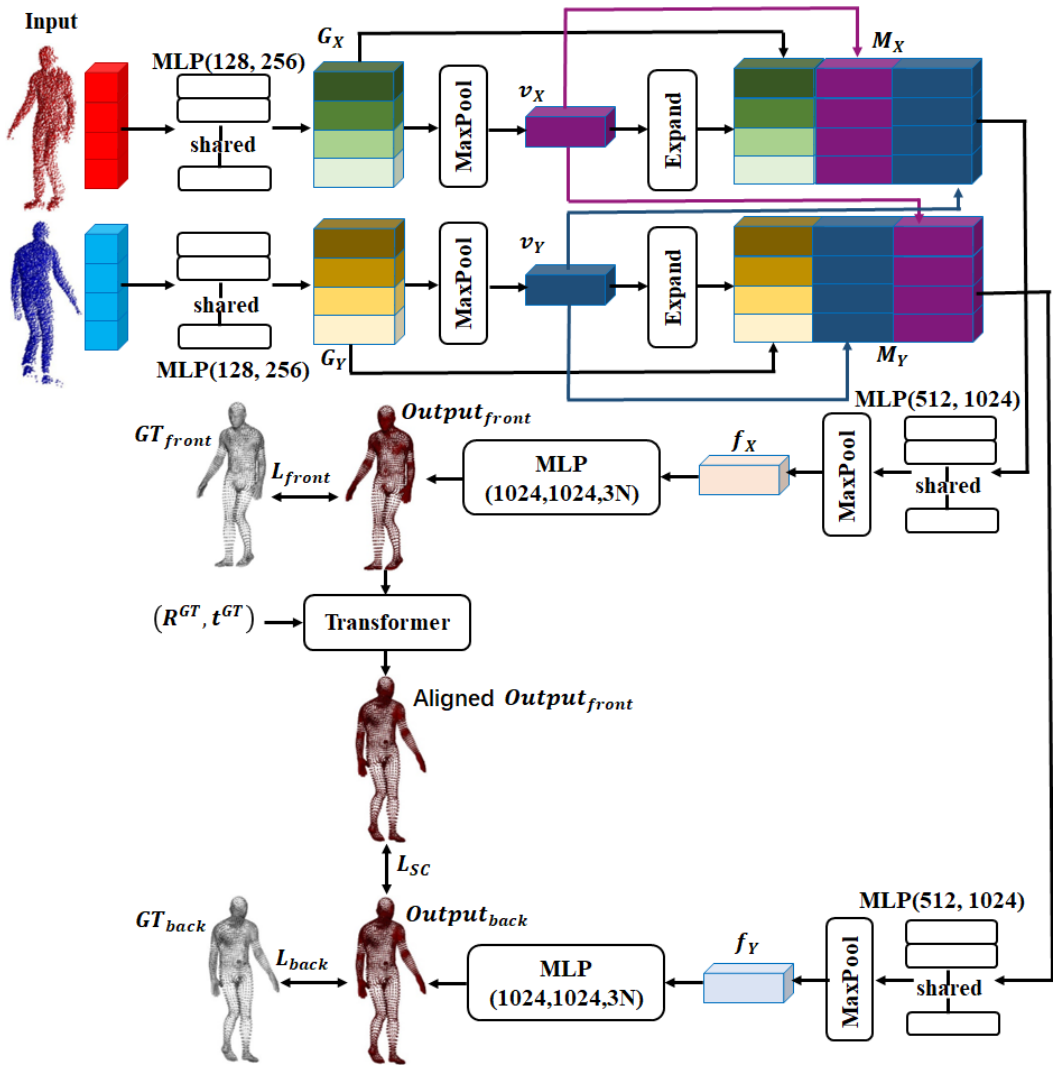

Fig. 3. Proposed network architecture.

body shapes from single partial point clouds, which is an ill-posed problem; and (ii) for the same subject, the body shape should be the same no matter if it is being observed from front- or back-facing views (our third insight).

To address this problem, we followed a two-fold strategy. Firstly, we proposed the above two-stream encoder and feature fusion strategy. The communication information between X and Y can be passed to the decoder. Secondly, we added a Transformer to align the two outputs $output_{front}$ and $output_{back}$ to efficiently compare the error of predicted parametric body vertices from the front- and back-facing partial body point clouds. Despite of the fact that many advanced decoder candidates are available (e.g. [33], [16]), we use the same MLP decoder as in [27] in order to validate our idea and to fairly compare our method with the work of [27]. Main conceptual improvements over [27] include (i) a two-stream decoder architecture, and (ii) we add a Transformer that transforms $\tau(X)$ from the front-facing view-based coordinate to the back-facing view-based coordinate by the ground-truth transformation. This Transformer plays an important role in our study as we can enforce a powerful constraint $\tau(X) = \zeta(Y)$. Note that the Transformer is necessary in the training phase, but does not contribute in the inference phase. This property is intuitive as no ground-truth transformation is available in the inference phase.

D. Loss function

We propose a customized loss function for efficiently supervising the learning of the proposed network. It consists of three terms: front-facing vertex loss, back-facing vertex loss, and a consistency loss.

Front-facing Vertex Loss. From the front-facing partial point clouds of bodies, our network outputs SMPL vertices, which are aligned with the front-facing partial point cloud. The prediction error is computed by comparing the ground truth against the reconstructed body. We define the front-facing vertex loss as:

$$L_{front} = \frac{1}{N}\sum_{i=1}^{N} ||\tau(X)_i - \tau(X)_i^{GT}||^2 \quad (3)$$

where $\tau(X)_i$ represents the i^{th} vertex of the reconstructed body and $\tau(X)_i^{GT}$ represents the ground-truth vertex of $\tau(X)_i$.

Back-facing Vertex Loss. Similar to L_{front}, our network also outputs SMPL vertices from the back-facing partial point clouds. We, thus, define the back-facing vertex loss as:

$$L_{back} = \frac{1}{N}\sum_{i=1}^{N} ||\zeta(Y)_i - \zeta(Y)_i^{GT}||^2 \tag{4}$$

Consistency Loss. Since the two partial point clouds are obtained by scanning the same subject, we conclude that $\tau(X) = \zeta(Y)$. Note that $\tau(X)$ and $\zeta(Y)$ cannot be directly compared as they are not aligned. Thanks to the proposed Transformer, we define a shape-shared loss to constrain the variations between the two reconstructed bodies:

$$L_{SC} = \frac{1}{N}\sum_{i=1}^{N} ||\tau(X)_i - \zeta(Y)_i||^2 \tag{5}$$

Complete Loss. Our complete loss is defined as:

$$Loss = \alpha \cdot L_{front} + \beta \cdot L_{back} + \omega \cdot L_{SC} \tag{6}$$

where α, β and ω are the weights that control the contributions of each term.

E. Registration

Once our network is trained, virtual correspondences $(\tau(X), \tau(Y))$ are obtained. Equation 2 is rewritten as:

$$\begin{bmatrix} R_{xy} & t_{xy} \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} \tau(X) \\ ones(N) \end{bmatrix} = \begin{bmatrix} \tau(Y) \\ ones(N) \end{bmatrix} \tag{7}$$

where $ones(N)$ represents the operation that creates a row vector filled with N $ones$. By normal equation, the transformation can be directly obtained:

$$\begin{aligned} \begin{bmatrix} R_{xy} & t_{xy} \\ 0 & 1 \end{bmatrix} = &\left(\begin{bmatrix} \tau(Y) \\ ones(U) \end{bmatrix} \times \begin{bmatrix} \tau(X) \\ ones(U) \end{bmatrix}^T \right) \\ &\times \left(\begin{bmatrix} \tau(X) \\ ones(U) \end{bmatrix} \times \begin{bmatrix} \tau(X) \\ ones(U) \end{bmatrix}^T \right)^{-1} \end{aligned} \tag{8}$$

IV. Experimental Results

A. Training dataset and setup

Considering that human beings wear clothes in the real life, we trained our model on the BUG (**B**ody **U**nder virtual **G**arments) dataset [16]. BUG is a large-scale synthetic dressed body dataset consiting of $100K$ male and $100K$ female dressed bodies, realistic dressed body scans and ground-truth body shapes in motion. Same as [27], simulated partial scans and ground-truth transformation are obtained. We randomly selected $99.5K$ male samples for training our model and $0.5K$ male samples for the test. The training is carried out using the Adam optimizer [34] 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 [35]. We set $\alpha = 1$, $\beta = 1$ and $\omega = 1$ in the loss.

B. Results on the PDT13 data

Despite our model is trained merely on the synthetic data, it is designed for dealing with the real-world data. To validate its effectiveness for the real-world data, we test the proposed algorithm on the PDT13 dataset [36]. The PDT13 dataset consists of front- and back-facing scans of subjects obtained using a Kinect camera. Figure 4 depicts our results, and it can be seen that the non-overlapping two partial body scans can be visually well aligned even noises exists.

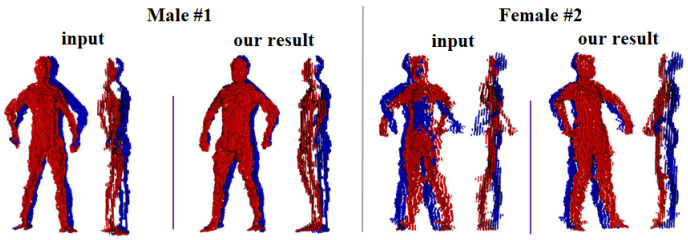

Fig. 4. Results on the PDT13 data using our method. The front- and back-facing partial body point clouds are denoted by the red (the source) and blue (the target) color, respectively.

C. Comparisons

In this experiment, we compare our algorithm against popular ICP [20], recent deep learning-based registration methods (deep global registration (DGR) [25], non-overlapping partial registration (NO-PR) [27]), the recent partial-based parametric body fitting method IP-Net [17], and the template-based body fitting method 3D-CODED [23]. Given the ground truth rotation R^{GT} and translation t^{GT}, the rotation error RE and translation error TE are defined as:

$$RE(R, R^{GT}) = \arccos\left(\frac{trace(R^{-1}R^{GT}) - 1}{2}\right) \tag{9}$$

$$TE(t, t^{GT}) = ||t - t^{GT}|| \tag{10}$$

where R and t represent the estimated rotation matrix and translation vector, respectively.

D. Results on the BUG data

We first perform the comparisons based on the synthetic data. We randomly select 50 samples from the unseen testing BUG data. Figure 5 and Table I compare the results of different methods. It can be seen that our method outperforms the other methods.

E. Results on the BUFF data

For the quantitative evaluation, the predicted transformation should be compared with the ground-truth transformation. However, no usable real-world dataset containg the ground-truth transformation exists in the literature. One potential solution is to scan the subjects by calibrated dual Kinect cameras facing each other. However, calibration errors cannot be avoided that may result in unfair comparisons; in addition, it is expensive and time-consuming to scan many subjects in

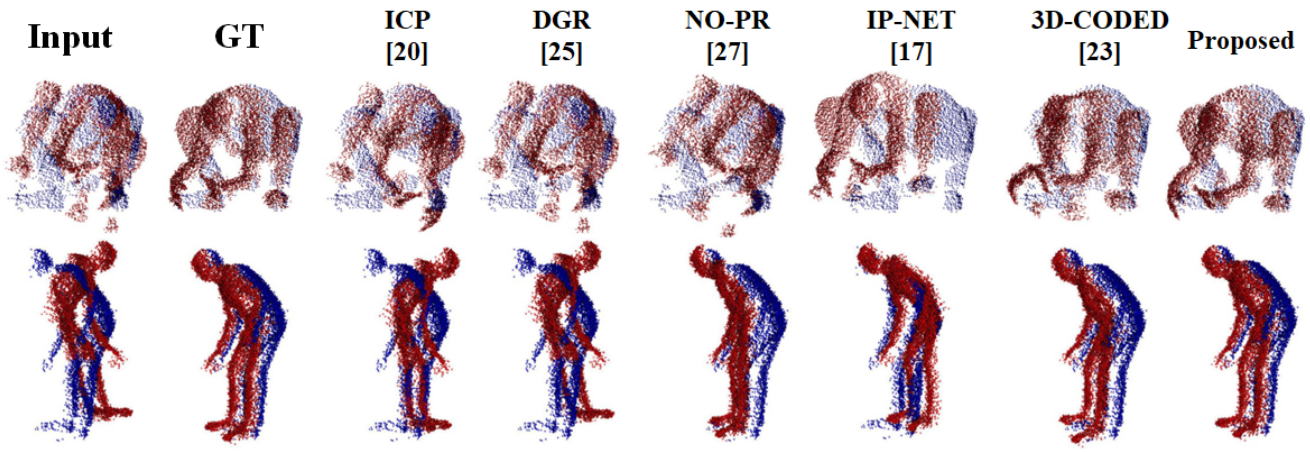

Fig. 5. Comparison with different methods based on BUG data.

TABLE I
COMPARISON OF AVERAGE ROTATION AND TRANSLATION ERRORS WITH DIFFERENT METHODS BASED ON THE UNSEEN BUG DATA.

Methods	ICP[20]	DGR [25]	NO-PR [27]	IP-Net [17]	3D-CODED [23]	Our method
RE	162.056°	162.012°	4.254°	20.279°	5.6°	1.702°
TE	186.222 mm	171.056 mm	13.429 mm	87.213 mm	29.02 mm	7.21 mm

order to generate a real-world body dataset with partial scans. To alleviate this problem, we made use of the fact that many scanned body models are publicly available. We propose thus a simple yet efficient approach dubbed RealParialScan to extract partial body point clouds directly from these scanned body models, as shown in Algorithm 1. RealPartialScan provides a step towards generating the large-scale real-world data for training and quantitatively evaluating deep learning-based algorithms that take requires partial data as input, including shape completion [37], partial registration [27], view synthesis [38], and multi-view tasks [39], [40]. Figure 6 shows an example of the obtained partial scans using our method.

Algorithm 1 Real-world partial body point cloud generation algorithm.

Input:
S: complete scanned body meshes or point clouds;
Output:
P^{GT}: real-world partial body point clouds;
1: rendering partial point clouds P from S using a rendering system (e.g. Blender)
2: for each point x in S, find its closet point y in P
3: if $dist\,(x, y) < threshold$
4: $x \in P^{GT}$
5: **return** P^{GT};

To quantitatively compare our algorithm against related methods, we treat the front- and back-facing partial body point clouds as the source and the target, respectively. Figure 7 depicts the visual comparisons on the BUFF data [41], and Table II illustrates the registration errors. It can be seen that ICP and DCR methods fail to perform the registration when no overlapping exists. Our method achieves the best performance. More results are illustrated in Figure 8.

F. Ablation Study

To verify the effect of each proposed component, we performed an ablation study based on test data containing 500 samples that are not included in the training phase.

Shape-interrelated Features. As Table III shows, the proposed strategy of offering communication between two partial inputs can reduce the rotation error and the translation error.

Consistency Loss. The proposed transformer is used to create the shape constraint for the two output shapes by minimizing their per-vertex errors. Therefore, the transformer works together with the designed shape-consistent loss. Table IV shows that the proposed consistency loss reduce the registration error.

V. CONCLUSIONS

We proposed a novel deep learning method for reconstructing omnidirectional body point clouds by aligning two non-overlapping partial body scans acquired with a single Kinect camera. A novel two-stream encoder-decoder network architecture, shape-interrelated features and a shape-constraint loss are proposed. Our model was trained on a synthetic dataset but it generalizes well to unseen real-world data. Experimental results show that our method outperforms state-of-the-art approaches. In the future, we are interested to extend our work to multi-view non-rigid body point cloud registration, and study the effect of overlap ratios on the registered result.

VI. ACKNOWLEDGEMENTS

This work was supported in part by the Innoviris under Project AI43D in close collaboration with Spenty's and in part by FWO under Project G084117.

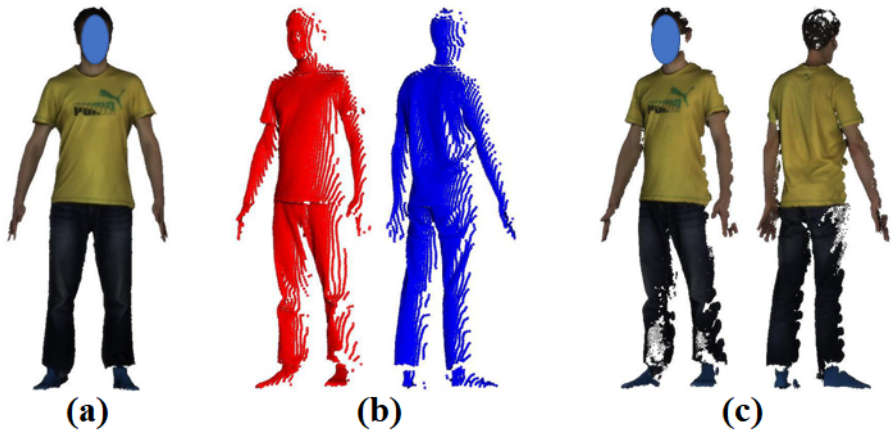

Fig. 6. An example of generating partial body scans using our method: (a) The complete scanned body model, (b) Synthetic partial scans, (c) Real-world partial scans using our method.

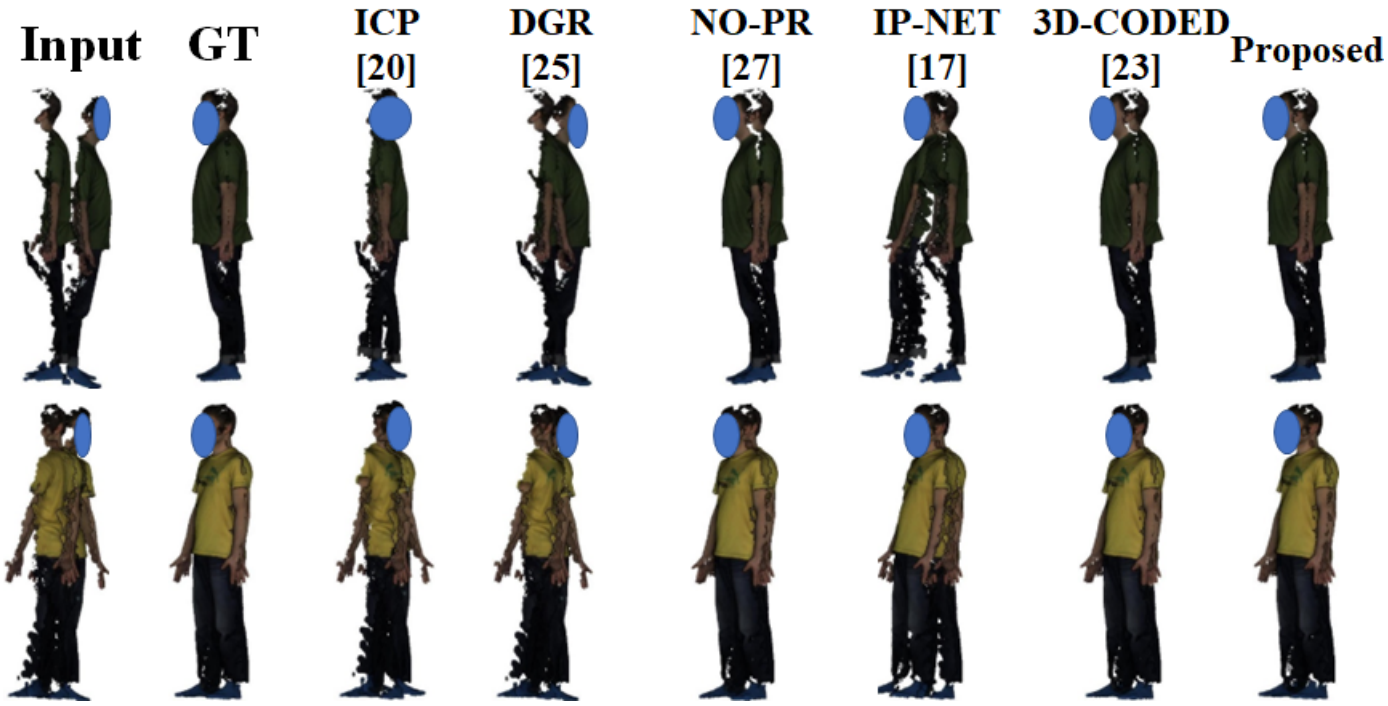

Fig. 7. Comparison with different registration methods based on BUFF data.

TABLE II
COMPARISON OF AVERAGE ROTATION AND TRANSLATION ERRORS WITH DIFFERENT REGISTRATION METHODS BASED ON BUFF DATA.

Methods	ICP [20]	DGR [25]	NO-PR [27]	IP-Net [17]	3D-CODED [23]	Our method
RE	160.035°	160.047°	1.759°	20.894°	2.06°	0.951°
TE	386.1 mm	389.3 mm	60.5 mm	504.3 mm	54.78 mm	34 mm

TABLE III
ABLATION STUDY ON THE PROPOSED SHAPE-INTERRELATED FEATURES.

Feature	μ , RE	σ , RE	μ , TE	σ , TE
With feature fusion	1.515°	1.637°	6.001 mm	4.156 mm
Without feature fusion	2.087°	2.261°	9.835 mm	7.456 mm

REFERENCES

[1] 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.

[2] J.-M. Lu and M.-J. J. Wang, "The evaluation of scan-derived anthropometric measurements," *IEEE Transactions on instrumentation and*

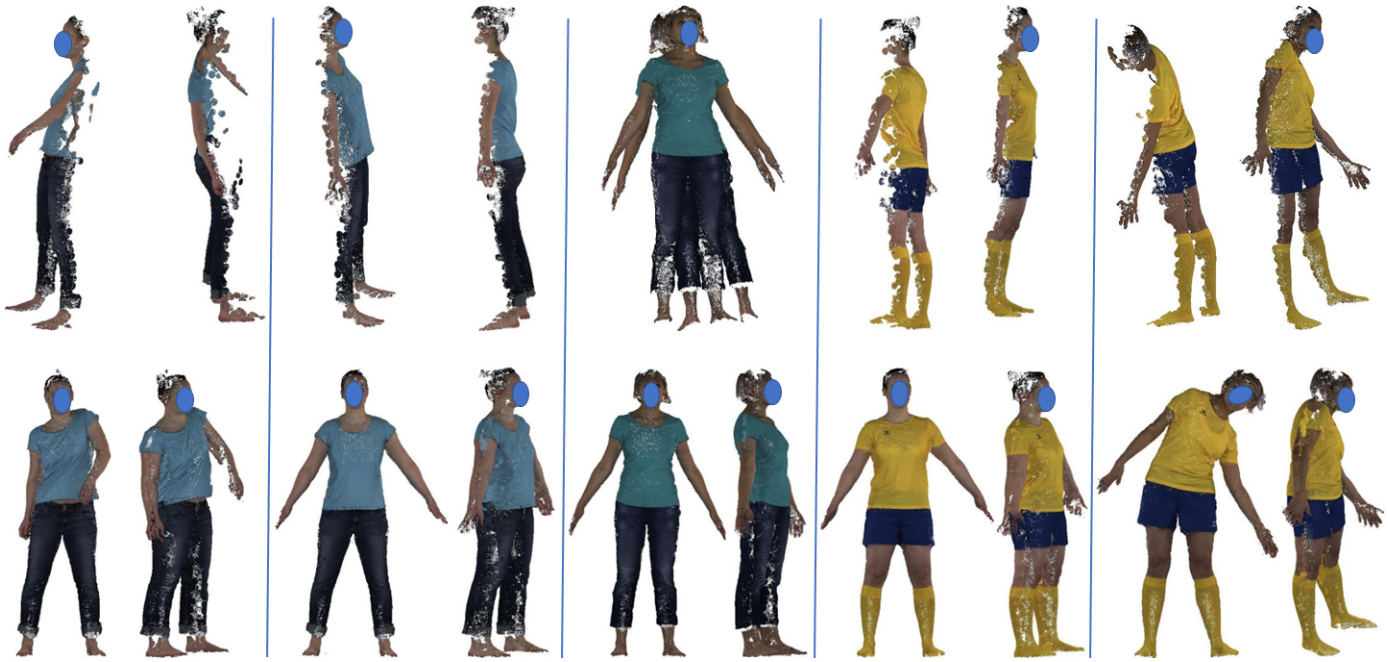

Fig. 8. More results using the proposed method based on BUFF data. Top: source and target partial non-overlapping scans. Bottom: Registration results using the proposed method.

TABLE IV
ABLATION STUDY ON THE PROPOSED SHAPE-CONSTRAINT LOSS.

Loss	μ , RE	σ , RE	μ , TE	σ , TE
$L_{front} + L_{back}$ (without Transformer)	1.694°	1.640°	6.850 mm	4.680 mm
$L_{front} + L_{back} + L_{SC}$ (with Transformer)	1.515°	1.637°	6.001 mm	4.156 mm

measurement, vol. 59, no. 8, pp. 2048–2054, 2010.

[3] 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.

[4] M. Kowalski, J. Naruniec, and M. Daniluk, "Livescan3D: A fast and inexpensive 3D data acquisition system for multiple Kinect V2 sensors," in *2015 international conference on 3D vision*. IEEE, 2015, pp. 318–325.

[5] Z. Liu, J. Huang, S. Bu, J. Han, X. Tang, and X. Li, "Template deformation-based 3d reconstruction of full human body scans from low-cost depth cameras," *IEEE Transactions on Cybernetics*, vol. 47, no. 3, pp. 695–708, 2016.

[6] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison *et al.*, "KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011, pp. 559–568.

[7] 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.

[8] 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.

[9] R. Huang, Y. Xu, W. Yao, L. Hoegner, and U. Stilla, "Robust global registration of point clouds by closed-form solution in the frequency domain," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 171, pp. 310–329, 2021.

[10] Q.-Y. Zhou, J. Park, and V. Koltun, "Fast global registration," in *European conference on computer vision*. Springer, 2016, pp. 766–782.

[11] P. Hu, E. S.-l. Ho, and A. Munteanu, "3DBodyNet: Fast reconstruction of 3D animatable human body shape from a single commodity depth camera," *IEEE Transactions on Multimedia*, 2021.

[12] B. Allen, B. Curless, and Z. Popović, "The space of human body shapes: reconstruction and parameterization from range scans," *ACM transactions on graphics (TOG)*, vol. 22, no. 3, pp. 587–594, 2003.

[13] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3d deformation model for tracking faces, hands, and bodies," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8320–8329.

[14] 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.

[15] A. Feng, D. Casas, and A. Shapiro, "Avatar reshaping and automatic rigging using a deformable model," in *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, 2015, pp. 57–64.

[16] 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.

[17] B. L. Bhatnagar, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "Combining implicit function learning and parametric models for 3d human reconstruction," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 311–329.

[18] Y. Cui, Y. An, W. Sun, H. Hu, and X. Song, "Memory-augmented point cloud registration network for bucket pose estimation of the intelligent mining excavator," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[19] Y. Wang, Y. Liu, Q. Xie, Q. Wu, X. Guo, Z. Yu, and J. Wang, "Density-invariant registration of multiple scans for aircraft measurement," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–15, 2020.

[20] P. J. Besl and N. D. McKay, "Method for registration of 3D shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. International Society for Optics and Photonics, 1992, pp. 586–606.

[21] S. Rusinkiewicz and M. Levoy, "Efficient variants of the ICP algorithm," in *Proceedings third international conference on 3D digital imaging and modeling*. IEEE, 2001, pp. 145–152.

[22] L. Wei, Q. Huang, D. Ceylan, E. Vouga, and H. Li, "Dense human body

correspondences using convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1544–1553.
[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] Y. Wang and J. M. Solomon, "Deep closest point: Learning representations for point cloud registration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3523–3532.
[25] C. Choy, W. Dong, and V. Koltun, "Deep global registration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2514–2523.
[26] Z. Gojcic, C. Zhou, J. D. Wegner, L. J. Guibas, and T. Birdal, "Learning multiview 3D point cloud registration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1759–1769.
[27] P. Hu and A. Munteanu, "Method for registration of 3D shapes without overlap for known 3D priors," *Electronics Letters*, vol. 57, no. 9, pp. 357–359, 2021.
[28] S. Chen, L. Nan, R. Xia, J. Zhao, and P. Wonka, "Plade: A plane-based descriptor for point cloud registration with small overlap," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 4, pp. 2530–2540, 2019.
[29] 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.
[30] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *arXiv preprint arXiv:1706.02413*, 2017.
[31] P. Guerrero, Y. Kleiman, M. Ovsjanikov, and N. J. Mitra, "Pcpnet learning local shape properties from raw point clouds," in *Computer Graphics Forum*, vol. 37, no. 2. Wiley Online Library, 2018, pp. 75–85.
[32] Y. Ben-Shabat and S. Gould, "Deepfit: 3d surface fitting via neural network weighted least squares," in *European Conference on Computer Vision*. Springer, 2020, pp. 20–34.
[33] Y. Yang, C. Feng, Y. Shen, and D. Tian, "Foldingnet: Point cloud auto-encoder via deep grid deformation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 206–215.
[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[35] 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.
[36] 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.
[37] 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.
[38] E. Penner and L. Zhang, "Soft 3d reconstruction for view synthesis," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, pp. 1–11, 2017.
[39] K. Gupta, S. Jabbireddy, K. Shah, A. Shrivastava, and M. Zwicker, "Improved modeling of 3d shapes with multi-view depth maps," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, pp. 71–80.
[40] X. Long, L. Liu, W. Li, C. Theobalt, and W. Wang, "Multi-view depth estimation using epipolar spatio-temporal networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8258–8267.
[41] 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.

Pengpeng Hu is currently a Researcher at the Electronics and Informatics Department, Vrije Universiteit Brussel (VUB), Ixelles, Belgium. In 2016, he was a Visiting Scholar with the School of Informatics of the 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 has 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 journals of MEASUREMENT, MEASUREMENT: SENSORS, JOURNAL OF TEXTILE RESEARCH, and JOURNAL OF SILK. He is also the Editorial Member for JOURNAL OF MODERN INDUSTRY AND MANUFACTURING and is the Topical Advisory Panel Member for journals of MDPI SENSORS and MDPI DESIGNS. He was the Guest Editor of the MDPI SENSORS, the Technical Support Chair of BMVC 2018, and the member of Program Committee in SKIMA 2017, SKIMA 2018, and SKIMA 2019. He is the outstanding paper winner of the Emerald Literati Award 2019.

Edmond S.L. Ho is currently a Senior Lecturer at School of Computing Science, University of Glasgow. He was the Programme Leader for BSc (Hons) Computer Science and an Associate Professor in the Department of Computer and Information Sciences at Northumbria University, Newcastle, UK. Prior to joining Northumbria University in 2016, he was a Research Assistant Professor in the Department of Computer Science at Hong Kong Baptist University. He received the BSc degree in Computer Science from the Hong Kong Baptist University, the MPhil degree from the City University of Hong Kong, and the PhD degree from the University of Edinburgh. His research interests include Computer Graphics, Computer Vision, Robotics, Motion Analysis, and Machine Learning.

Adrian Munteanu received the M.Sc. degree in electronics and telecommunications from the Politehnica University of Bucharest, Bucharest, Romania, in 1994, the M.Sc. degree in biomedical engineering from the University of Patras, Patras, Greece, in 1996, and the Doctorate degree in applied sciences from Vrije Universiteit Brussel, Ixelles, Belgium, in 2003. He is currently a Professor with the Electronics and Informatics Department of the Vrije Universiteit Brussel. In the period 2004–2010, he was a Postdoctoral Fellow with the Fund for Scientific Research - Flanders (FWO), Belgium, and since 2007, he has been a Professor with VUB. He made contribution to more than 400 publications and holds seven patents. He was the recipient of the 2004 BARCO-FWO Prize for the Ph.D. work, the Co-Recipient of the Most Cited Paper Award from Elsevier for 2007. He was an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA and is currently an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING.