PatientHandNet: 3-D Open-Palm Hand Reconstruction From Sparse Multiview Depth Images

Dai, X., Zhao, R., Hu, P. & Munteanu, A.

Original citation & hyperlink:
Dai, X, Zhao, R, Hu, P & Munteanu, A 2023, 'PatientHandNet: 3-D Open-Palm Hand Reconstruction From Sparse Multiview Depth Images', IEEE Transactions on Instrumentation and Measurement, vol. 72, 5023613.
https://dx.doi.org/10.1109/tim.2023.3304706

DOI 10.1109/tim.2023.3304706
ISSN 0018-9456
ESSN 1557-9662

Publisher: Institute of Electrical and Electronics Engineers

PatientHandNet: 3D Open-palm Hand Reconstruction from Sparse Multi-view Depth Images

Xinxin Dai, Ran Zhao, Pengpeng Hu, and Adrian Munteanu

***Abstract*—Accurately reconstructing 3D hand shapes of patients is important for immobilization device customization, artificial limb generation, and hand disease diagnosis. Traditional 3D hand scanning requires multiple scans taken around the hand with a 3D scanning device. These methods require the patients to keep an open-palm posture during scanning, which is painful or even impossible for patients with impaired hand functions. Once multi-view partial point clouds are collected, expensive post-processing is necessary to generate a high-fidelity hand shape. To address these limitations, we propose a novel deep-learning method dubbed PatientHandNet to reconstruct high-fidelity hand shapes in a canonical open-palm pose from multiple-depth images acquired with a single-depth camera. The hand poses in the depth images may vary, hand movements are allowed, facilitating the 3D scanning process in particular for patients with difficult conditions. The proposed method has strong operability since it is insensitive to the input pose, allowing for pose variations in the input depth images. We also proposed two novel datasets: a large-scale synthetic dataset to train our model and a real-world dataset with ground-truth hand biometrics extracted by an experienced anthropometrist. Extensive experimental results on the unseen synthetic data and real-world data demonstrate that the proposed method provides robust and easy-to-use hand shape reconstruction and outperforms the state-of-the-art methods in biometric accuracy terms.**

***Index Terms*—3D hand shape, 3D hand reconstruction, multi-view depth processing, multi-scale features, point cloud processing, hand biometrics.**

I. Introduction

RECONSTRUCTION of the 3D hand shape plays an important role in hand-centric medical applications, such as hand immobilization device design [1–3], artificial limb generation [4, 5], and osteoarthritis evaluation [6, 7], all of which usually require patients to open their hands for accurate anthropometric measurements [8]. However, it is difficult for patients with impaired hand functions [9, 10] to keep an open palm due to the fact that their hands usually exhibit complex hand poses caused by tenosynovitis, fracture, or other types of diseases. Researchers also observed that different hand postures cause substantial changes in hand geometry [11]. Therefore, there is a demand for reconstructing the patient's hand shape in an open-palm pose while not having access to such a pose during scanning.

Researchers have attempted to reconstruct hand shapes from different signals, such as hand anthropometric measurements, RGB images, depth images, and point clouds. The measurement-based methods need to enter dozens of different dimensions into a hand sizing system for estimating hand shape [1]. However, obtaining these measurement values highly relies on the expertise of the anthropometrist, which is expensive, time-consuming and prone to errors. The RGB-based methods require 2D to 3D representations for predicting 3D hand shapes from a single RGB image [12–14]. Such methods may result in perspective distortions and scale ambiguity due to the lack of depth information.

With the development of commodity depth cameras, depth images/point clouds can be easily obtained, encouraging researchers to devise low-cost hand scanners. These methods are mainly classified into two categories: non-parametric methods and parametric methods. Non-parametric methods correspond to traditional 3D hand scanning technologies, including single-camera and multi-camera scanning. Single-camera scanning [15, 16] moves a camera around a hand for data acquisition, which is time-consuming and may fail due to the fast movement of the camera. Multi-camera scanning [17–19] allows multiple cameras to capture a hand from different views simultaneously but relies on the quality of external calibration and requires a professional operation, which makes it unportable. More importantly, non-parametric methods usually require the hand to be in a static open-palm pose during scanning. However, maintaining an open-palm pose may induce pain or maybe simply not possible for patients with certain conditions. Parametric methods are developed based on the parametric hand models, which usually factor out the hand using parameters that control the shape and the pose [20]. Such methods mainly aim to fit the parametric hand models (e.g. SMPL-X) to the input data in order to obtain the optimal model parameters. Then an open-palm hand can be generated by setting the pose parameters to zero. However, such methods are not robust since a slight change in a parameter may result in a significant shape/pose variation. To address this issue, some researchers proposed to regress model vertices instead of parametric parameters [21, 22]. However, most of these works pay more attention to hand pose estimation while ignoring the accuracy of the shape.

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

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

Pengpeng Hu is with the Centre for Computational Science and Mathematical Modelling, Coventry University, Coventry, UK, email: ae1133@coventry.ac.uk (corresponding author).

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

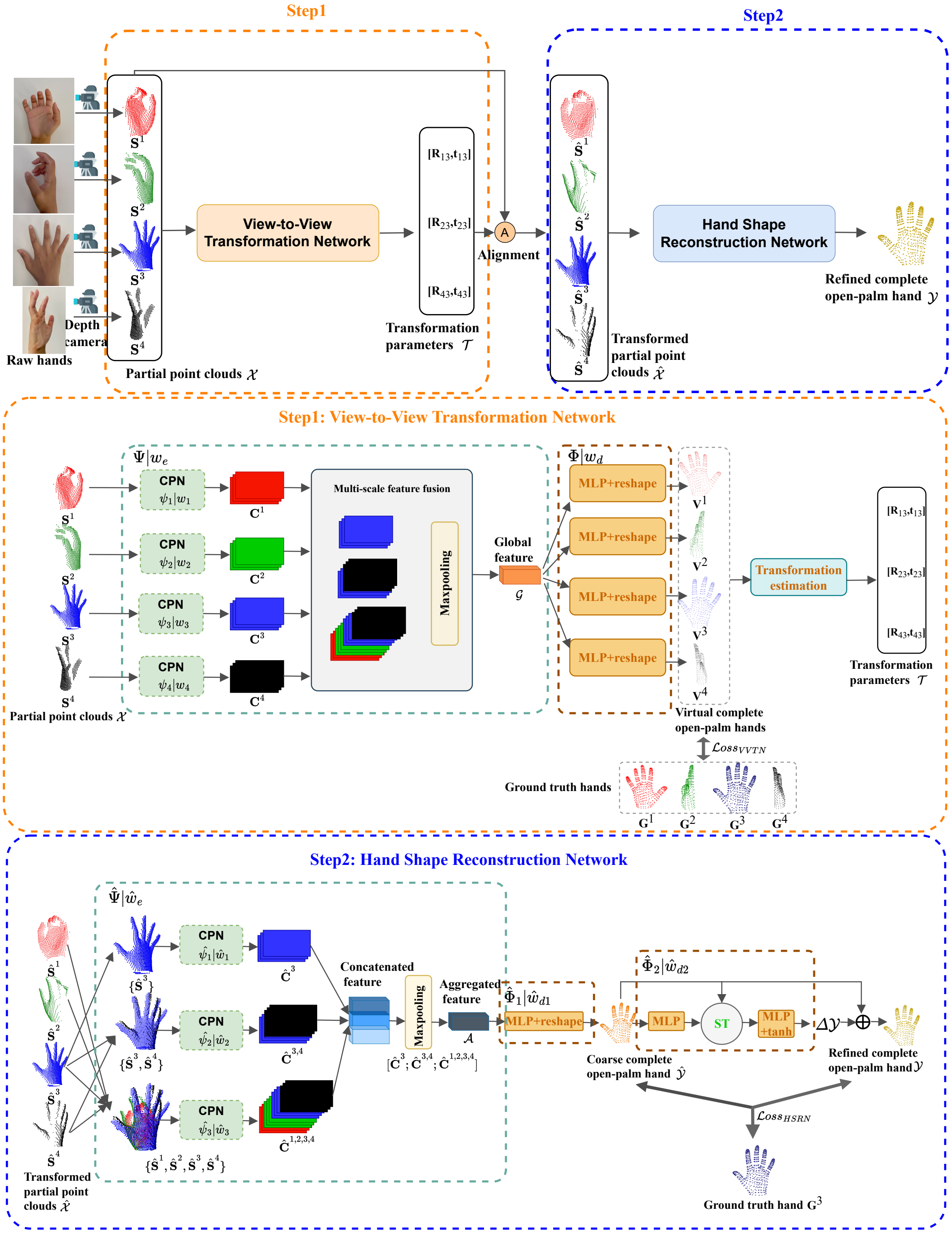

Step1
Step2
Depth camera
Raw hands
S^1
S^2
S^3
S^4
Partial point clouds $\mathcal{X}$
View-to-View Transformation Network
$[R_{13},t_{13}]$
$[R_{23},t_{23}]$
$[R_{43},t_{43}]$
Transformation parameters $\mathcal{T}$
A
Alignment
$\hat{S}^1$
$\hat{S}^2$
$\hat{S}^3$
$\hat{S}^4$
Transformed partial point clouds $\hat{\mathcal{X}}$
Hand Shape Reconstruction Network
Refined complete open-palm hand $\mathcal{Y}$
Step1: View-to-View Transformation Network
$\Psi|w_e$
CPN $\psi_1|w_1$
CPN $\psi_2|w_2$
CPN $\psi_3|w_3$
CPN $\psi_4|w_4$
C^1
C^2
C^3
C^4
Multi-scale feature fusion
Maxpooling
Global feature
$\mathcal{G}$
$\Phi|w_d$
MLP+reshape
MLP+reshape
MLP+reshape
MLP+reshape
V^1
V^2
V^3
V^4
Virtual complete open-palm hands
Transformation estimation
$[R_{13},t_{13}]$
$[R_{23},t_{23}]$
$[R_{43},t_{43}]$
Transformation parameters $\mathcal{T}$
$\mathcal{L}oss_{VVTN}$
Ground truth hands
G^1
G^2
G^3
G^4
Step2: Hand Shape Reconstruction Network
$\hat{\Psi}|\hat{w}_e$
$\hat{S}^1$
$\hat{S}^2$
$\hat{S}^3$
$\hat{S}^4$
Transformed partial point clouds $\hat{\mathcal{X}}$
$\{\hat{S}^3\}$
$\{\hat{S}^3,\hat{S}^4\}$
$\{\hat{S}^1,\hat{S}^2,\hat{S}^3,\hat{S}^4\}$
CPN $\hat{\psi}_1|\hat{w}_1$
CPN $\hat{\psi}_2|\hat{w}_2$
CPN $\hat{\psi}_3|\hat{w}_3$
$\hat{C}^3$
$\hat{C}^{3,4}$
$\hat{C}^{1,2,3,4}$
Concatenated feature
$[\hat{C}^3;\hat{C}^{3,4};\hat{C}^{1,2,3,4}]$
Maxpooling
Aggregated feature
$\mathcal{A}$
$\hat{\Phi}_1|\hat{w}_{d1}$
MLP+reshape
Coarse complete open-palm hand $\hat{\mathcal{Y}}$
$\hat{\Phi}_2|\hat{w}_{d2}$
MLP
ST
MLP +tanh
$\Delta\mathcal{Y}$
$\oplus$
Refined complete open-palm hand $\mathcal{Y}$
$\mathcal{L}oss_{HSRN}$
Ground truth hand G^3

Fig. 1: Schematic diagram of PatientHandNet for open-palm hand shape reconstruction. Our method consists of two steps, taking as input four partial point clouds from palmar view, dorsal view, ulnar view and radial view of the hand and regressing the vertices of an open-palm hand. In the 1st step, View-to-View Transformation Network (VVTN) consumes the four partial hand point clouds and generates four virtual complete hands in the same open-palm pose. The obtained four virtual complete hands are used to calculate transformation parameters between palmar view, ulnar view, radial view and dorsal view, which enables to roughly align the input four partial hands to form a holistic representation of hand shape. In the 2nd step, Hand Shape Reconstruction Network takes as input the four aligned partial point clouds and learns an aggregated feature through Multi-scale and Multi-view Feature Aggregation module (MMFA). Then the aggregated feature of MMFA is fed into the Hand Vertex Regression module (HVR) in order to regress the vertices of the open-palm hand shape in a coarse-to-fine manner.

To address the issues mentioned above, we aim to devise an algorithm to reconstruct hand shapes in a canonical open-palm pose from sparse multi-view depth images. Specifically, we propose a novel method termed PatientHandNet to reconstruct a high-fidelity open-palm hand shape from four-depth images. Note that we do not restrict the hand pose from the four views to be the same. The patient does not necessarily need to adopt an open-palm pose nor is required to maintain a *rigid* pose in the four scans. This is illustrated in Fig. 1(a), where we collected four depth images of the hand from palmar, dorsal, ulnar, and radial views using a single commodity depth camera. These depth images cover the whole hand geometry with minimal overlap, which implicitly accounts for the spatial dependency between adjacent views.

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

- We proposed, to the best of our knowledge, the first deep learning-based method, dubbed PatientHandNet, for reconstructing a high-fidelity open-palm 3D hand shape from four depth images captured by a single commodity depth camera. The method allows for various pose variations among the camera shots.
- We proposed a large-scale multi-view synthetic dataset with a wide variety of hand shapes and hand poses and corresponding ground truth hand shapes in a canonical open-palm pose.
- We collected a novel real-world dataset by capturing 18 subjects (13 males and 5 females) via a structure sensor Mark I employed in an iPad and hired a professional anthropometrist to obtain the ground-truth hand biometrics.
- Extensive experiments conducted on both synthetic and real-world data demonstrate that our method outperforms the state-of-the-art methods both qualitatively and quantitatively.

The rest of the paper is organized as follows. Section II briefly reviews related works that include non-parametric and neural parametric hand shape reconstruction. Section III describes the proposed method and proposed dataset. Section IV presents the experimental setup and comparison results among different methods on both synthetic data and real-world data. In Section V, we discuss the effectiveness of different modules of our method via detailed ablation studies. The paper is concluded in Section VI.

II. Related work

A. Non-parametric hand shape reconstruction

As aforementioned, non-parametric methods can be classified into multi-camera scanning and single-camera scanning according to the number of employed sensors. Multi-camera scanning [17–19, 23] captures the hand surface using multiple cameras fixed in different positions. The main advantage of multi-camera scanning is that it enables fast data acquisition and reduces the negative effect of involuntary movements by simultaneously acquiring the hand surface from different viewpoints. However, it highly depends on the quality of external calibration and needs a professional operation. In addition, it lacks mobility due to its expensive setup, configuration and calibration procedures. In contrast, single-camera scanning is a more cost-effective alternative to multi-camera scanning due to its requirement of only one scanner. This scanner held in the hand moved around the hand to acquire the complete shape [15, 16]. However, for successful 3D reconstruction, the acquisition procedure requires successive viewpoints to contain sufficient redundant information, resulting in time-consuming computation and potential failures due to the fast movement of the camera.

Apart from these limitations, non-parametric methods require the hand to be static and in an open-palm pose during scanning in order to provide accurate hand measurements; as mentioned, maintaining a rigid pose is impractical for regular patients and impossible for patients with impaired hand functions. To address these problems, we proposed a high-efficiency, posture-immune method which is improved as follows: 1) we only employ four depth images captured by a single camera and the hand poses of these depth images can be different; 2) our method outputs the hand shapes in an open-palm pose, which is essential for downstream applications such as hand biometrics extraction.

B. Neural parametric hand shape reconstruction

With the rapid development of deep learning, neural networks have been introduced to address the problem of 3D hand reconstruction. By literature review, we observe that most existing methods focus on predicting the hand pose from RGB images [24–26] or depth images [27–30]. A limited number of researchers have attempted to reconstruct the hand shape, but they estimate the pose and shape of the hand simultaneously. [31] proposed an end-to-end framework for hand shape reconstruction from a depth image, which embedded model-based hand pose and shape layers to generate 3D joint positions and hand meshes. Although such a method can predict hand mesh directly from the depth image, it suffers from artifacts due to the difficulty in optimizing complex parameters. [12] is a self-supervised 3D hand reconstruction network, which not only predicts 3D mesh but also outputs texture. [32] proposed an identity-aware hand mesh estimation model by regressing the parameters of MANO, which can incorporate the identity information to calibrate the shape parameters. [33] can find a balance between a parametric and non-parametric model to improve the accuracy of hand shape and pose. [34] can predict detailed hand mesh from a single image using the proposed frequency decomposition loss. However, all of them adopt parameter regression of the MANO model [35] to predict the hand pose and shape. Such techniques are less robust since a slight change in the parameters can significantly change the shape and pose. [36, 37] directly regressed the 3D joint positions and mesh vertices of the hand model from RGB images. [36] can estimate hand mesh from an RGB image with occlusion and [37] has lightweight stacked structures that can employ mobile to estimate hand shape from an RGB image. However, such RGB-based methods are easy to result in perspective distortions and scale ambiguity in the estimated outputs. To address the problems of perspective distortions and scale ambiguity, [38] proposed a vertex regression-based method that effectively establishes a one-to-one mapping between the voxelized depth map and the voxelized hand model. However, the voxelized depth map includes redundant information, distracting the attention of the network from extracting valuable features [39]. Unlike these methods, we focus on hand shape reconstruction by outputting the hand shapes in a canonical open-palm pose. Such a strategy will force the neural network to pay more attention to the shape instead of the pose.

III. Proposed Method

A. Problem Statement

In this section, we formulate the problem in this study. Given a set of uncalibrated partial point clouds of a hand $\mathcal{X} = \{\mathbf{S}^n\}_{n=1}^{N}$, where $\mathbf{S}^n = \{s_i^n \in \mathbb{R}^3 | i = 1, 2, ..., I^n\}$ denotes a set of points with I^n points captured from the nth view, N is the number of input partial point clouds and is set to 4 in this study (corresponding to four views of the hand: palmar, ulnar, dorsal, and radial views). Note that the poses of these partial hand point clouds can be the same as well different and the point numbers of each $\mathbf{S}^n$ can also be different. The goal is to devise a low-cost and easy-to-use method to reconstruct a high-fidelity open-palm hand mesh with J points $\mathcal{Y} = \{(y_j \in \mathbb{R}^3, e_z \in \mathbb{Z}^2)| j = 1, 2, ..., J, z = 1, 2, ..., Z\}$ from $\mathcal{X}$, where y_i and e_m are the vertices and edges of $\mathcal{Y}$ respectively. We proposed a two-step deep learning architecture for this task. Firstly, a transformation network is trained to estimate the relative transformation relationship $\mathcal{T}$ between $\mathcal{X}$. $\mathcal{T}$ is used to align $\mathcal{X}$ into $\hat{\mathcal{X}}$, which is represented as $\hat{\mathcal{X}} = \mathcal{X}\mathcal{T}$. Then we use a shape reconstruction network to predict open-palm hand shape $\mathcal{Y}$ from $\hat{\mathcal{X}}$. Therefore, there establish two mappings $M_1 : \mathcal{X} \longmapsto \mathcal{T}$ and $M_2 : \hat{\mathcal{X}} \longmapsto \mathcal{Y}$. $\mathcal{T}$ estimated by M_1 can strengthen dependency between adjacent views, enabling M_2 to learn the more robust geometric and spatial representation of hand shape. To further improve the performance, we also propose a coarse-to-fine decoder in M_2 to utilize local information to refine the vertex coordinates of the hand. Ultimately, this process yields a high-fidelity open-palm hand mesh derived from four independent partial point clouds.

As shown in Fig. 1, PatientHandNet we proposed mainly consists of two networks: view-to-view transformation network (VVTN) and hand shape reconstruction network (HSRN). Their procedures are detailed in Algorithm 1.

B. Feature Extractor

Similar to the encoder design of [40, 41], we stack two simplified PointNet (PN) modules to build the feature extractor, which is named the Combined PointNet (CPN). The architecture of CPN is depicted in Fig. 2. The first PN is a shared multilayer perceptron (MLP) consisting of two layers with ReLU activation. It takes as input I^n points represented as a $I^n \times 3$ matrix where each row denotes the 3D position of a point s_i^n and converts s_i^n into a point feature vector f_i^l. The set of f_i^l can be represented as a feature matrix F^l where each row is f_i^l. By means of the point-wise maxpooling operation, F^l is aggregated into a global feature $\boldsymbol{g}^l$. Next, we concatenate $\boldsymbol{g}^l$ to each f_i^l to obtain a updated feature matrix $\hat{F}^l$. $\hat{F}^l$ is further fed into the second PN to learn the final global feature $\boldsymbol{g}^h$ following the similar processing of the first PN. $\boldsymbol{g}^l$ and $\boldsymbol{g}^h$ are then concatenated, forming the combined latent vector $\mathbf{C}$. Such a strategy ensures that $\mathbf{C}$ contains both low-level and high-level features extracted from the input point clouds.

Algorithm 1 PatientHandNet

```
    Input:$\mathcal{X}$ : { $\mathbf{S}^1, \mathbf{S}^2, \mathbf{S}^3, \mathbf{S}^4$}
  Output:$\mathcal{Y}$
procedure VVTN        ▷ Input:$\mathcal{X}$        ▷ Output: $\mathcal{T}$
    features of $\mathcal{X} : \mathcal{G}^{(0)} \leftarrow \Psi_1(\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3, \mathbf{S}_4)$
    initialize step $t \leftarrow 0$
    $\mathbf{V}_t^1, \mathbf{V}_t^2, \mathbf{V}_t^3, \mathbf{V}_t^4 \leftarrow \Phi_1(\mathcal{G}^{(t)})$
    while $\mathcal{L}_{VVTN}((\mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3, \mathbf{G}_4), (\mathbf{V}_t^1, \mathbf{V}_t^2, \mathbf{V}_t^3, \mathbf{V}_t^4))$ not converged do
    $t \leftarrow t + 1$
    $\mathbf{V}_t^1, \mathbf{V}_t^2, \mathbf{V}_t^3, \mathbf{V}_t^4 \leftarrow$AdamOptimizer($\mathbf{V}_{t-1}^1, \mathbf{V}_{t-1}^2, \mathbf{V}_{t-1}^3, \mathbf{V}_{t-1}^4$)
    end while
    $\mathbf{V}^1, \mathbf{V}^2, \mathbf{V}^3, \mathbf{V}^4 \leftarrow \mathbf{V}_t^1, \mathbf{V}_t^2, \mathbf{V}_t^3, \mathbf{V}_t^4$
    $\mathcal{T} \leftarrow \arg\min_{\mathbf{R}_{n3}, \mathbf{t}_{n3}} \sum_{j=1}^{J} ||\mathbf{R}_{n3} v_j^n + \mathbf{t}_{n3} - \phi(v_j^n, \mathbf{V}^3)||^2, n = 1, 2, 4$
    return $\mathcal{T}$
procedure HSRN        ▷ Input:$\hat{\mathcal{X}}$        ▷ Output: $\mathcal{Y}$
    features of $\mathcal{X} : \mathcal{A}^{(0)} \leftarrow \hat{\Psi}(\hat{\mathbf{S}}_1, \hat{\mathbf{S}}_2, \hat{\mathbf{S}}_3, \hat{\mathbf{S}}_4)$
    initialize step $t \leftarrow 0$
    $\hat{\mathcal{Y}}_t \leftarrow_1(\mathcal{A}^{(t)})$
    $\Delta\mathcal{Y}_t \leftarrow \hat{\Phi}_2(\hat{\mathcal{Y}}_t)$
    $\mathcal{Y}_t \leftarrow_t + \Delta\mathcal{Y}_t$
    while $\mathcal{L}_{HSRN}(\mathbf{G}_3, \hat{\mathcal{Y}}_t, \mathcal{Y}_t)$ not converged do
    $t \leftarrow t + 1$
    $\mathcal{Y}_t \leftarrow$AdamOptimizer( $\mathcal{Y}_{t-1}$)
    end while
    $\mathcal{Y} \leftarrow \mathcal{Y}_t$
    return $\mathcal{Y}$
```

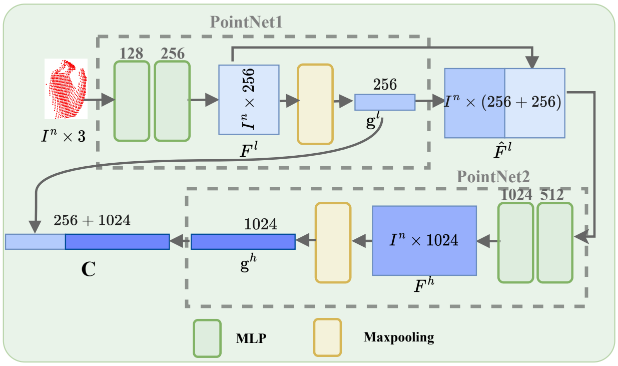
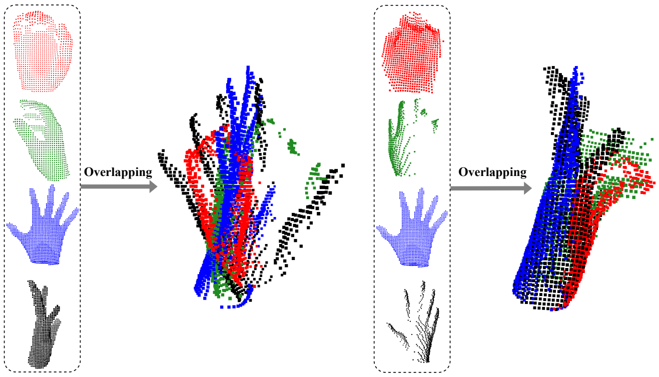


Fig. 2: Architecture of the proposed CPN.

C. View-to-View Transformation Network

As Fig. 1(b) shows, we design a four-branch encoder since the proposed method takes four partial point clouds $\mathcal{X} = \{\mathbf{S}^1, \mathbf{S}^2, \mathbf{S}^3, \mathbf{S}^4\}$ as input. An intuitive alternative is



(a) (b)

Fig. 3: Visual comparisons of overlapping the four partial scans without and with pre-alignment. (a) Raw partial point clouds and the overlapping of them by translating their centroids to the origin. (b) Pre-aligned partial point clouds and the overlapping of them.

to simply leverage four CPNs to consume individual partial point clouds to obtain four global features of partial point clouds and then apply a maxpooling operation to fuse these features to obtain the final global feature. However, such a solution ignores the spatial relationship among these point clouds. The input four partial point clouds are captured from the same hand but uncalibrated (Fig. 3(a)). Our insight is that the final hand reconstruction can be significantly improved if the four partial point clouds are pre-aligned (Fig. 3(b)) before extracting their features (comparisons can be seen in the following ablation studies). However, it is challenging to pre-align the four partial point clouds because of two main reasons: (1) low overlaps exist between the palmar-view point clouds and ulnar-view/radial-view point clouds, and no overlaps exist between the palmar-view and the dorsal-view point clouds, which makes the popular Iterative Closest Point (ICP) algorithm [42] and its variants fail; (2) the hand pose of the four partial point clouds are different, which is hard to be addressed using rigid registration. To tackle this problem, we propose the View-to-View Transformation network (VVTN) inspired by the work of [43], as shown in Fig. 1(b).

Given the four partial point clouds $\mathbf{S}^1, \mathbf{S}^2, \mathbf{S}^3, \mathbf{S}^4$, VVTN first uses four feature extractors to learn four combined latent vectors, respectively:

$$\begin{array}{ll} \mathbf{C}_1 = \psi_1(\mathbf{S}_1|w_1) & \mathbf{C}_2 = \psi_2(\mathbf{S}_2|w_2) \\ \mathbf{C}_3 = \psi_3(\mathbf{S}_3|w_3) & \mathbf{C}_4 = \psi_4(\mathbf{S}_4|w_4) \end{array} \tag{1}$$

where $\psi_n(\cdot|w_n)$ represents the feature extraction function with weight w_n learned by the CPN, n is the index of views. Then, it employs a maxpooling operation to fuse these four combined latent vectors into a global feature:

$$\mathcal{G} = maxpooling(\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3, \mathbf{C}_4; \mathbf{C}_3, \mathbf{C}_4; \mathbf{C}_3) \tag{2}$$

Therefore, the feature extractor of VVTN Ψ consists of four $\psi_n(\cdot|w_n)$ and one maxpooling:

$$\mathcal{G} = \Psi(\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3, \mathbf{S}_4|w_e) \tag{3}$$

Next, $\mathcal{G}$ is further fed into the decoders Φ that consists of four MLP with reshape operations to predict virtual correspondences [43] represented by four complete hands in the same open-palm pose:

$$\mathbf{V}^1, \mathbf{V}^2, \mathbf{V}^3, \mathbf{V}^4 = \Phi(\mathcal{G}|w_d) \tag{4}$$

where $\mathbf{V}^n = \{v_j^n \in \mathbb{R}^3 | j = 1, 2, ..., J\}$ denotes the set of points with J points corresponding to $\mathbf{S}^n$, n is the index of views. Since $\mathbf{V}^1, \mathbf{V}^2, \mathbf{V}^3, \mathbf{V}^4$ have the same number and order of points, and the same hand shape and pose, the virtual correspondences are built among $\mathbf{V}^1, \mathbf{V}^2, \mathbf{V}^3$, and $\mathbf{V}^4$. Given such virtual correspondences, the transformation parameters among $\mathbf{S}^1, \mathbf{S}^2, \mathbf{S}^3$, and $\mathbf{S}^4$ can be easily calculated and can be used to align $\mathbf{S}^1, \mathbf{S}^2, \mathbf{S}^3, \mathbf{S}^4$ despite they have different hand poses and little overlaps. Specifically, we take $\mathbf{V}^3$ as a reference view (or target view) corresponding to our final reconstructed $\mathcal{Y}$. The transformation $\mathcal{T} = \{\hat{\mathbf{R}}_{n3}, \hat{\mathbf{t}}_{n3} | n = 1, 2, 4\}$ of $\{\mathbf{V}^3, \mathbf{V}^n | n = 1, 2, 4\}$ can be estimated by solving the following optimization:

$$\hat{\mathbf{R}}_{n3}, \hat{\mathbf{t}}_{n3} = \underset{\mathbf{R}_{n3}, \mathbf{t}_{n3}}{\arg\min} \sum_{j=1}^{J} ||\mathbf{R}_{n3} v_j^n + \mathbf{t}_{n3} - \phi(v_j^n, \mathbf{V}^3)||^2 \tag{5}$$

where $\mathbf{R}_{n3} \in SO(3)$ and $\mathbf{t}_{n3} \in \mathbb{R}^3$ are the estimated rotation matrix and translation vector respectively from the nth view to the 3th view. Suppose $\hat{\mathbf{R}}_{n3}$ and $\hat{\mathbf{t}}_{n3}$ are the optimal rotation and translation, $\phi(v_j^n, \mathbf{V}^3)$ is a correspondence function that maps the point v_j^n in $\mathbf{V}^n$ to its corresponding point in $\mathbf{V}^3$. Each $\mathbf{S}^n$ in $\mathcal{X}$ can be aligned to $\mathbf{S}^3$, which is denoted as $\hat{\mathbf{S}}^n$.

$$\hat{\mathbf{S}}^n = \mathbf{S}^n[\hat{\mathbf{R}}_{n3}, \hat{\mathbf{t}}_{n3}] \tag{6}$$

Since $\mathbf{S}^3$ is a reference view, $\hat{\mathbf{S}}^3$ is still $\mathbf{S}^3$. Therefore, $\hat{\mathcal{X}}$ is the composition of $\hat{\mathbf{S}}^1, \hat{\mathbf{S}}^2, \hat{\mathbf{S}}^3, \hat{\mathbf{S}}^4$.

D. Hand Shape Reconstruction Network

The hand shape reconstruction network (HSRN) is an encoder-decoder architecture. The proposed encoder learns the aggregated feature in a multi-scale and multi-view manner, named Multi-scale and Multi-view Feature Aggregation (MMFA). The proposed decoder is able to regress 3D model vertices using a coarse-to-fine strategy, called Hand Vertex Regression (HVR).

1) Multi-scale and Multi-view Feature Aggregation: The pre-aligned partial point clouds $\hat{\mathcal{X}} = \{\hat{\mathbf{S}}^1, \hat{\mathbf{S}}^2, \hat{\mathbf{S}}^3, \hat{\mathbf{S}}^4\}$ are further fed into MMFA. As its name shows, the MMFA is designed to learn a feature by aggregating the multi-scale and multi-view information extracted from the input partial point clouds. We notice that the palmar-view and dorsal-view point clouds contain more information of hand length and breadth, while the ulnar-view and radial-view point clouds are more informative regarding the hand thickness. Therefore, the MMFA is designed to include three steps. Firstly, a CPN is used to extract the feature of a single-view point cloud $\{\hat{\mathbf{S}}^3\}$. Next, a CPN is applied to learn the feature of the concatenation of two point clouds $\{\hat{\mathbf{S}}^3, \hat{\mathbf{S}}^4\}$. Last, another CPN is leveraged to obtain the feature of the concatenation of all point clouds $\{\hat{\mathbf{S}}^1, \hat{\mathbf{S}}^2, \hat{\mathbf{S}}^3, \hat{\mathbf{S}}^4\}$. These features can be represented as:

$$\begin{aligned}\hat{\mathbf{C}}^{3} &= \hat{\psi}_1(\hat{\mathbf{S}}^{3}|\hat{w}_1)\\ \hat{\mathbf{C}}^{3,4} &= \hat{\psi}_2(cat(\hat{\mathbf{S}}^{3},\hat{\mathbf{S}}^{4})|\hat{w}_2)\\ \hat{\mathbf{C}}^{1,2,3,4} &= \hat{\psi}_3(cat(\hat{\mathbf{S}}^{1},\hat{\mathbf{S}}^{2},\hat{\mathbf{S}}^{3},\hat{\mathbf{S}}^{4})|\hat{w}_3)\end{aligned} \tag{7}$$

where $\hat{\psi}_n(\cdot|\hat{w}_n)$ represents the feature extraction function with weight $\hat{w}_n$ learned by the CPN and $cat(\cdot)$ represents the concatenation operation, and $\hat{\mathbf{C}}^{3}, \hat{\mathbf{C}}^{3,4}, \hat{\mathbf{C}}^{1,2,3,4}$ are three multi-scale multi-view features, which describe the hand shape from geometric and spatial perspectives. $\hat{\mathbf{C}}^{3}$ contains geometric information on the length and breadth of the hand. $\hat{\mathbf{C}}^{3,4}$ not only improves the hand thickness relative to $\hat{\mathbf{C}}^{3}$, but also establishes a relative spatial relationship between $\hat{\mathbf{S}}^{3}$ and $\hat{\mathbf{S}}^{4}$. $\hat{\mathbf{C}}^{1,2,3,4}$ fuses the information from all the views, which further enhances the geometric and spatial representation of the 3D hand. All these features are finally fused by the maxpooling operation into a comprehensive representation called aggregated feature $\mathcal{A}$:

$$\mathcal{A} = maxpooling(\hat{\mathbf{C}}^{3}; \hat{\mathbf{C}}^{3,4}; \hat{\mathbf{C}}^{1,2,3,4}) \tag{8}$$

Therefore, the feature extractor of HSRN $\hat{\Psi}$ consists of three $\hat{\psi}_n(\cdot|\hat{w}_n)$ and one maxpooling:

$$\mathcal{A} = \hat{\Psi}(\hat{\mathbf{S}}_1, \hat{\mathbf{S}}_2, \hat{\mathbf{S}}_3, \hat{\mathbf{S}}_4|\hat{w}_e) \tag{9}$$

2) Hand Vertex Regression: The Hand Vertex Regression (HVR) takes the aggregated feature $\mathcal{A}$ as input and aims to output a high-fidelity open-palm hand, as shown in Fig. 1(c). To this end, an MLP is used to directly regress the vertices of the open-palm hand. However, such a global strategy ignores the local information of hand. To overcome this limitation, we proposed a refinement method by embedding a local self-transformer (ST) with the MPL into our HVR module to locally refine the vertex coordinates of the reconstructed hand. We first use one MLP with reshape operation $\hat{\Phi}_1$ to predict the initially reconstructed hand shape, which is represented as $\hat{\mathcal{Y}}$:

$$\hat{\mathcal{Y}} = \hat{\Phi}_1(\mathcal{A}|\hat{w}_{d1}) \tag{10}$$

Then, each point $\hat{y}$ of $\hat{\mathcal{Y}}$ will be refined through ST, which is depicted in Fig. 4. Given jth point $\hat{y}_j$ of $\hat{\mathcal{Y}}$, we first calculate the attention scores $\{\mathbf{a}_{j,k}|k = 1, 2, ..., K\, k \neq j\}$ between $\hat{y}_j$ and its k-nearest neighboring points (k-NN) $\{\hat{y}_{j,k}|k = 1, 2, ..., K, k \neq j\}$ as:

$$\mathbf{a}_{j,k} = \frac{exp(M_Q(\hat{y}_j) \ominus M_K(\hat{y}_{j,k}))}{\sum_{k=1}^{K} exp(M_Q(\hat{y}_j) \ominus M_K(\hat{y}_{j,k}))} \tag{11}$$

where M_Q, M_K denote the MLPs with different parameters, and $\ominus$ denotes the element-wise subtraction. Finally, the point spatial feature $\mathbf{p}_j$ of $\hat{y}_j$ can be obtained by:

$$\mathbf{p}_j = M_V(\hat{y}_j) \oplus \sum_{k=1}^{K} a_{j,k} \odot M_V(\hat{y}_{j,k}) \tag{12}$$

where M_V denotes the MPL. $\oplus$ and $\odot$ denote element-wise addition and Hadamard product, respectively. The point spatial feature $\mathbf{p}_j$ is fed to the MLP to produce the displacement feature $\mathbf{d}_j$ of $\hat{y}_j$. $\mathbf{d}_j$ is further used for generating the point displacement $\Delta\hat{y}_j$, which is formulated as:

$$\Delta\hat{y}_j = \tanh(MLP(\mathbf{d}_j)) \tag{13}$$

where tanh is the hyper-tangent activation. Therefore, all $\hat{y}_j$ of $\hat{\mathcal{Y}}$ have displacements, which is formulated as:

$$\Delta\mathcal{Y} = \hat{\Phi}_2(\hat{\mathcal{Y}}|\hat{w}_{d2}) \tag{14}$$

where $\hat{\Phi}_2(\cdot|\hat{w}_{d2})$ is the ST module with weight $\hat{w}_{d2}$. Finally, $\hat{\mathcal{Y}}$ is refined by $\Delta\mathcal{Y}$:

$$\mathcal{Y} = \hat{\mathcal{Y}} + \Delta\mathcal{Y} \tag{15}$$

where $\mathcal{Y}$ is the refined hand shape.

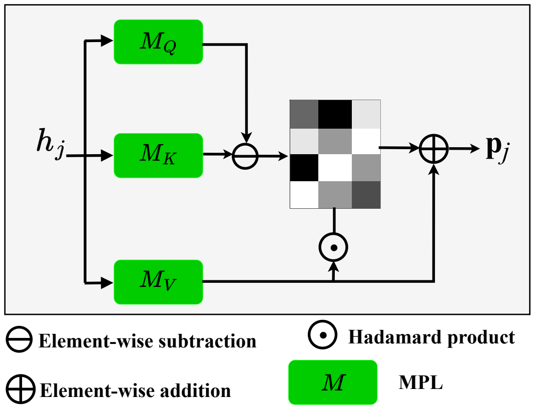


Fig. 4: The structure of the self-transformer module.

E. Loss Functions

The proposed model was trained in a supervised end-to-end manner. As shown in Fig. 1, the proposed PatientHandNet includes two neural networks. To train our model, we define the following loss functions to evaluate the reconstruction errors:

1) VVTN: VVTN has four branches to predict virtual correspondences $\{\mathbf{V}^1, \mathbf{V}^2, \mathbf{V}^3, \mathbf{V}^4\}$ corresponding to $\{\mathbf{S}^1, \mathbf{S}^2, \mathbf{S}^3, \mathbf{S}^4\}$. These branches are supervised by four SMPL-X hand models with J points in open palm pose, respectively, which are denoted as $\{\mathbf{G}^1, \mathbf{G}^2, \mathbf{G}^3, \mathbf{G}^4\}$ corresponding to the ground truth of $\{\mathbf{S}^1, \mathbf{S}^2, \mathbf{S}^3, \mathbf{S}^4\}$. Therefore, the loss function of VVTN is defined for these four branches:

$$\begin{aligned}\mathcal{L}_{VVTN} &= \mathcal{L}(\mathbf{V}^1, \mathbf{G}^1) + \mathcal{L}(\mathbf{V}^2, \mathbf{G}^2) + \mathcal{L}(\mathbf{V}^3, \mathbf{G}^3) + \mathcal{L}(\mathbf{V}^4, \mathbf{G}^4)\\ &= \frac{1}{J}\sum_{j=1}^{J}||v_j^1 - g_j^1||^2 + \frac{1}{J}\sum_{i=1}^{J}||v_j^2 - g_j^2||^2\\ &+ \frac{1}{J}\sum_{i=1}^{J}||v_j^3 - g_j^3||^2 + \frac{1}{J}\sum_{i=1}^{J}||v_j^4 - g_j^4||^2\end{aligned} \tag{16}$$

where $v_j \in \mathbf{V}, g_j \in \mathbf{G}$ are the j-th corresponding points pair in $\mathbf{V}, \mathbf{G}$.

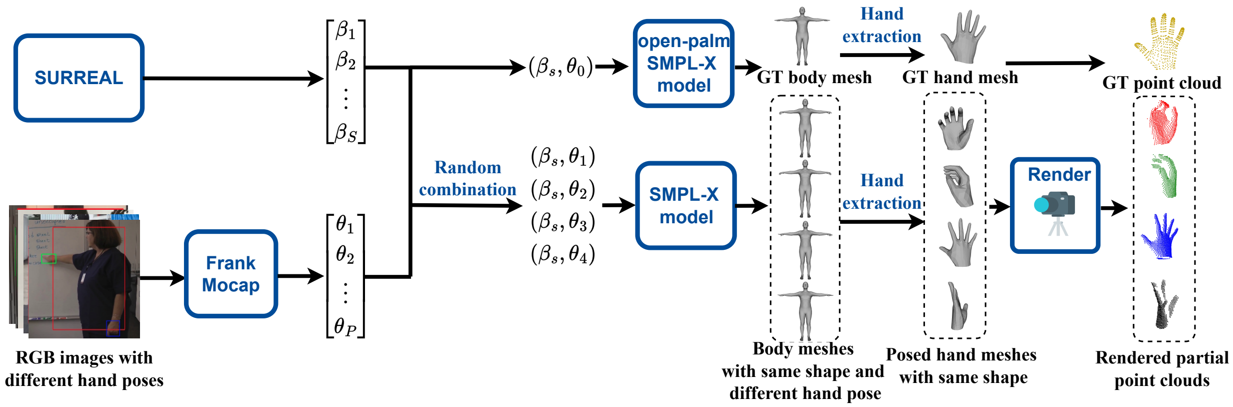


Fig. 5: The proposed pipeline of synthetic dataset generation. We first utilize FrankMocap to extract pose parameters (θ) of the SMPL-X model from RGB images. The shape parameters (β) of the SMPL-X model are taken from the SURREAL. Next, we randomly combine β and θ to generate four posed bodies with the same identity, from which we easily obtain the hand meshes. The ground-truth open-palm hand meshes can be obtained via resetting the random θ to be the pre-defined θ_0. Finally, we render realistic partial scans from the posed hands via the open-source Blensor.

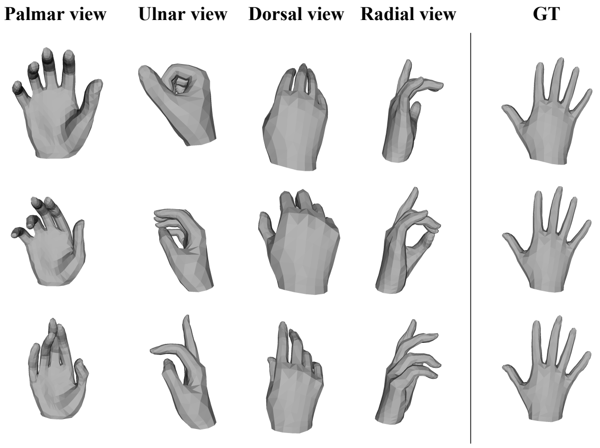


Fig. 6: Our synthetic meshes with hand shape and pose variations. Each row is from the same subject with different hand poses, the last column depicts the ground truth hand meshes.

2) HSRN: HSRN is supervised by the open-palm hand $\mathbf{G}^3$ in a coarse-to-fine manner. We, thus, define the loss as:

$$\begin{aligned}\mathcal{L}_{HSRN} &= \mathcal{L}(\hat{\mathcal{Y}}, \mathbf{G}^3) + \mathcal{L}(\mathcal{Y}, \mathbf{G}^3) \\ &= \frac{1}{J}\sum_{j=1}^{J} \|\hat{y}_j - g_j^3\|^2 + \frac{1}{J}\sum_{i=1}^{J} \|y_j - g_j^3\|^2 \end{aligned} \quad (17)$$

where $\hat{y}_j \in \hat{\mathcal{Y}}, y_j \in \mathcal{Y}$ and $g_j^3 \in \mathbf{G}^3$are the j-th corresponding points in $\hat{\mathcal{Y}}, \mathcal{Y}$ and $\mathbf{G}^3$,respectively.

3) Complete loss: Our complete loss is defined by a weighted sum of $\mathcal{L}_{VVTN}$ and $\mathcal{L}_{HSRN}$:

$$\mathcal{L} = \alpha_1 \times \mathcal{L}_{VVTN} + \alpha_2 \times \mathcal{L}_{HSRN} \quad (18)$$

where α_1 and α_2 are the weights that control the contribution of each term.

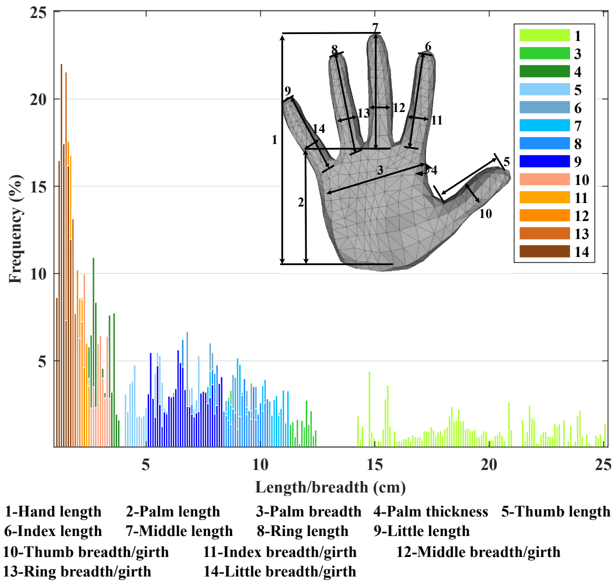


Fig. 7: The 13 dimensions of hand accompanying with their IDs on the hand template and the statistics of lengths of 13 dimensions in our synthetic dataset.

F. Proposed dataset

1) Synthetic dataset: In order to train our model, a large-scale dataset consisting of inputs with four partial scans of the same hand and corresponding ground-truth open-palm hand meshes is necessary. Moreover, the four partial scans may have the same or different postures. Unfortunately, none of the existing public hand datasets meets these requirements. The existing hand datasets lack ground-truth open-palm hand shapes and most of them only contain single palmar-facing depth images. An intuitive solution is to collect the dataset by scanning real-world subjects. However, such a procedure is extremely expensive and time-consuming for a large-scale dataset (e.g., 100K samples). To address this problem, we propose an innovative method to generate the appropriate dataset needed for training the proposed PatientHandNet. Fig. 5 illustrates the proposed pipeline for generating such a dataset.

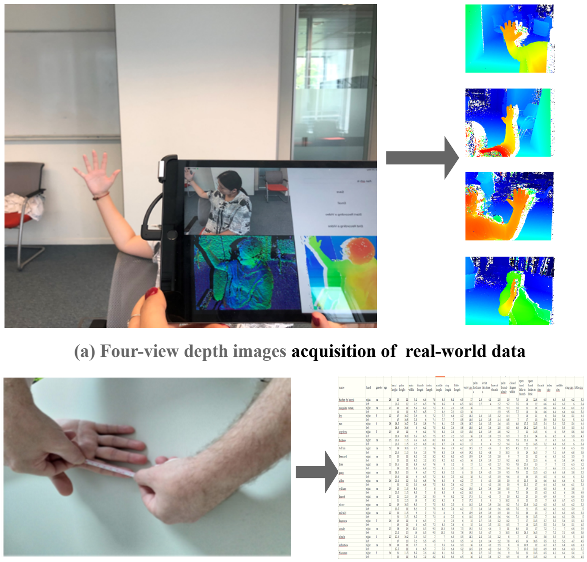
Generating 3D hand meshes. First, we collected $12K$ images online consisting of female and male subjects performing different motions. Second, a state-of-the-art RGB-based hand and motion capture system, FrankMocap [44], was used to extract the θ and β values of SMPL-X from these images. SMPL-X [45] is a generative body model with two key parameters: the pose parameter θ and the shape parameter β. Given a pair of θ and β, a new human body mesh with specified body posture, hand gesture and facial expression can be obtained. Therefore, $12K$ θs and $12K$ βs were obtained via FrankMocap. We observed that the estimated poses looked good, while the estimated shapes were less accurate due to the inherent scale ambiguity associated with the 2D to 3D conversion via FrankMocap. We, thus, dropped the $12K$ βs and used 2103 βs from the SURREAL dataset [46]. Next, a large-scale body dataset was obtained by a random combination of βs and θs. Specifically, we randomly selected a β and four θs to create four body models, which had the same body shape but different hand poses. By assigning the pre-defined open-palm pose (denoted by θ_0), another body model with the designed open-palm hand shapes was built. The left and right hands were easily extracted from these body models to build a sample of our dataset (four pairs of posed hands as input and a pair of open-palm hands as the ground truth). Using this procedure, we generated $300K$ samples for the proposed dataset. Fig. 6 depicts samples of the proposed dataset.

Rendering multi-view partial scans. The open-source Blender Sensor Simulation plugin Blensor [47] was used to render the partial scans of hands. In this study, we set the camera as Microsoft Kinect V2 and the scanning distance are randomly selected at intervals from 0.5 to 0.8 meters. To mimic real-world scanning, we also added a random rotation set at the intervals from -10^o to 10^o in all x, y, z directions; the noise was added as well. We rendered the left hand and the right hand separately.

As aforementioned, each sample of our hand *mesh* dataset consists of four posed hands and one open-palm ground-truth hand. By rendering one partial scan for each posed hand mesh from one of the four designed views (palmar, ulnar, dorsal, and radial views) completes the data generation process consisting of multi-view partial hand scans and corresponding ground-truth open-palm hand meshes.

Statistics of the synthetic hand dataset. To analyze the hand shapes of our proposed dataset, we measured lengths of the 13 dominant dimensions [48] based on the open-palm hand pose, as shown in Fig. 7. Take hand length as an example (Measurement Type #1), 50% of its length is distributed between 17 and 22cm, which is consistent with the real-world hand biometrics statistical result in [49]. Moreover, the minimum length and maximum length of hand length are 13 cm and 25.2 cm, which widely covers the hand length from a child to an adult. The other dimensions have similar statistical properties.

2) *Real-world dataset:* Our method also can generalize well to real-world data, but there exists no publicly available multi-view real-world dataset that includes ground-truth open-palm hands. Therefore, we built a novel real-world dataset, as shown in Fig. 8(a). An iPad with an embedded structure sensor Mark I was used to scanning 18 different hands (13 males and 5 females) in arbitrary posture from four views of the hand. A professional anthropometrist extracted hand measurements according to the 14 dimensions described in Fig. 8(b); these can be considered as the ground truth values of the real-world dataset.



(a) Four-view depth images acquisition of real-world data

(b) Measurement values acquisition of real-world data

Fig. 8: (a) An iPad embedded with a structure sensor Mark I is employed to scan a hand and produce four-view depth images. (b) A professional anthropometrist uses tape to extract biometrics measurements according to the 14 dimensions detailed in Fig. 7. The table shows the measurement values.

IV. Experiments

We compared the proposed method against the state-of-the-art shape reconstruction methods based on synthetic data as well as real-world data, including the following methods: HandAR [13], S^2HAND [12], HandOccNet [36], SnowflakeNet [50], L2H [51] and MobRecon [37].

When we use the unseen synthetic dataset mentioned in Section.III-G, we randomly select 720 samples, which have open palm hand as ground truth and no overlapping with training and validation datasets. These selected data are represented as point clouds, which can be the input of point-based methods (SnowflakeNeta [50] and L2H [51]). However, RGB-based methods (HandAR [13], S^2HAND [12], HandOccNet [36], MobRecon [37]) require RGB images as input. To ensure a fair comparison, we obtain RGB images by rendering the ground-truth open-palm mesh corresponding to the selected data.

We also evaluate the performance on real-world data. As described in Section.III-G, the real-world data we collected

TABLE I: Mean of absolute and relative anthropometric measurements errors on 5 male real-world data for various noise levels (Unit:mm)

Measurements	$\sigma = 0$		$\sigma = 0.1\%$		$\sigma = 0.2\%$		$\sigma = 0.3\%$	
	E_{AE}	E_{RE}	E_{AE}	E_{RE}	E_{AE}	E_{RE}	E_{AE}	E_{RE}
1-Hand length	1.60	0.74%	2.20	1.13%	2.60	1.21%	3.60	1.60%
2-Palm length	2.20	1.91%	3.00	2.33%	3.80	3.29%	3.60	3.24%
3-Palm width	2.20	2.39%	2.40	2.49%	2.60	2.86%	3.40	3.49%
4-Thumb length	2.00	2.83%	2.00	3.20%	2.00	2.98%	2.20	3.07%
5-Index length	1.40	1.71%	1.80	2.52%	2.80	3.54%	3.00	3.57%
6-Middle length	2.00	2.23%	2.40	2.97%	3.60	4.06%	4.20	4.49%
7-Ring length	2.80	3.44%	3.10	4.11%	3.70	4.57%	4.40	5.23%
8-Little length	2.00	3.23%	2.80	4.78%	3.80	6.00%	4.60	6.90%
9-Palm thickness	1.20	3.90%	1.80	6.57%	3.00	9.87%	3.80	11.81%
10-Thumb girth	2.00	2.66%	2.40	3.52%	3.20	4.30%	3.60	5.08%
11-Index girth	1.20	1.66%	1.40	1.89%	1.40	2.02%	2.20	2.87%
12—Middle girth	1.80	2.61%	2.00	2.42%	2.20	3.18%	3.00	4.03%
13—Ring girth	1.40	2.19%	1.60	2.89%	1.00	1.63%	1.40	2.58%
14—Little girth	2.20	3.94%	2.20	4.04%	2.40	4.48%	2.60	4.53%

TABLE II: Mean of absolute and relative anthropometric measurements errors on 5 female real-world data in different density (Unit:mm)

Measurements	No sampling		2048		1024		512	
	E_{AE}	E_{RE}	E_{AE}	E_{RE}	E_{AE}	E_{RE}	E_{AE}	E_{RE}
1-Hand length	2.00	1.09%	2.00	1.10%	2.60	1.41%	3.20	1.73%
2-Palm length	2.20	2.13%	2.60	2.55%	3.20	3.14%	3.50	3.53%
3-Palm breadth	1.60	1.97%	1.00	1.25%	1.40	1.75%	1.80	2.25%
4-Thumb length	1.80	2.96%	2.20	3.66%	2.00	3.33%	1.80	3.00%
5-Index length	0.80	1.10%	1.20	1.65%	1.60	2.20%	2.00	2.76%
6-Middle length	3.20	3.99%	3.20	3.99%	3.60	4.49%	3.90	4.96%
7-Ring length	2.40	3.29%	2.80	3.84%	3.00	4.15%	3.00	4.11%
8-Little length	2.80	4.74%	3.00	5.06%	3.20	5.43%	3.60	6.13%
9-Palm thickness	1.20	4.99%	1.40	5.76%	2.00	8.08%	2.80	11.17%
10-Thumb girth	2.20	3.41%	2.00	3.08%	2.40	3.72%	3.20	4.97%
11-Index girth	2.60	4.46%	2.60	4.85%	2.70	4.95%	3.00	5.19%
12-Middle girth	2.40	4.06%	2.40	4.06%	2.00	3.43%	2.20	3.83%
13-Ring girth	1.40	2.53%	1.40	2.53%	1.60	2.95%	1.40	2.59%
14-Little girth	2.80	6.37%	3.00	6.79%	3.40	7.64%	3.40	7.67%

consist of partial point clouds and corresponding ground-truth hand measurement values extracted by a professional anthropometrist. Furthermore, the corresponding RGB images are also captured. Therefore, the performance of point-based methods and RGB-based methods can be evaluated on real-world data with ground-truth hand measurement values.

In the end, we conduct ablation experiments based on the 720 samples to understand the effectiveness of key components in our proposed method, including multi-view input, pre-aligned inputs, multi-scale and multi-view feature learning and point-based refinement strategy.

A. Experimental Setup

The training is operated on a desktop PC with Intel(R) Core i9-10900 CPU @ 2.80GHz and GeForce RTX 3090Ti. Our training has 200 epochs, and we set $\alpha_1 = 1, \alpha_2 = 0$ for the first 50 and set $\alpha_1 = 1, \alpha_2 = 1$ for the rest of the epochs.

B. Evaluation metrics

1) Evaluation Metrics on synthetic data: Since the synthetic dataset contains the ground-truth open-palm hand, errors can be computed by comparing the predicted hand shape and the ground-truth hand shape. To this end, we employ the widely-used reconstruction evaluation metric: Chamfer Distance (CD) [52]. The CD error measures the average nearest squared distance between predicted hand $\mathcal{Y}$ and the ground truth hand $\mathbf{G}^3$, which is defined as:

$$\begin{aligned} E_{CD}(\mathcal{Y}, \mathbf{G}^3) = \frac{1}{2}[\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \min_{g^3 \in \mathbf{G}^3} ||y - g^3||_2^2 \\ + \frac{1}{|\mathbf{G}^3|} \sum_{g^3 \in \mathbf{G}^3} \min_{y \in \mathcal{Y}} ||g^3 - y||_2^2] \end{aligned} \tag{19}$$

2) Evaluation Metrics on real-world data: Unlike the synthetic dataset, the real-world dataset only contains the multi-view partial scans as input but lacks the ground-truth open-palm hand shapes, which makes CD fail. Fortunately, our collected real-world dataset contains the hand measurement values measured by a professional anthropometrist, which are ground-truth measurement values. From the reconstructed hand shape, hand measurements can be easily extracted. Therefore, the absolute error (AE) and relative error (RE) are applied to measure the measurement errors by comparing the predicted measurement values M_{pred} and ground-truth measurement values M_{gt}. The AE and RE are defined as:

$$E_{AE} = |M_{pred} - M_{gt}| \tag{20}$$

$$E_{RE} = \frac{|M_{pred} - M_{gt}|}{M_{gt}} \times 100 \tag{21}$$

3) Results on synthetic data: We first compare our results with different methods on the unseen synthetic data. Fig. 9 depicts quantitative comparisons and each point is colorized by the point-to-point errors in millimeters between the reconstructed hand shape and the ground-truth shape. It can be seen that the results of S^2HAND [12] and HandOccNet [36] do not have the open palm posture but have severe bending on the fingers. The results of HandAR [13] produced a severe deformity on the little finger and the ninth result of them collapses. The results of SnowflakeNet [50], L2H [51] and MobRecon [37] have good geometric shapes. However, their point-to-point errors are larger than our proposed method, especially MobRecon [37], which demonstrates RGB-based methods are easy to result in scale ambiguity in the shape.

We also calculate the average value μ, average standard derivation σ and maximum distance max of Eq.(19) to quantitatively evaluate each method and summarize the result in Table III. From this we can have two main observations. First, the performance of the depth-based methods are better compared to RGB-based methods. For example, our method leads to 12.17 mm, 3.68 mm and 41.55mm improvements of μ, σ, and max compared with HandAR [13], respectively. This verifies that 3D information is desired for more accurate hand shape reconstruction. Second, compared with different depth image-based methods, our results achieves the best performance in term of all metrics. Specifically, our method significantly outperforms SnowflakeNet [50] and L2H [51] in terms of both μ and σ, which implies that our method has a higher accuracy for the global hand shape. The improvement of max brought by our method is also obvious, which indicates that our method achieves the most accurate local hand shape reconstruction.

4) Results on real-world data: Once our model is successfully trained, it generalizes well to real-world data. We further qualitatively compare the proposed method against different methods based on our collected real-world dataset, as shown in Fig. 10. Note that the results of S^2HAND [12] and HandOccNet [36] do not have the open palm posture but have bent

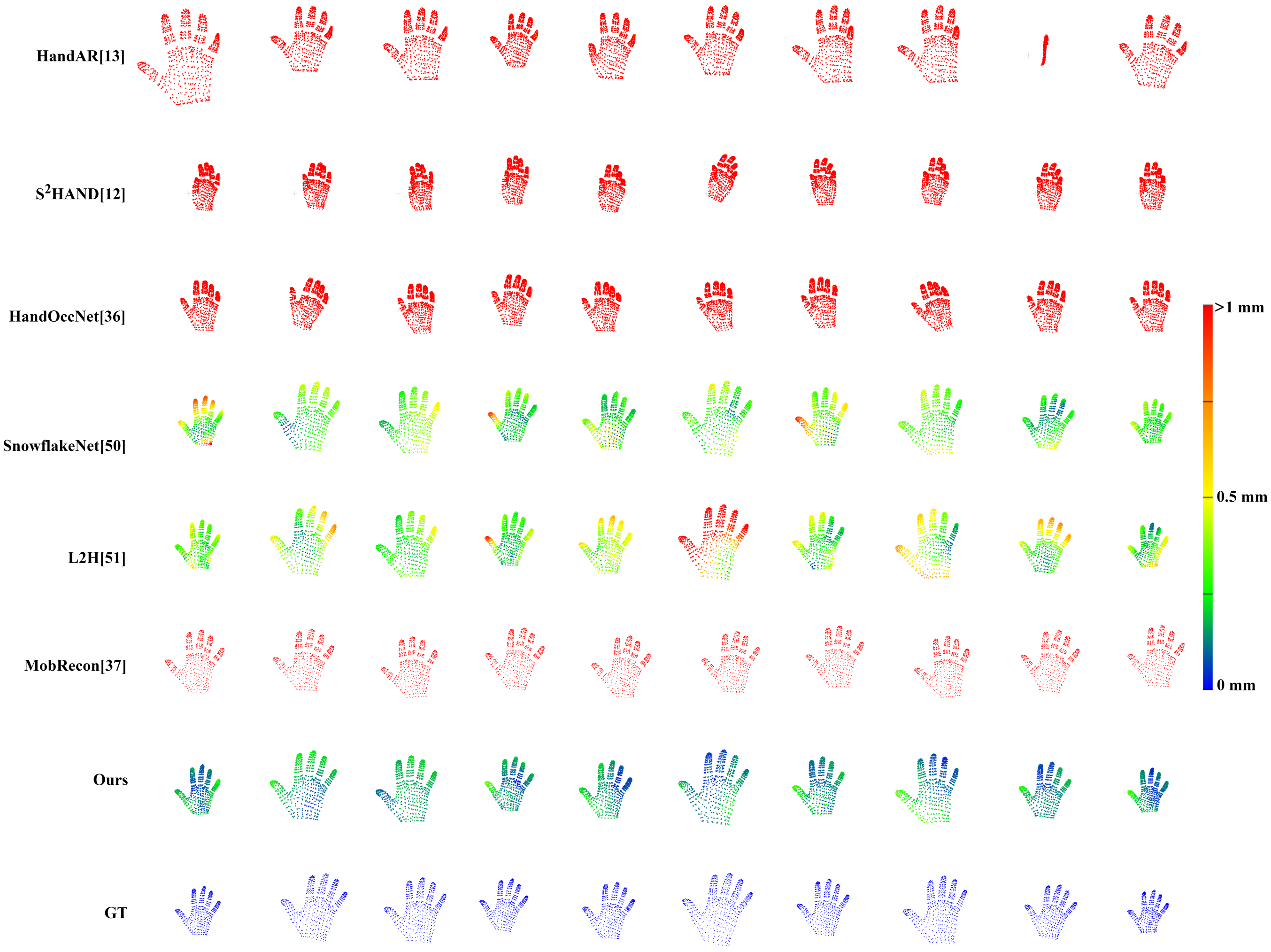


Fig. 9: The visual comparison of hand shape reconstruction error with state-of-the-art methods on synthetic data. From top to bottom: the results of HandAR[13], S^2HAND[12], HandOccNet[36], SnowflakeNet[50], L2H[51], MobRecon[37], our results and the ground truth. The colored points represent the point-to-point errors in millimeters between reconstructed hand shape and GT.

fingers, and the results of HandAR [13] obviously have severe deformation in shape. The results of MobRecon [37] have various hand postures instead of open palm poses. Despite the results of SnowflakeNet [50] and L2H [51] looking similar to ours, our method actually achieves the best results according to Table IV that shows the quantitative hand measurement comparisons for five females and five males. Besides, the average AE and RE of RGB-based methods are larger than 5 mm and 9% respectively, the average AE and RE of depth-based methods of SnowflakeNet [50] and L2H [51] are larger than 4 mm and 5%, while our results are less than 3 mm and 4%. The results on real-world data are consistent with those on unseen synthetic data, which further demonstrates the effectiveness of the proposed method.

C. Robustness

In this section, we demonstrate the robustness of our proposed method on real-world data in different formats.

TABLE III: The average mean, standard deviation and maximum distance of chamfer distance on synthetic data (Unit:mm)

Methods	μ	σ	max
HandAR[13]	12.38	4.31	43.84
S^2HAND [12]	13.34	3.63	21.93
HandOccNet [36]	13.41	4.90	27.57
SnowflakeNe[50]	0.34	0.75	2.50
L2H [51]	0.43	0.81	2.62
MobRecon [37]	5.98	3.19	19.18
Ours	**0.21**	**0.63**	**2.29**

We randomly selected 5 male real-world subjects' scans and introduce different levels of Gaussian noise, specifically $\sigma = 0.1\%, 0.2\%, 0.3\%$, to study the effect of noise on our method. The error results are presented in Table I. Increasing levels of noise will lead to higher errors. However, the maximum absolute error still less than 5 mm even with the presence of

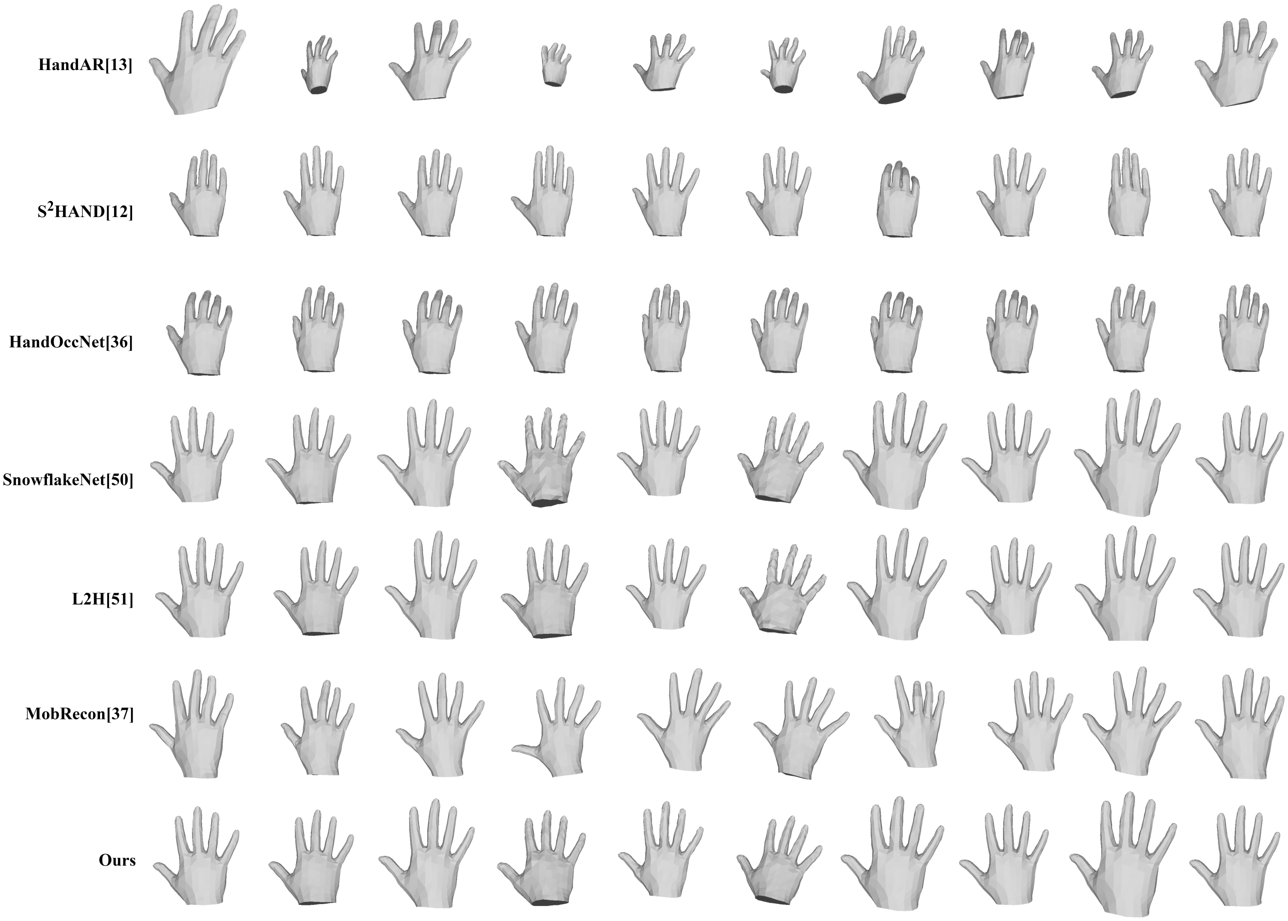


Fig. 10: The visual comparison of hand shape reconstruction error with state-of-the-art methods on real-world data. From top to bottom: the results of 3 HandAR[13], S^2HAND[12], HandOccNet[36], SnowflakeNet[50], L2H[51], MobRecon[37], and our results.

TABLE IV: Mean of absolute and relative anthropometric measurements errors comparison on real-world data (Unit:mm)

Measurements	Female (5 subjects)														Male (5 subjects)													
	HandAR[13]		S^2HAND[12]		HandOccNet[36]		SnowflakeNe[50]		L2H[51]		MobRecon [37]		Ours		HandAR[13]		S^2HAND [12]		HandOccNet[36]		SnowflakeNe[50]		L2H[51]		MobRecon [37]		Ours	
	E_{AE}	E_{RE}	E_{AE}	E_{RE}	E_{AE}	E_{RE}	E_{AE}	E_{RE}	E_{AE}	E_{RE}	E_{AE}	E_{RE}	E_{AE}	E_{RE}	E_{AE}	E_{RE}	E_{AE}	E_{RE}	E_{AE}	E_{RE}	E_{AE}	E_{RE}	E_{AE}	E_{RE}	E_{AE}	E_{RE}	E_{AE}	E_{RE}
1- Hand length	63.8	35%	10.0	5.4%	7.6	4.2%	5.2	2.8%	4.8	2.6%	15.4	8.2%	**2.0**	**1.1%**	43.0	20.0%	34.4	16.1%	23.6	10.8%	6.6	3.2%	4.6	2.1%	25.4	11.6%	**1.6**	**0.7%**
2- Palm length	33.8	33%	8.8	8.6%	5.4	5.4%	5.4	5.2%	5.0	5.0%	10.6	10.2%	**2.2**	**2.1%**	26.2	22.1%	18.8	15.6%	14.0	11.5%	5.8	4.9%	3.6	3.0%	17.8	14.6%	**2.2**	**0.9%**
3- Palm breadth	29.2	35.9%	8.6	10.5%	8.2	10.0%	3.0	3.7%	3.2	4.1%	11.6	14.2%	**1.6**	**2.0%**	21.6	22.3%	21.0	22.2%	18.6	19.5%	4.8	5.1%	5.8	5.9%	17.8	18.6%	**2.2**	**2.4%**
4- Thumb length	25.6	42.6%	11.0	18.2%	7.8	12.9%	4.2	6.9%	4.4	7.3%	9.6	15.8%	**1.8**	**3.0%**	17.2	24.7%	15.8	23.4%	11.8	16.8%	4.4	6.5%	4.6	6.6%	14.0	0.0%	**2.0**	**2.8%**
5- Index length	27.4	37.5%	10.4	14.3%	8.2	11.3%	3.2	4.4%	2.6	3.6%	3.8	5.2%	**0.8**	**1.1%**	14.8	17.6%	17.4	21.4%	15.6	19.1%	3.0	3.8%	3.2	4.0%	9.0	10.8%	**1.4**	**1.7%**
6- Middle length	29.6	36.5%	7.2	8.8%	4.2	5.1%	4.4	5.5%	3.6	4.5%	7.2	8.7%	**3.2**	**4.0%**	18.8	20.0%	15.2	16.4%	12.0	12.8%	3.4	4.0%	4.2	4.7%	9.0	9.6%	**2.0**	**2.2%**
7- Ring length	25.4	34.4%	9.8	13.2%	7.6	10.3%	3.4	4.7%	**1.8**	**3.3**%	6.6	8.8%	2.4	**3.3%**	18.4	21.5%	10.6	24.4%	13.4	15.5%	4.6	5.7%	5.0	6.1%	13.0	14.9%	**2.8**	**3.4%**
8- Little length	23.4	39.6%	9.0	15.2%	7.4	12.5%	3.8	6.4%	3.0	5.1%	3.4	5.8%	**2.8**	**4.7%**	16.8	24.7%	16.0	23.9%	11.0	16.3%	3.6	5.8%	4.4	6.9%	7.4	10.9%	**2.4**	**3.8%**
9- Palm thickness	7.2	27.8%	4.6	17.7%	6.0	23.0%	4.6	18.3%	4.2	16.6%	2.8	10.6%	**1.2**	**5.0%**	8.8	28.6%	9.2	30.2%	9.0	29.7%	2.2	7.1%	3.6	12.1%	6.2	20.9%	**1.2**	**3.9%**
10- Thumb girth	12.8	19.7%	7.8	12.0%	6.8	10.5%	4.4	6.8%	5.6	8.7%	6.0	9.2%	**2.2**	**3.4%**	13.8	17.8%	18.2	23.5%	16.4	21.0%	4.0	5.4%	4.4	6.1%	13.2	17.0%	**2.0**	**2.7%**
11- Index girth	10.0	17.1%	7.0	11.9%	5.6	9.5%	5.2	8.9%	5.2	9.0%	4.0	6.7%	**2.6**	**4.5%**	12.6	17.1%	16.4	23.2%	13.4	18.7%	5.6	7.8%	4.8	6.8%	11.0	15.0%	**1.2**	**1.7%**
12- Middle girth	11.2	18.9%	6.6	11.0%	4.8	8.0%	4.4	7.5%	4.2	7.1%	3.8	6.3%	**2.4**	**4.1%**	12.8	17.9%	15.8	22.7%	12.2	17.2%	5.2	7.6%	3.4	4.9%	8.2	11.4%	**1.8**	**2.6%**
13- Ring girth	9.0	16.2%	5.6	10.1%	6.4	11.6%	3.4	6.2%	4.8	7.8%	5.0	8.9%	**1.4**	**2.5%**	9.6	15.0%	12.2	19.2%	10.8	17.0%	5.2	8.3%	4.8	7.6%	7.2	11.1%	**1.4**	**2.2%**
14- Little girth	9.2	19.8%	5.0	10.9%	5.2	11.3%	4.0	9.1%	4.4	10.0%	4.6	9.9%	**2.8**	**6.4%**	10.4	18.2%	12.4	22.0%	11.0	19.5%	5.2	9.3%	4.0	7.1%	7.0	12.3%	**2.2**	**3.9%**
Average	22.7	29.6%	8.0	21.7%	6.5	10.4%	4.2	6.9%	4.1	6.8%	6.7	9.2%	**2.1**	**3.4%**	17.5	20.5%	17.4	21.7%	13.8	17.5%	4.5	6.0%	4.3	6.0%	11.9	14.2%	**1.9**	**2.6%**

0.3% noise, which is deemed acceptable, which is deemed acceptable. These results indicate that our method effectively filters out irrelevant or erroneous information, enabling it to focus on extracting valuable features for reliable predictions. Additionally, we conducted experiments on another 5 female real-world dataset, which we subsampled into three different point cloud densities: 2048, 1024, and 512. The purpose was to examine how the density of the point cloud affects the performance of our method. The error results are shown in Table II, from which we can see that the errors with different densities are very close. Therefore, our method can handle sparse or dense data without affecting its ability to extract valuable information. In conclusion, our method is able to adapt well to real-world scenarios where noise and data irregularities are prevalent.

D. *Ablation studies*

We take the average value μ, average standard derivation σ and maximum value max of CD errors to evaluate the performance of critical components. Details are provided next.

1) Effectiveness of multi-view input: In this section, we explored the performance of different inputs based on the proposed method. PatientHandNet takes four depth images (i.e., palmar, ulnar, dorsal, and radial views of hand) as the

TABLE V: Comparisons of reconstruction errors with different number of depth images (Unit:mm)

	single depth image	two depth images	four depth images
μ	0.443	0.430	**0.401**
σ	0.153	0.146	**0.099**
max	1.106	1.034	**0.598**

TABLE VI: Comparisons of reconstruction errors without and with pre-aligning raw inputs (Unit:mm)

	Wihtout pre-alignment	With pre-alignment
μ	0.468	**0.401**
σ	0.158	**0.099**
max	1.134	**0.598**

input since they are the minimal captures necessary in order to obtain the complete geometry of the hand. We retrain our network based on a single depth image (dorsal view), two depth images (dorsal view and radial view) and four depth images, respectively. Table V compares the reconstruction errors of them. It can be seen that the results from two depth images are better than the results from a single depth image, and the results from four depth images are the best.

2) Effectiveness of pre-aligned inputs: One of our insights is that the reconstruction accuracy is improved if the raw inputs are pre-aligned. To validate it, we compared the reconstruction accuracy for the proposed method without and with the pre-alignment. As Table VI illustrates, it can be seen that the pre-alignment of raw inputs can significantly reduce reconstruction errors. With the help of the proposed VVT module, the input partial point clouds are aligned to obtain more accurate global shape information of the complete hand and also introduce the spatial relationship among inputs.

3) Effectiveness of multi-scale and multi-view feature learning: MMFA was proposed in our model to learn hierarchically geometric and spatial features from multi-view inputs for hand shape reconstruction. To validate its effectiveness, two variants of our method were proposed, which adopt a single CPN and multi-scale CPNs to capture geometric and spatial context, respectively. The variant of using a single CPN inputs all of views into a CPN to learn features; the multi-scale CPNs are described in Section III-D. The experimental results are reported in Table VII, from which we can see that the multi-scale CPNs outperform a single CPN in terms of all metrics. This implies that multi-scale CPNs are able to learn more robust geometric and spatial features from different dimensions and different number of views compared to a single CPN.

TABLE VII: Comparisons of reconstruction errors with a single CPN and multi-scale CPNs feature learning methods (Unit:mm)

	Without MMFA	With MMFA
μ	0.525	**0.401**
σ	0.179	**0.099**
max	1.343	**0.598**

TABLE VIII: Comparisons of reconstruction errors with and without point-based refinement (Unit:mm)

	Without refinement	With refinement
μ	0.440	**0.401**
σ	0.143	**0.099**
max	1.113	**0.598**

4) Effectiveness of point-based refinement strategy: We also proposed a point-based refinement strategy to obtain an accurate reconstructed hand shape. Table VIII compares the reconstruction errors with and without the refinement step. It can be seen that the proposed point-based refinement helps to improve the accuracy of our result significantly. This is mainly because the self-transformer in the refinement concentrates on the local spatial context between each point and its neighbors to further optimize the position of each point.

V. Conclusions

In this work, we have proposed a novel deep learning-based framework, PatientHandNet, to reconstruct the open-palm hand shape from sparse multi-view depth images captured by a single commodity depth camera. Compared to the existing methods, our method has the following advantages: 1) we only require four depth images as input acquired with a single depth camera; 2) the subjects are allowed to change their hand poses during data acquisition; and 3) we output a high-fidelity hand mesh in a canonical open-palm pose. This makes it accurate and convenient for hand biometrics extraction of patients with impaired hand functions. We also synthesized a large-scale dataset for training the proposed model and built a novel real-world dataset that includes multi-view partial hand scans of 18 subjects and their ground-truth hand biometrics. Extensive results based on both synthetic data and real-world scans validated the effectiveness of our method and showed that it outperforms the state-of-the-art methods.

VI. Acknowledgments

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

References

[1] C.-C. Kuo, H.-Y. Kung, H.-C. Wu, and M.-J. Wang, "Developing a hand sizing system for a hand exoskeleton device based on the kansei engineering method," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–13, 2020.

[2] H. K. Yap, P. M. Khin, T. H. Koh, Y. Sun, X. Liang, J. H. Lim, and C.-H. Yeow, "A fully fabric-based bidirectional soft robotic glove for assistance and rehabilitation of hand impaired patients," *IEEE Robotics and Automation Letters*, vol. 2, no. 3, pp. 1383–1390, 2017.

[3] N. P. Oess, J. Wanek, and A. Curt, "Design and evaluation of a low-cost instrumented glove for hand function assessment," *Journal of neuroengineering and rehabilitation*, vol. 9, no. 1, pp. 1–11, 2012.

[4] N. E. Krausz, R. A. Rorrer *et al.*, “Design and fabrication of a six degree-of-freedom open source hand,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 5, pp. 562–572, 2015.

[5] Z. Xu and E. Todorov, “Design of a highly biomimetic anthropomorphic robotic hand towards artificial limb regeneration,” in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3485–3492.

[6] D. Magu, A. Aggarwal, P. Behera, and A. Khurana, “Use of finger length ratio as a marker for knee osteoarthritis: A case-control study of 2,456 patients,” *medRxiv*, 2020.

[7] S. M. Hussain, Y. Wang, D. C. Muller, A. E. Wluka, G. G. Giles, J. T. Manning, S. Graves, and F. M. Cicuttini, “Association between index-to-ring finger length ratio and risk of severe knee and hip osteoarthritis requiring total joint replacement,” *Rheumatology*, vol. 53, no. 7, pp. 1200–1207, 2014.

[8] X. Li, R. Wen, Z. Shen, Z. Wang, K. D. K. Luk, and Y. Hu, “A wearable detector for simultaneous finger joint motion measurement,” *IEEE transactions on biomedical circuits and systems*, vol. 12, no. 3, pp. 644–654, 2018.

[9] S. Tang, E. P. Sabonghy, A. Chaudhry, P. S. Shajudeen, M. T. Islam, N. Kim, F. J. Cabrera, J. Reddy, E. Tasciotti, and R. Righetti, “A model-based approach to investigate the effect of a long bone fracture on ultrasound strain elastography,” *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2704–2717, 2018.

[10] E. B. Brokaw, I. Black, R. J. Holley, and P. S. Lum, “Hand spring operated movement enhancer (handsome): a portable, passive hand exoskeleton for stroke rehabilitation,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 19, no. 4, pp. 391–399, 2011.

[11] S. H. Nasir and O. Troynikov, “Influence of hand movement on skin deformation: A therapeutic glove design perspective,” *Applied Ergonomics*, vol. 60, pp. 154–162, 2017.

[12] Y. Chen, Z. Tu, D. Kang, L. Bao, Y. Zhang, X. Zhe, R. Chen, and J. Yuan, “Model-based 3d hand reconstruction via self-supervised learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 451–10 460.

[13] X. Tang, T. Wang, and C.-W. Fu, “Towards accurate alignment in real-time 3d hand-mesh reconstruction,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 698–11 707.

[14] P. Chen, Y. Chen, D. Yang, F. Wu, Q. Li, Q. Xia, and Y. Tan, “I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 929–12 938.

[15] P. Volonghi, G. Baronio, and A. Signoroni, “3d scanning and geometry processing techniques for customised hand orthotics: an experimental assessment,” *Virtual and Physical Prototyping*, vol. 13, no. 2, pp. 105–116, 2018.

[16] D. Koutny, D. Palousek, T. Koutecky, A. Zatocilova, J. Rosicky, and M. Janda, “3d digitalization of the human body for use in orthotics and prosthetics,” *International Journal of Biomedical and Biological Engineering*, vol. 6, no. 12, pp. 690–697, 2012.

[17] M. Carfagni, R. Furferi, L. Governi, M. Servi, F. Uccheddu, Y. Volpe, and K. Mcgreevy, “Fast and low cost acquisition and reconstruction system for human hand-wrist-arm anatomy,” *Procedia Manufacturing*, vol. 11, pp. 1600–1608, 2017.

[18] B. Ceulemans, S.-P. Lu, G. Lafruit, and A. Munteanu, “Robust multiview synthesis for wide-baseline camera arrays,” *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2235–2248, 2018.

[19] A. Yu, K. L. Yick, S. Ng, and J. Yip, “2d and 3d anatomical analyses of hand dimensions for custom-made gloves,” *Applied ergonomics*, vol. 44, no. 3, pp. 381–392, 2013.

[20] X. Deng, Y. Zhu, Y. Zhang, Z. Cui, P. Tan, W. Qu, C. Ma, and H. Wang, “Weakly supervised learning for single depth-based hand shape recovery,” *IEEE Transactions on Image Processing*, vol. 30, pp. 532–545, 2020.

[21] P. Hu, N. N. Kaashki, V. Dadarlat, and A. Munteanu, “Learning to estimate the body shape under clothing from a single 3-d scan,” *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 3793–3802, 2020.

[22] R. Zhao, X. Dai, P. Hu, and A. Munteanu, “Posenormnet: Identity-preserved posture normalization of 3d body scans in arbitrary postures,” *IEEE transactions on industrial informatics*, 2023.

[23] S. E. Ovur, H. Su, W. Qi, E. De Momi, and G. Ferrigno, “Novel adaptive sensor fusion methodology for hand pose estimation with multileap motion,” *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–8, 2021.

[24] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt, “Ganerated hands for real-time 3d hand tracking from monocular rgb,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 49–59.

[25] A. El-Sawah, N. D. Georganas, and E. M. Petriu, “A prototype for 3-d hand tracking and posture estimation,” *IEEE Transactions on Instrumentation and Measurement*, vol. 57, no. 8, pp. 1627–1636, 2008.

[26] Y. Cai, L. Ge, J. Cai, N. M. Thalmann, and J. Yuan, “3d hand pose estimation using synthetic data and weakly labeled rgb images,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 3739–3753, 2020.

[27] L. Ge, H. Liang, J. Yuan, and D. Thalmann, “Robust 3d hand pose estimation from single depth images using multi-view cnns,” *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4422–4436, 2018.

[28] B. Wang, Y. Jin, Y. Chen, Z. Sun, M. Duan, H. Chen, X. Fan, and J. Zheng, “Gaze tracking 3d reconstruction of object with large-scale motion,” *IEEE Transactions on Instrumentation and Measurement*, 2023.

[29] L. Ge, H. Liang, J. Yuan, and D. Thalmann, “Real-time 3d hand pose estimation with 3d convolutional neural networks,” *IEEE transactions on pattern analysis and*

machine intelligence, vol. 41, no. 4, pp. 956–970, 2018.

[30] F. Albiol, A. Corbi, and A. Albiol, "Geometrical calibration of x-ray imaging with rgb cameras for 3d reconstruction," *IEEE transactions on medical imaging*, vol. 35, no. 8, pp. 1952–1961, 2016.

[31] J. Malik, A. Elhayek, F. Nunnari, K. Varanasi, K. Tamaddon, A. Heloir, and D. Stricker, "Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 110–119.

[32] D. Kong, L. Zhang, L. Chen, H. Ma, X. Yan, S. Sun, X. Liu, K. Han, and X. Xie, "Identity-aware hand mesh estimation and personalization from rgb images," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*. Springer, 2022, pp. 536–553.

[33] Z. Yu, C. Li, L. Yang, X. Zheng, M. B. Mi, G. H. Lee, and A. Yao, "Overcoming the trade-off between accuracy and plausibility in 3d hand shape reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 544–553.

[34] T. Luan, Y. Zhai, J. Meng, Z. Li, Z. Chen, Y. Xu, and J. Yuan, "High fidelity 3d hand shape reconstruction via scalable graph frequency decomposition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 795–16 804.

[35] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *arXiv preprint arXiv:2201.02610*, 2022.

[36] J. Park, Y. Oh, G. Moon, H. Choi, and K. M. Lee, "Handoccnet: Occlusion-robust 3d hand mesh estimation network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1496–1505.

[37] X. Chen, Y. Liu, Y. Dong, X. Zhang, C. Ma, Y. Xiong, Y. Zhang, and X. Guo, "Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 544–20 554.

[38] J. Malik, S. Shimada, A. Elhayek, S. A. Ali, C. Theobalt, V. Golyanik, and D. Stricker, "Handvoxnet++: 3d hand shape and pose estimation using voxel-based neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8962–8974, 2021.

[39] C. Liu, D. Kong, S. Wang, J. Li, and B. Yin, "Dlgan: depth-preserving latent generative adversarial network for 3d reconstruction," *IEEE Transactions on Multimedia*, vol. 23, pp. 2843–2856, 2020.

[40] 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.

[41] 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.

[42] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. Spie, 1992, pp. 586–606.

[43] 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.

[44] Y. Rong, T. Shiratori, and H. Joo, "Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1749–1759.

[45] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 975–10 985.

[46] 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.

[47] M. Gschwandtner, R. Kwitt, A. Uhl, and W. Pree, "Blensor: Blender sensor simulation toolbox," in *International Symposium on Visual Computing*. Springer, 2011, pp. 199–208.

[48] Y. Yang, H. Zhou, Y. Song, and P. Vink, "Identify dominant dimensions of 3d hand shapes using statistical shape model and deep neural network," *Applied Ergonomics*, vol. 96, p. 103462, 2021.

[49] F. Arifi, M. S. AmetiV, and Z. Metaj, "Stature and its estimation utilizing length of hand measurements of both gender adolescents from western region of kosovo," *Sports Injr Med*, vol. 5, p. 171, 2021.

[50] P. Xiang, X. Wen, Y.-S. Liu, Y.-P. Cao, P. Wan, W. Zheng, and Z. Han, "Snowflakenet: Point cloud completion by snowflake point deconvolution with skip-transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5499–5509.

[51] P. Hu, R. Zhao, X. Dai, and A. Munteanu, "Predicting high-fidelity human body models from impaired point clouds," *Signal Processing*, vol. 192, p. 108375, 2022.

[52] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 605–613.

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.

Pengpeng Hu is 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, and MEASUREMENT: SENSORS. He is also 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.

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.