MeasureXpert: Automatic Anthropometric Measurement Extraction from Two Unregistered, Partial, Posed, and Dressed Body Scans

Supplementary Material

6. Detailed Architecture

In the MeasureXpert pipeline (Figure 1), we employ four encoder blocks (within light grey rectangles), two multi-decoder blocks (within yellow rectangles), and five decoder blocks (consisting of three green rectangular slabs. Each green rectangular slab represents a fully connected layer). All encoder blocks share a common structure, as do the sub-decoders within the multi-decoders and the other decoders. We begin by introducing the encoder and decoder structures, followed by a step-by-step formulation of the overall architecture. Then, we establish the body segmentation and introduce the corresponding notations to help understand multi-decoder blocks. Lastly, we provide a detailed formulation of the loss functions.

6.1. Encoder

The encoder process begins with a specific point set passing through an MLP consisting of two layers with 128 and 256 neurons, denoted as MLP_{128}^{256}. Following this MLP, a max pooling operation is applied to obtain a global feature vector of the input point set. To enrich the feature representation, this vector is concatenated with the pre-pooling output from the MLP. And the concatenated features are processed by a second MLP with layers of 512 and 1024 neurons, denoted as MLP_{512}^{1024}. A final max pooling operation solidifies the ultimate global feature.

6.2. Decoder

The decoder consists of three fully connected layers. The first two layers consist of 1024 neurons, denoted as FC_1^{1024} and FC_2^{1024}, respectively, while the third layer has neurons depending on what we want to output. If we want to obtain a point set, the number of neurons is equal to three times the number of points N (to account for the x, y, and z coordinates), denoted as $FC_3^{3\times N}$, while if we want to output values, the number of neurons is the number of values, denoted as $FC_3^{|values|}$. Additionally, if the last layer is $FC_3^{3\times N}$, a reshape operation is executed to output the 3D coordinates of the expected point set.

6.3. TrioNet

Encoder. TrioNet proceeds from two partial point clouds $\mathcal{S}^f$ and $\mathcal{S}^b$ of one individual from front- and back-view, respectively. The shared encoder function $\mathcal{E}$ with corresponding weights (Eq. 2) follows the following steps:

$$MLP_{128}^{256}(\mathcal{S}^i) = \Omega_1^i, i \in \{f, b\} \tag{10}$$

$$Maxpool(\Omega_1^i) = \Omega^i, i \in \{f, b\} \tag{11}$$

$$Maxpool(MLP_{512}^{1024}([\Omega_1^i, \Omega^i])) = \mathcal{F}^i, i \in \{f, b\} \tag{12}$$

extracting the global features $\mathcal{F}^f$ and $\mathcal{F}^b$, where Ω_1^i represents the local features, Ω^i denotes the global feature, and $\mathcal{F}^i$ is the final feature of the input point cloud.

Input branches. Each sub-decoder $\mathcal{D}'$ via corresponding weights to decode the global feature to the corresponding posed body point cloud following:

$$\begin{aligned}\mathcal{D}'(\mathcal{F}^i|\omega_x^i) &= FC_3^{3\times N_x}(FC_2^{1024}(FC_1^{1024}(\mathcal{F}^i))).reshape(-1,3)\\ &= \widetilde{\mathcal{S}}_x^i, \qquad i \in \{f, b\}\end{aligned} \tag{13}$$

where $x \in \{T, H, RA, LA, RL, LL\}$ represents the specific segment of body $\widetilde{\mathcal{S}}_x^i$ (refer to Section 6.5), while ω_x^i is the weights for $\widetilde{\mathcal{S}}_x^i$ decoder and $N_x = |\widetilde{\mathcal{S}}_x^i|$. All six segments form the comlpete posed body point cloud $\widetilde{\mathcal{S}}^i$, $i \in \{f, b\}$.

Shape branch. The shape branch initiates with USV extraction (Algorithm 1) to extract $\mathcal{F}_s$. In this process, the shared regression step utilizes a decoder block with corresponding weights following:

$$Regression(\mathcal{F}^f) = FC_3^{512}(FC_2^{1024}(FC_1^{1024}(\mathcal{F}^f))) = \vec{\omega}^f, \tag{14}$$

$$Regression(\mathcal{F}^b) = FC_3^{512}(FC_2^{1024}(FC_1^{1024}(\mathcal{F}^b))) = \vec{\omega}^b. \tag{15}$$

to obtain the two PIVs $\vec{\omega}^f$ and $\vec{\omega}^b$.

The extracted USV $\mathcal{F}_s$, extracted via Algorithm 1, is fed into a decoder block to learn the function Eq. 3 mentioned in Section 3.1:

$$\begin{aligned}\mathcal{D}(\mathcal{F}_s) &= FC_3^{3\times 10364}(FC_2^{1024}(FC_1^{1024}(\mathcal{F}_s))).reshape(-1,3)\\ &= [\mathcal{T}_c, \mathcal{L}mk_c]^T := \mathcal{TL}_c.\end{aligned} \tag{16}$$

Therefore, the reshaped output has 10364 points, in which the first 6890 points are T-posed body mesh vertices $\mathcal{T}_c$, and the last 3456 points constitute 21 groups of landmarks $\mathcal{L}mk_c$ (refer to Section 7).

6.4. OR-Net

The offset prediction takes $\mathcal{TL}_c$ as input and outputs the offsets to be added to $\mathcal{TL}_c$, yielding refined coordinates $\mathcal{TL}$

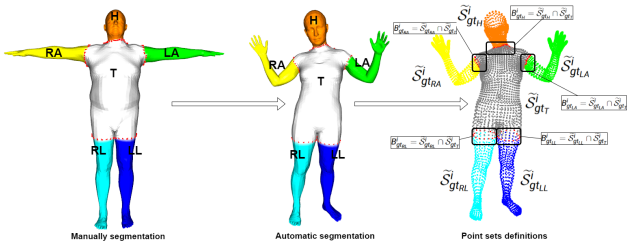

Figure 6. Examples of body segmentation and corresponding notations.

for the vertices $\mathcal{T}$ and landmarks $\mathcal{L}mk$ of the T-posed mesh following:

$$MLP_{128}^{256}(\mathcal{TL}_c) = \Omega_{TL_1}, \tag{17}$$

$$Maxpool(\Omega_{TL_1}) = \Omega_{TL}, \tag{18}$$

$$Maxpool(MLP_{512}^{1024}([\Omega_{TL_1}, \Omega_{TL}])) = \mathcal{F}_{TL}, \tag{19}$$

$$\mathcal{TL} = \mathcal{TL}_c + FC_3^{3\times 10364}(FC_2^{1024}(FC_1^{1024}(\mathcal{F}_{TL}))).reshape(-1,3) \tag{20}$$

where Ω_{TL_1} represents the local features, Ω_{TL} denotes the global feature, and $\mathcal{F}_{TL}$ is the final feature of $\mathcal{TL}_c$.
The refined landmarks, denoted as $\mathcal{L}mk = \{l_i\}_{i=1}^{21}$, consist of 21 groups of points that are then fed into the subsequent encoder-decoder regression module to output measurement values following:

$$MLP_{128}^{256}(l_i) = \Omega_{l_i 1}, \tag{21}$$

$$Maxpool(\Omega_{l_i 1}) = \Omega_{l_i}, \tag{22}$$

$$Maxpool(MLP_{512}^{1024}([\Omega_{l_i 1}, \Omega_{l_i}])) = \mathcal{F}_{l_i}, \tag{23}$$

$$Regression(\mathcal{F}_{l_i}) = FC_3^1(FC_2^{1024}(FC_1^{1024}(\mathcal{F}_{l_i}))) = V_i, \tag{24}$$

where $i \in \{1, 2, ..., 21\}$, $\Omega_{l_i 1}$ represents the local features, Ω_{l_i} denotes the global feature, and $\mathcal{F}_{l_i}$ is the final feature of l_i.

6.5. Body segmentation

As Figure 6 shows, a T-posed SMLP mesh can be manually segmented into six different segments $P = \{T, H, LA, RA, LL, RL\}$, where T is the torse, H is the head, LA is the left arm with hand, RA is the right arm with hand, LL is the left leg with foot, and RL is the right leg with foot. Considering all SMLP meshes have the same topology, based on the previous manual segmentation, an arbitrary SMLP body can be segmented automatically. Taking the ground-truth posed body point sets $\widetilde{\mathcal{S}}_{gt}^i = \{\tilde{s}_{gt_j}^i | j = 1, 2, \ldots, \widetilde{M}_{gt}^i\}$, $i \in \{f, b\}$, as example, after segmentation, we classified $\widetilde{\mathcal{S}}_{gt}^i$ into: the torse point set $\widetilde{\mathcal{S}}_{gt_T}^i$, the head point set $\widetilde{\mathcal{S}}_{gt_H}^i$, the left arm with hand point set $\widetilde{\mathcal{S}}_{gt_{LA}}^i$, the right arm with hand point set $\widetilde{\mathcal{S}}_{gt_{RA}}^i$, the left leg with foot point set $\widetilde{\mathcal{S}}_{gt_{LL}}^i$, and the right leg with foot point set $\widetilde{\mathcal{S}}_{gt_{RL}}^i$. The point $\tilde{s}_{gt_x,j}^i$ in each segment $\widetilde{\mathcal{S}}_{gt_x}^i$ can be classified into two types: common points and boundary points. Common points are exclusive to $\widetilde{\mathcal{S}}_{gt_x}^i$, while boundary points are shared by $\widetilde{\mathcal{S}}_{gt_x}^i$ and $\widetilde{\mathcal{S}}_{gt_y}^i$, where $x \neq y$. We use $B_{gt_x}^i$ to represent the set of boundary point pairs $(b_{gt_x,j}^i, b_{gt_T,j}^i)$, where $B_{gt_x}^i = \widetilde{\mathcal{S}}_{gt_x}^i \cap \widetilde{\mathcal{S}}_{gt_T}^i$, $x \in P \setminus \{T\}$, and $B_{gt_x}^i \in B_{gt}^i = \{\widetilde{\mathcal{S}}_{gt_H}^i \cap \widetilde{\mathcal{S}}_{gt_T}^i, \widetilde{\mathcal{S}}_{gt_{RA}}^i \cap \widetilde{\mathcal{S}}_{gt_T}^i, \widetilde{\mathcal{S}}_{gt_{LA}}^i \cap \widetilde{\mathcal{S}}_{gt_T}^i, \widetilde{\mathcal{S}}_{gt_{RL}}^i \cap \widetilde{\mathcal{S}}_{gt_T}^i, \widetilde{\mathcal{S}}_{gt_{LL}}^i \cap \widetilde{\mathcal{S}}_{gt_T}^i\}$. Each pair includes two points: the first point, $b_{gt_x,j}^i$, is a boundary point of one specific body part excluding the torso, and the second point, $b_{gt_T,j}^i$, is a corresponding boundary point on the torso.

6.6. Loss functions

We give the formulations of the loss functions mentioned in Section 6.6.
TrioNet. For two input branches, they share the same loss functions: local loss (Eq. 25), global loss (Eq. 26), and inter-connective loss (Eq. 27).

$$\mathcal{L}_{part}^i = \sum_{x \in P} \frac{1}{|\widetilde{\mathcal{S}}_x^i|} \sum_{\tilde{s}_{x_j}^i \in \widetilde{\mathcal{S}}_x^i} ||\tilde{s}_{x_j}^i - \tilde{s}_{gt_x,j}^i||^2, \tag{25}$$

$$\begin{aligned}\mathcal{L}_{global}^i = &\frac{1}{|\widetilde{\mathcal{S}}^i|} \sum_{\tilde{s}_j^i \in \widetilde{\mathcal{S}}^i} \min_{\tilde{s}_{gt_j}^i \in \widetilde{\mathcal{S}}_{gt}^i} \|\tilde{s}_j^i - \tilde{s}_{gt_j}^i\|^2 \\ &+ \frac{1}{|\widetilde{\mathcal{S}}_{gt}^i|} \sum_{\tilde{s}_{gt_j}^i \in \widetilde{\mathcal{S}}_{gt}^i} \min_{\tilde{s}_j^i \in \widetilde{\mathcal{S}}^i} \|\tilde{s}_{gt_j}^i - \tilde{s}_j^i\|^2,\end{aligned} \tag{26}$$

$$\mathcal{L}_{boundary}^i = \sum_{x \in P \setminus \{T\}} \frac{1}{|B_x^i|} \sum_{(b_{x_j}^i, b_{T_j}^i) \in B_x^i} ||b_{x_j}^i - b_{T_j}^i||^2, \tag{27}$$

where $P = \{T, H, LA, RA, LL, RL\}$ is the body segments set, $\tilde{s}_{x_j}^i \in \widetilde{\mathcal{S}}_x^i$ represents the predicted body segment point corresponding to the ground-truth point $\tilde{s}_{gt_x,j}^i \in \widetilde{\mathcal{S}}_{gt_x}^i$, $\widetilde{\mathcal{S}}^i$ denotes the set of all predicted body points, and $(b_{x_j}^i, b_{T_j}^i) \in B_x^i$ is the boundary point pair corresponding to the ground-truth pair $(b_{gt_x,j}^i, b_{gt_T,j}^i) \in B_{gt_x}^i$. Additionally, $i \in \{f, b\}$ indicates whether the target of supervision for the loss is the front (f) or the back (b) of the body.
For the shape branch, we formulate $\mathcal{L}_{tpose}^c$ and $\mathcal{L}_{lmk}^c$ as following:

$$\mathcal{L}_{tpose}^c = \frac{1}{|\mathcal{T}_c|} \sum_{i=1}^{|\mathcal{T}_c|} ||v_i' - v_{gt_i}||^2, \tag{28}$$

$$\mathcal{L}^c_{lmk} = \sum_{i=1}^{21} \frac{1}{|l'_i|} \sum_{x \in l'_i} \min_{y \in l_{gt_i}} \|x-y\|^2 + \frac{1}{|l_{gt_i}|} \sum_{y \in l_{gt_i}} \min_{x \in l'_i} \|y-x\|^2, \tag{29}$$

where $v_{gt_i} \in \mathcal{T}_{gt}$ is a vertex in the ground-truth T-posed body mesh corresponding to the predicted vertex $v'_i \in \mathcal{T}_c$, while $l_{gt_i} \in \mathcal{L}mk_{gt}$ is the ground-truth point set corresponding to the predicted point set of one level landmarks $l'_i \in \mathcal{L}mk_c$.

OR-Net. The loss functions to supervise mesh $\mathcal{L}_{tpose}$ and landmarks $\mathcal{L}_{lmk}$ optimization are similar to previous prediction loss functions $\mathcal{L}^c_{tpose}$ and $\mathcal{L}^c_{lmk}$, respectively.

$$\mathcal{L}_{tpose} = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} ||v_i - v_{gt_i}||^2, \tag{30}$$

$$\mathcal{L}_{lmk} = \sum_{i=1}^{21} \frac{1}{|l_i|} \sum_{x \in l_i} \min_{y \in l_{gt_i}} \|x-y\|^2 + \frac{1}{|l_{gt_i}|} \sum_{y \in l_{gt_i}} \min_{x \in l_i} \|y-x\|^2, \tag{31}$$

where $v_i \in \mathcal{T}$ is a refined vertex of the predicted T-posed mesh, while $l_i \in \mathcal{L}mk$ is a refined point set of one level predicted landmarks. Beside $\mathcal{L}_{tpose}$ and $\mathcal{L}_{lmk}$, OR-Net has another L1 loss to supervise value predictions defined as:

$$\mathcal{L}_{value} = \sum_{i=1}^{21} ||V_i - V_{gt_i}||, \tag{32}$$

where $V_{gt_i} \in V_{gt}$ represents the ground-truth measurement value corresponding to $V_i \in V$.

7. Datasets

In this section, we first present the detailed synthesizing process of the BWM dataset, utilized for training, validation, and testing. Then we outline the extension process applied to two open-access real-world datasets, FAUST [6] and 4D-Dress [59].

7.1. BWM dataset

Following Figure 2, the synthesizing process of the BWM has four steps: i) unclothed bodies generation, ii) dressing the posed unclothed bodies, iii) rendering the dressed bodies, and iv) measurement annotation.

Unclothed bodies. We initiate our process by utilizing the SMPL model to generate synthetic unclothed human body meshes. We utilize the extensive pose and shape parameters provided by the SURREAL dataset [57] in our generation process. By randomly combining these pose and shape parameters, we create tuples of human bodies, each consisting of three meshes: two with randomly arbitrary poses $\widetilde{\mathcal{S}}^f_{gt}$ and $\widetilde{\mathcal{S}}^b_{gt}$, and one in a canonical "T" pose $\mathcal{T}_{gt}$. Notably, all bodies within a tuple share identical shape parameters but differ in pose parameters.

Dressing. To simulate clothed human bodies, we adopt the method proposed in the BUG dataset [20]. We apply the same clothing mesh to the two posed bodies within each tuple, effectively generating clothed body models while maintaining consistent underlying body shapes.

Rendering partial point clouds. Subsequently, we use BlenSor [18] to simulate the scanning process, rendering partial point clouds $\mathcal{S}^f$ and $\mathcal{S}^b$ from the front and back views of the clothed, posed bodies, respectively. We employ a single-camera scanner that incorporates inherent noise perturbations typical of real-world depth images.

Ground-truth measurement annotation. In accordance with the ISO 8559 standard [25], we establish 16 measurement levels on the bodies in the canonical pose (see Figure 8). These levels include the bust, underbust, waist, hip, middle thighs, knees, calves, upper arms, elbows, and wrists. For fashion design, the measurement should also consider the preferred waist level [49], which means the level a person would prefer to wear the waist of his or her pants. Additionally, we further define five more measurement levels in the abdominal region as Figure 2 shows, centrally located on the torso, to accommodate these discrepancies. Therefore, we establish 21 measurement levels in total.

We employ plane at each measurement level to segment the unclothed body in the canonical pose to obtain the measurement landmarks $\mathcal{L}mk_{gt} = \{l_{gt_i}\}_{i=1}^{21}$. To facilitate training, we manually assigned a fixed number of landmarks of each level through random sampling based on the variation in girth. Specifically, 256 landmark points are allocated for the bust, underbust, hip, and six waist levels; 128 landmark points are assigned to the legs, including two mid-thigh, two knee, and two calf levels; and 64 landmark points are designated for the arms, covering two upper arm, two elbow, and two wrist levels, resulting in a total of 3456 landmark points. We then extract precise anthropometric measurements $V_{gt} = \{V_{gt_i}\}_{i=1}^{21}$ via the function Convex Hull in OpenCV [8] working on the corresponding landmark point set $\{l_{gt_i}\}_{i=1}^{21}$.

In summary, each sample in the BWM includes three unclothed body models: two in different postures and one in a canonical posture, accompanied by two clothed body mesh with scans (front and back views) corresponding to the posed unclothed bodies and measurement landmarks with ground-truth values. In addition to estimating body shape under clothing, the proposed dataset is also suitable for investigating other research problems such as non-rigid point cloud registration and point cloud completion. In this study, we generated $150K$ male and $150K$ female data groups. We split the dataset into 99% for training, 0.3% for

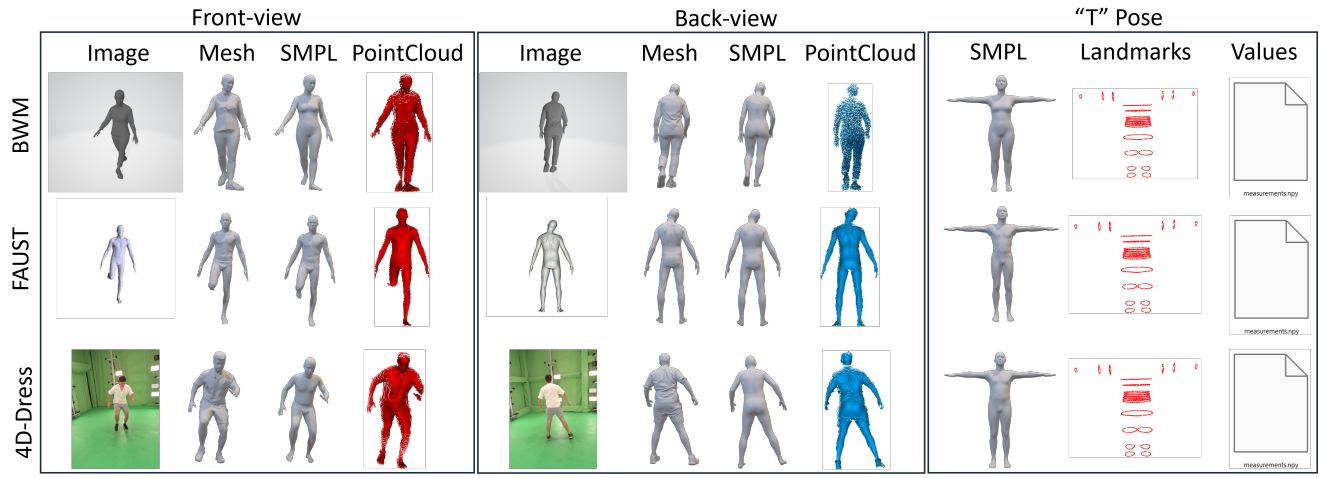

Figure 7. Examples of data we used in our experiments from BWM, extended FAUST and 4D-Dress.

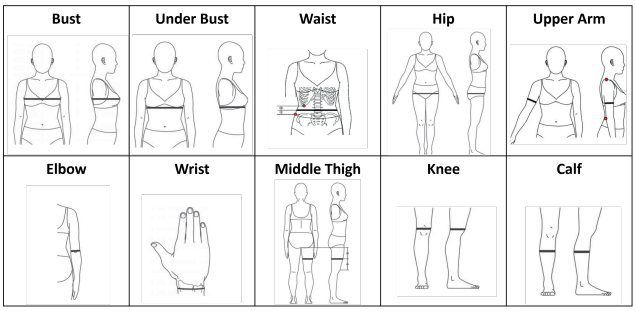

Figure 8. Measurement definitions from ISO8559 [25]. Besides the first four definitions on the torse, the other six definitions are defined on both the right and left sides.

testing, and 0.7% for validation.

7.2. Real-world datasets extension

4D-Dress. The 4D-Dress dataset [59] consists of high-resolution 4D scans of clothed humans captured over time, providing dynamic image and mesh sequences that represent body movements and clothing deformations. In addition to the dressed body meshes, the dataset includes underlying SMPL body models with associated pose and shape parameters. However, unlike the BWM dataset, 4D-Dress does not provide body models in a canonical posture, partial dressed body scans, or anthropometric measurements. To obtain the underlying body model in a canonical T-pose, all pose parameters can simply be set to zero within the SMPL model. For the scanning and measurement, we adopt the RealPartialScan [22] method to extract the front and back views from the dressed meshes and OpenCV to calculate the corresponding anthropometric values.

FAUST. The FAUST dataset [6] contains high-resolution 3D scans of 10 unclothed human bodies in various poses, along with their corresponding fitted SMPL models. However, FAUST does not provide the SMPL parameters (shape and pose), preventing direct acquisition of T-posed mesh for each subject by simply setting the pose parameters to zero. To overcome this, we employ PoseNormNet to perform posture normalization on the posed SMPL models for each subject, bringing them into a canonical posture ("T" pose). Once normalized, we apply the Iterative Closest Point (ICP) to align all the normalized models for the same subject. We then compute the average coordinates for each vertex across all the normalized models to obtain a "mean" model, which serves as the canonical T-posed mesh for the subject. For anthropometric measurements, we follow the approach used in the BWM dataset, measuring each normalized model at predefined levels. The mean values across all models at each measurement level are taken as the subject's anthropometric values. For the partial scans, we also use the RealPartialScan [22] method to extract the front and back views from the real-world meshes and export the resulting point clouds.

8. Experiments

This section of supplementary materials is organized into three parts: (i) additional content for comparative experiments, (ii) demonstration of the by-products generated by this method, and (iii) extended ablation studies and results, providing a more comprehensive verification of the method's scientific validity and effectiveness.

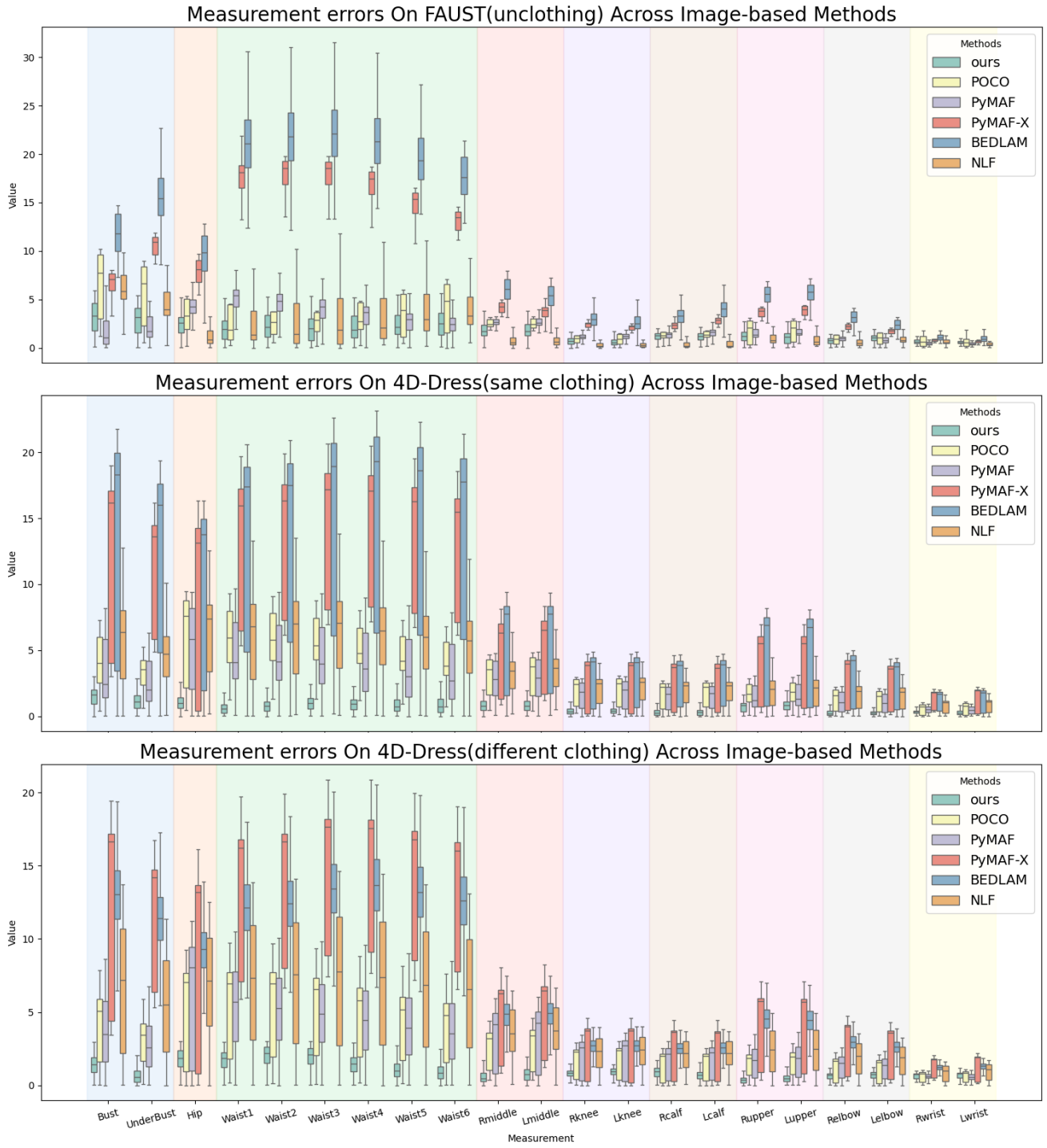

Figure 9. These figures present box plots of absolute errors (in cm) for body measurements on the FAUST and 4D-Dress datasets across different image-based methods: POCO [15], NLF [53], PyMAF-X [65], BEDLAM [5] PyMAF [64], and ours.

8.1. Evaluation Metrics

As supplementary materials for Section 4, we continue to use absolute errors to evaluate the error of the measured values. For ablation studies, in addition to Chamfer Distance, we incorporate Mean Squared Error (MSE) and Mean Absolute Error (MAE) to provide a more comprehensive and

Measurement	Ours		BEDLAM [5]		NLF [53]		POCO [15]		PyMAF [64]		PyMAF-X [65]	
Unit (CM)	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
bust	3.25	**1.62**	13.33	5.68	7.3	4.42	9.02	7.32	**2.93**	5.23	7.1	3.58
hip	2.38	**1.24**	10.52	5.78	**1.93**	2.45	4.34	3.74	4.78	2.38	7.14	2.75
left calf	1.12	**0.52**	3.74	1.45	**0.58**	0.71	1.54	0.85	1.67	0.7	2.52	1
left elbow	**1**	**0.43**	2.66	1.32	1.05	0.76	1.36	1.3	**0.97**	0.85	1.89	1.17
left knee	0.68	**0.46**	2.7	1.45	**0.59**	0.76	1.17	1.05	1.31	0.64	1.88	0.74
left middle thigh	1.79	**0.92**	5.88	3.45	**1.16**	1.36	3.19	1.78	3	1.36	3.58	1.02
left upper arm	**1.06**	**0.64**	6.48	2.94	1.23	1.53	2.54	2.58	1.98	1.61	3.8	1.76
left wrist	0.58	**0.21**	1.01	0.46	0.54	0.44	0.69	0.62	**0.5**	0.37	0.74	0.49
right calf	1.15	**0.48**	3.23	1.38	**0.56**	0.74	1.45	0.85	1.47	0.68	2.14	0.8
right elbow	**0.73**	**0.38**	3.46	1.52	0.86	0.95	1.47	1.55	1.17	0.96	2.21	1.11
right knee	0.72	**0.45**	3.01	1.38	**0.59**	0.81	1.19	0.99	1.28	0.63	2.11	0.85
right middle thigh	1.77	**0.88**	6.43	3.38	**1.15**	1.4	3.22	1.74	3.05	1.36	3.82	1.24
right upper arm	**1.2**	**0.59**	6.29	2.97	1.41	1.57	2.5	2.78	1.83	1.68	3.65	1.66
right wrist	0.65	**0.23**	1.07	0.34	0.73	0.42	0.77	0.62	**0.54**	0.35	0.82	0.45
under bust	**2.95**	**1.5**	15.93	4.3	6.06	4.5	8.18	7.5	3.21	4.92	10.81	4.12
waist1	**1.99**	**1.4**	21.06	5.99	3.23	4.43	5.09	6.63	5.71	3.83	16.82	4.9
waist2	**2.3**	**1.44**	21.51	5.71	3.41	4.58	5.67	7.4	5.42	4.38	17.63	5.24
waist3	**2.09**	**1.47**	22.15	5.92	3.85	5.09	6.34	8.33	5.09	4.92	17.73	5.17
waist4	**2.17**	**1.42**	21.8	5.97	4.46	5.3	6.67	8.5	4.59	5.03	16.82	5.04
waist5	**2.33**	**1.4**	20.06	5.64	5.02	5.08	6.92	8.31	4.01	5.02	15.01	4.84
waist6	**2.49**	**1.43**	18.47	5.43	5.36	4.73	7.09	8.03	3.57	4.97	13.38	4.78

Table 1. Comparison of measurement errors across different image-based methods on FAUST.

Measurement	Ours		BEDLAM [5]		NLF [53]		POCO [15]		PyMAF [64]		PyMAF-X [65]	
Unit (CM)	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
bust	**1.55**	**0.76**	12.32	8.49	5.47	3.23	4.18	1.89	3.29	2.3	11.1	6.52
hip	**1.04**	**0.58**	9	6.53	6.1	3.21	5.63	3.32	4.89	3.24	7.9	6.84
left calf	**0.35**	**0.31**	2.64	1.78	1.91	0.97	1.66	0.93	1.41	0.88	2.26	1.81
left elbow	**0.28**	**0.22**	2.44	1.81	1.45	0.9	1.16	0.76	0.96	0.65	2.23	1.73
left knee	**0.45**	**0.3**	2.69	1.93	2.13	1.06	1.84	1.09	1.56	1.06	2.32	1.96
left middle thigh	**0.82**	**0.45**	5.3	3.42	3.33	1.43	3.05	1.48	2.75	1.52	4.63	2.71
left upper arm	**0.81**	**0.38**	4.37	3.29	1.95	1.09	1.8	0.76	1.46	0.93	3.56	2.7
left wrist	**0.28**	**0.18**	1.17	0.84	0.81	0.52	0.55	0.44	0.45	0.29	1.18	0.89
right calf	**0.32**	**0.27**	2.57	1.79	1.85	0.98	1.59	0.99	1.34	0.89	2.29	1.82
right elbow	**0.29**	**0.27**	2.76	2.04	1.54	0.96	1.24	0.79	1.06	0.71	2.51	1.86
right knee	**0.41**	**0.3**	2.69	1.95	1.98	1.11	1.82	1.02	1.5	1.01	2.32	1.94
right middle thigh	**0.82**	**0.47**	5.22	3.5	3.16	1.44	3	1.35	2.72	1.44	4.35	2.78
right upper arm	**0.72**	**0.35**	4.45	3.34	1.89	1.07	1.7	0.73	1.39	0.9	3.61	2.66
right wrist	**0.35**	**0.15**	1.19	0.74	0.77	0.49	0.53	0.38	0.49	0.24	1.23	0.69
under bust	**1.13**	**0.66**	11.53	6.94	4.44	2.28	3.34	1.07	2.54	1.73	10.52	4.27
waist1	**0.67**	**0.56**	12.31	7.53	6.01	3.24	5.87	2.19	4.61	2.59	12.32	5.33
waist2	**0.8**	**0.55**	12.68	7.42	6.17	3.3	5.85	2.08	4.48	2.52	12.87	5.06
waist3	**0.97**	**0.55**	13.74	7.94	6.24	3.35	5.65	1.91	4.33	2.53	13.69	5.06
waist4	**0.92**	**0.56**	14.07	8.07	5.9	3.16	5.14	1.68	3.96	2.47	13.69	4.87
waist5	**0.88**	**0.63**	13.57	7.76	5.5	2.93	4.6	1.57	3.52	2.32	13	4.68
waist6	**0.9**	**0.69**	12.97	7.44	5.22	2.76	4.22	1.53	3.22	2.19	12.22	4.62

Table 2. Comparison of measurement errors across different image-based methods on 4D-Dress (same clothing).

Measurement	Ours		BEDLAM [5]		NLF [53]		POCO [15]		PyMAF [64]		PyMAF-X [65]	
Unit (CM)	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
bust	**1.38**	**0.71**	12.11	3.84	6.7	4.21	4.17	2.17	3.65	2.25	12.52	6.21
hip	**1.77**	**0.79**	8.09	3.7	7.03	3.23	5.12	3.33	6.03	4.02	9	6.29
left calf	**0.69**	**0.32**	2.27	0.97	2.15	0.92	1.43	0.92	1.68	1.08	2.45	1.68
left elbow	**0.7**	**0.26**	2.35	0.91	1.75	0.96	1.13	0.78	1.18	0.69	2.53	1.6
left knee	**0.91**	**0.29**	2.36	1.09	2.37	1	1.71	1.08	1.97	1.33	2.52	1.8
left middle thigh	**0.76**	**0.46**	4.31	2.02	3.81	1.57	2.66	1.41	3.28	2.07	4.9	2.49
left upper arm	**0.49**	**0.34**	4.05	1.57	2.41	1.44	1.62	0.78	1.67	1.05	4.06	2.51
left wrist	0.69	0.24	1.17	0.42	0.96	0.51	0.65	0.35	**0.57**	**0.22**	1.33	0.83
right calf	**0.87**	**0.4**	2.23	1	2.11	0.95	1.4	0.95	1.63	1.05	2.49	1.69
right elbow	**0.63**	**0.26**	2.65	1.01	1.87	1.05	1.18	0.82	1.3	0.75	2.84	1.73
right knee	**0.8**	**0.25**	2.38	1.04	2.25	1.06	1.67	1.01	1.9	1.25	2.56	1.74
right middle thigh	**0.57**	**0.44**	4.25	2.02	3.65	1.57	2.57	1.29	3.23	2	4.63	2.56
right upper arm	**0.39**	**0.3**	4.14	1.58	2.36	1.45	1.54	0.74	1.57	1	4.12	2.48
right wrist	0.66	0.18	1.18	0.32	0.89	0.51	**0.6**	0.34	**0.6**	**0.17**	1.34	0.64
under bust	**0.7**	**0.59**	11.07	2.7	5.4	3.32	3.09	1.39	2.68	1.6	11.65	4.09
waist1	**1.73**	**0.67**	11.7	2.93	7.04	4.09	5.41	3.02	5.29	2.8	13.28	4.75
waist2	**2**	**0.76**	12.06	2.8	7.17	4.21	5.45	2.92	5.04	2.59	13.83	4.51
waist3	**1.93**	**0.74**	13.05	3.03	7.33	4.39	5.25	2.71	4.75	2.44	14.79	4.59
waist4	**1.4**	**0.62**	13.32	3.09	7.05	4.3	4.77	2.43	4.41	2.34	14.83	4.45
waist5	**1.04**	**0.61**	12.86	2.97	6.63	4.08	4.28	2.25	3.99	2.2	14.14	4.33
waist6	**0.93**	**0.62**	12.28	2.85	6.3	3.88	3.95	2.12	3.68	2.09	13.36	4.33

Table 3. Comparison of measurement errors across different image-based methods on 4D-Dress (unseen styles).

Measurement	Ours		3DBodyNet [21]		ArtEq [16]		IP-Net [3]		IP-Net (Partial) [3]	
Unit (CM)	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
bust	3.25	**1.62**	4.76	2.02	**3.08**	3.71	4.2	3.75	3.42	2.84
hip	2.38	**1.24**	4.69	2.4	**2.32**	3.78	3.8	3.33	6.36	3.14
left calf	1.12	**0.52**	1.77	0.87	**1.02**	1.32	1.6	1.55	2.82	1.33
left elbow	**1**	**0.43**	1.02	0.52	1.15	0.82	1.25	1.25	1.6	0.92
left knee	**0.68**	**0.46**	1.73	0.93	1.01	0.97	1.64	1.5	2.44	1.06
left middle thigh	1.79	**0.92**	2.88	1.38	**1.72**	2.48	2.77	2.2	4.91	2.27
left upper arm	**1.06**	**0.64**	2.24	1.16	1.88	2.19	2.3	2.29	3.53	1.75
left wrist	0.58	**0.21**	**0.47**	0.32	0.7	0.28	1.23	0.71	1.42	0.55
right calf	1.15	**0.48**	1.38	0.74	**1.04**	1.18	1.45	1.43	2.6	1.23
right elbow	**0.73**	**0.38**	1.27	0.7	1.01	1.19	1.47	1.53	2.03	1.09
right knee	**0.72**	**0.45**	1.65	0.83	1.02	0.98	1.57	1.45	2.36	1.06
right middle thigh	**1.77**	**0.88**	2.9	1.34	1.86	2.43	2.77	2.26	5.09	2.37
right upper arm	**1.2**	**0.59**	1.93	1.07	1.63	2.15	2.19	2.12	3.1	1.57
right wrist	0.65	**0.23**	**0.49**	0.36	0.75	0.3	0.68	0.57	0.63	0.38
under bust	**2.95**	**1.5**	4.67	2.19	3.07	3.45	3.6	3.86	3.98	3.05
waist1	**1.99**	**1.4**	6.42	3.76	5.22	4.5	6.11	4.87	7.76	4.89
waist2	**2.3**	**1.44**	6.1	3.78	5.3	4.35	6.06	4.78	7.38	4.9
waist3	**2.09**	**1.47**	6.01	3.75	5.43	4.38	6.1	4.71	7.1	4.96
waist4	**2.17**	**1.42**	5.55	3.58	5.14	4.33	5.71	4.64	6.35	4.77
waist5	**2.33**	**1.4**	4.92	3.04	4.42	4.08	4.68	4.36	5.09	4.03
waist6	**2.49**	**1.43**	4.52	2.59	3.86	3.96	4	3.9	4.17	3.24

Table 4. Comparison of measurement errors across different point cloud-based methods on FAUST.

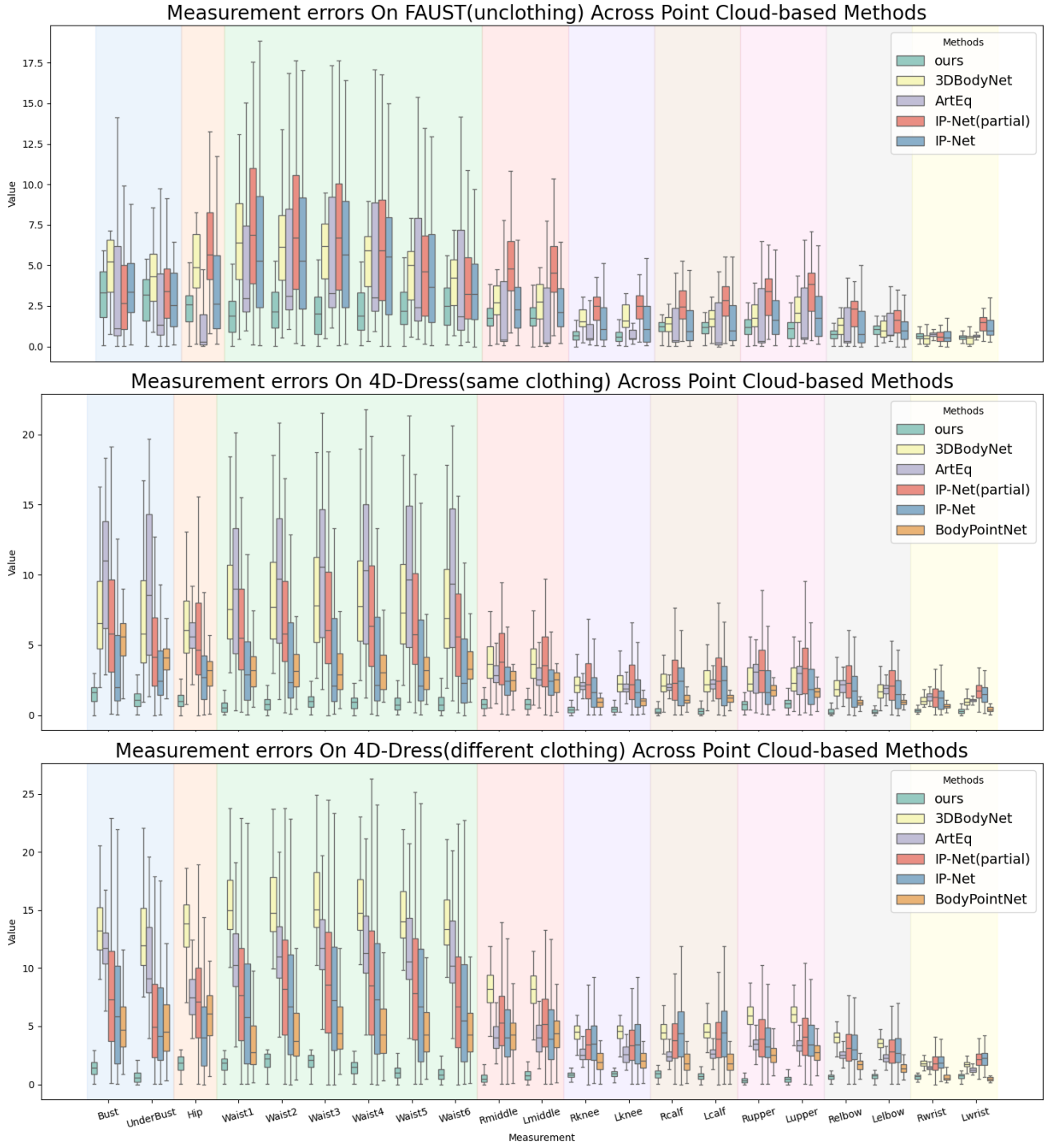

Figure 10. These figures present box plots of absolute errors (in cm) for body measurements on the FAUST and 4D-Dress datasets across different point cloud-based methods: 3DBodyNet [21], ArtEq [16], IP-Net (partial) and IP-Net [3], BodyPointNet [20], and ours.

detailed assessment.

The three metrics CD, MSE, and MAE, which we used to evaluate the performance of body shape estimation, are defined as follows:

$$CD(pred, gt) = \frac{1}{|pred|} \sum_{x \in pred} \min_{y \in gt} \|x - y\|^2 + \frac{1}{|gt|} \sum_{y \in gt} \min_{x \in pred} \|x - y\|^2 \tag{33}$$

Measurement	Ours		BodyPointNet [20]		3DBodyNet [21]		ArtEq [16]		IP-Net [3]		IP-Net (Partial) [3]	
Unit (CM)	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
bust	**1.55**	**0.76**	5.3	1.77	8.78	6.89	10.17	4.09	4.42	5.08	7.34	5.99
hip	**1.04**	**0.58**	2.93	1.3	7.51	5.12	5.67	1.65	3.83	3.92	5.8	4.33
left calf	**0.35**	**0.31**	1.19	0.43	2.82	1.94	2.21	0.55	2.28	1.66	2.8	1.89
left elbow	**0.28**	**0.22**	0.91	0.25	2.02	1.23	1.85	0.47	1.6	1.2	2.2	1.41
left knee	**0.45**	**0.3**	0.92	0.41	2.6	1.47	2.02	0.51	1.7	1.27	2.43	1.6
left middle thigh	**0.82**	**0.45**	2.31	0.97	4.34	2.69	2.77	1.02	2.76	1.97	4.03	2.49
left upper arm	**0.81**	**0.38**	1.58	0.58	3.16	2.46	2.58	1.25	2.33	2.04	3.54	2.25
left wrist	**0.28**	0.18	0.41	0.19	1.06	0.52	1.08	**0.17**	1.5	0.66	1.72	0.73
right calf	**0.32**	**0.27**	1.13	0.4	2.6	1.66	2	0.45	2.19	1.59	2.65	1.82
right elbow	**0.29**	**0.27**	0.89	0.29	2.28	1.44	2.06	0.57	1.84	1.39	2.43	1.58
right knee	**0.41**	**0.3**	0.89	0.45	2.56	1.52	2.07	0.48	1.77	1.35	2.46	1.68
right middle thigh	**0.82**	**0.47**	2.33	0.96	4.44	2.89	2.9	1.03	2.88	2.14	4.22	2.57
right upper arm	**0.72**	**0.35**	1.66	0.61	3.15	2.5	2.68	1.25	2.3	2.07	3.45	2.23
right wrist	**0.35**	**0.15**	0.65	0.21	1.22	0.69	1.3	0.3	1.21	0.84	1.29	0.87
under bust	**1.13**	**0.66**	3.98	1.54	8.37	7.45	9.37	5.17	4.45	4.81	5.84	5.79
waist1	**0.67**	**0.56**	3.2	1.56	9.95	7.83	8.91	5.04	5.27	6.36	7.62	7.07
waist2	**0.8**	**0.55**	3.25	1.62	10.08	7.96	9.61	4.96	5.32	6.56	7.97	7.29
waist3	**0.97**	**0.55**	3.19	1.73	10.17	8.07	10.26	5.01	5.41	6.8	8.39	7.55
waist4	**0.92**	**0.56**	3.21	1.77	10.08	8.02	10.19	5.4	5.41	6.83	8.54	7.68
waist5	**0.88**	**0.63**	3.28	1.8	9.65	7.7	9.91	5.48	5.1	6.43	8.15	7.47
waist6	**0.9**	**0.69**	3.55	1.75	9.22	7.38	9.75	5.35	4.56	5.68	7.25	6.89

Table 5. Comparison of measurement errors across different point cloud-based methods on 4D-Dress (same clothing).

Measurement	Ours		BodyPointNet [20]		3DBodyNet [21]		ArtEq [16]		IP-Net [3]		IP-Net (Partial) [3]	
Unit (CM)	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
bust	**1.38**	**0.71**	6.38	5.76	13.75	2.76	11.71	2.13	6.47	4.9	8.05	5.36
hip	**1.77**	**0.79**	7.07	5.55	13.71	2.3	7.55	1.88	4.47	3.25	7.43	4.35
left calf	**0.69**	**0.32**	2.32	2.03	4.8	1.09	2.65	0.51	4.45	2.38	3.93	2.13
left elbow	**0.7**	**0.26**	1.74	1.41	3.48	0.59	2.29	0.39	2.92	1.47	2.89	1.43
left knee	**0.91**	**0.29**	2.48	2.04	4.48	0.76	2.63	0.73	3.51	2.04	3.49	1.9
left middle thigh	**0.76**	**0.46**	4.89	3.12	8.17	1.57	4.02	1.34	4.38	2.74	5.47	3.02
left upper arm	**0.49**	**0.34**	2.96	1.56	5.96	0.98	3.33	0.75	3.96	2.09	4.25	2.3
left wrist	**0.69**	0.24	0.8	0.89	1.7	0.32	1.27	**0.22**	2.2	0.71	2.14	0.7
right calf	**0.87**	**0.4**	2.33	2.05	4.59	0.91	2.46	0.54	4.38	2.34	3.89	2.17
right elbow	**0.63**	**0.26**	1.95	1.32	3.97	0.63	2.54	0.42	3.25	1.59	3.23	1.62
right knee	**0.8**	**0.25**	2.4	2.08	4.49	0.7	2.61	0.57	3.61	1.91	3.58	1.93
right middle	**0.57**	**0.44**	4.74	3.16	8.29	1.6	4.09	1.14	4.52	2.72	5.62	3.06
right upper	**0.39**	**0.3**	4.66	9.11	5.93	1.02	3.42	0.73	3.79	2.06	4.14	2.3
right wrist	**0.66**	**0.18**	0.88	0.95	1.82	0.27	1.41	0.21	1.91	0.74	1.81	0.83
under bust	**0.7**	**0.59**	5.81	5.02	12.59	2.92	10.36	3.24	5.39	4.03	5.83	4.38
waist1	**1.73**	**0.67**	4.7	5.28	15.54	2.84	10.59	2.87	6.63	5.15	8.26	5.45
waist2	**2**	**0.76**	5.29	4.93	15.44	2.89	11.27	2.84	7.37	5.25	8.8	5.66
waist3	**1.93**	**0.74**	5.74	4.78	15.7	2.84	11.98	2.83	7.86	5.43	9.19	5.98
waist4	**1.4**	**0.62**	5.52	4.78	15.46	2.77	11.86	3.09	7.95	5.67	9.19	6.13
waist5	**1.04**	**0.61**	5.49	4.7	14.64	2.65	11.41	3.2	7.55	5.63	8.73	6.03
waist6	**0.93**	**0.62**	5.48	4.74	13.97	2.56	11.11	3.15	6.59	5.15	7.68	5.61

Table 6. Comparison of measurement errors across different point cloud-based methods on 4D-Dress (unseen styles).

$$MSE(pred, gt) = \frac{1}{|pred|} \sum_{x \in pred, y \in gt} (x - y)^2 \quad (34)$$

$$MAE(pred, gt) = \frac{1}{|pred|} \sum_{x \in pred, y \in gt} |x - y| \quad (35)$$

where $(pred, gt) \in \{(\mathcal{T}, \mathcal{T}_{gt}), (\widetilde{\mathcal{S}}^f, \widetilde{\mathcal{S}}^f_{gt}), (\widetilde{\mathcal{S}}^b, \widetilde{\mathcal{S}}^b_{gt})\}$.

8.2. Comparisons with different methods on real-world datasets

Due to space limitations, the main text presents comparisons only with the latest (2023/2024) methods and the most relevant approach, 3DBodyNet, which remains the most recent method using partial front- and back-view scans for body shape estimation, making it particularly relevant to our study. However, for a more comprehensive evaluation, our actual experiments include comparisons with relevant works from the past five years.

In the supplementary materials, we expand the comparison by categorizing methods into:

- Comparison with image-based methods: POCO (2024)[15], NLF (2024) [53], PyMAF-X (2023) [65], BEDLAM (2023) [5] and PyMAF (2021) [64];
- Comparison with point cloud-based methods: ArtEq (2023) [16], 3DBodyNet (2021) [21], IP-Net (partial) (2020) taking partial point cloud as input and IP-Net (2020) taking complete point cloud as input [3], and Body PointNet (2020) [20].

We first outline a method for conducting fairer comparisons across different approaches, accounting for variations in input data. As MeasureXpert utilizes two inputs containing body information from both front and back views, additional adjustments are necessary to ensure equitable comparison. For image-based methods, we used two RGB images of one individual from the front- and back-views, which can obtain two pairs of pose and shape parameters (θ_1, β_1, θ_2, β_2). We extracted the unclothed and T-posed body mesh of this individual from the mean shape parameters $\frac{\beta_1+\beta_2}{2}$ and measured the body mesh to obtain the anthropometric values. For ArtEq, IP-Net, and Body PointNet, which take complete point clouds as input, we predict two posed SMPL models from the front and back complete scans of an individual. For ArtEq and IP-Net, which output SMPL parameters, we follow a process similar to that of image-based methods, using the mean shape vector to generate the T-posed body as the final result. For Body PointNet, which directly regresses SMPL vertices instead of parameters, the predicted SMPL models lack explicit shape parameters, preventing the same processing as image-based methods. Instead, we first apply PoseNormNet to normalize the two models into T-posed SMPL meshes. These posture-normalized models are then registered and averaged vertex-to-vertex to produce a mean T-posed model, from which the final measurements are extracted.

Since all methods output SMPL bodies, we extract measurement values from the final results as described in Section 7.1 and compare them with the ground-truth values. To ensure a comprehensive evaluation, we use box plots to illustrate measurement error distributions (Figure. 9 and Figure. 10) and provide mean and standard deviation tables (Table 1-Tabel 6) for a clearer quantitative comparison of measurement accuracy across different methods. The red and bold values in each row highlight the lowest mean and standard deviation. These additional experiments further validate the effectiveness and advantages of our approach.

8.3. By-products

Figure 11 and Figure 12 report the performance on pose estimation based on FAUST and 4D-Dress dataset, respectively.

8.4. Evaluation on challenging scans from low-cost devices

In previous experiments, we relied on front- and back-view partial point clouds extracted with RealPartialScan from the public FAUST and 4D-Dress datasets. These datasets were acquired in multi-camera studios equipped with professional active-stereo or structured-light systems. The resulting meshes are high-resolution, have been carefully denoised and topologically cleaned. Consequently, the partial point clouds produced by RealPartialScan are relatively uniform, contain few holes, and have low noise.

By contrast, we captured a more challenging real-world dataset with two consumer devices: Orbbec Astra 2 (structured-light, 2.5–$3m$ range) and CR-Scan Otter (handheld, around $1m$ range). Astra 2 delivers a single depth frame per view at a 2.5–$3m$ range; the data suffers from artefacts, missing regions, and noise. CR-Scan Otter performs a one-to-two-minute handheld sweep. The working distance for the CR-Scan Otter is within 1 meter, which needs to be scanned by moving the scanner from head to feet slowly to capture many small point patches and register all the patches with global optimization to form a point cloud. During this scanning process, unavoidable subject micro-motion causes non-rigid mis-registration, accuracy fluctuations, and seam gaps (Check Table 7 for more information).

The dataset consists of 42 real-world front and back scan pairs from six volunteers: 24 pairs with an Orbbec Astra 2 and 18 pairs with a CR-Scan Otter. The resulting dataset (Fig. 13) comprises

- 26 pairs in unconstrained everyday poses;
- 8 pairs in which the same forearm is occluded in both views;
- 8 pairs with deliberate partial occlusion of the waist.

The results in Table 8 were obtained without any complex post-processing: each scan was fed directly into the network in its raw form, preserving depth-dependent noise (Astra 2), patch-fusion drift (Otter), holes, and other artifacts. Apart from a simple threshold-based filter to remove static back-

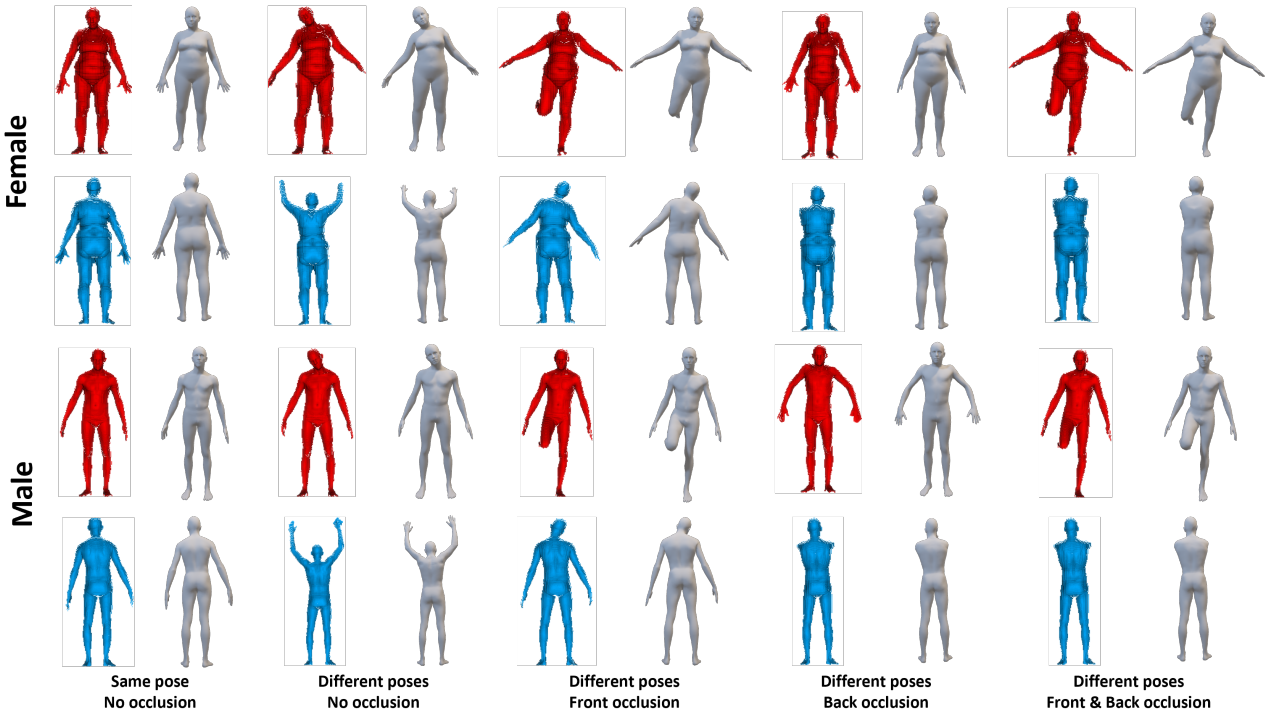

Figure 11. Pose estimation results of TrioNet based on FAUST dataset: the red point clouds are captured from the front-view of bodies in the FAUST dataset, while the blue point clouds are the back-view captures. The meshes following the point clouds are the posed bodies TrioNet predicted.

Table 7. Data characteristics and key challenges associated with two low-cost scanning devices.

Device	**Capture protocol**	**Key challenges**
Orbbec Astra 2 (structured-light, fixed setup) Price: $353	• Single-view depth capture at $\sim$ 2.5–3 m.	• Increased noise at longer capture distances. • Artefacts and missing regions (holes) in the scanned point cloud. • Insufficient accuracy and completeness of acquired geometry.
CR-Scan OTTER (hand-held, real-time registration) Price: $899	• 1–2 min hand-held scanning at around 1 m. • Continuous alignment of sequential point patches during device movement.	• Slight subject motion leads to non-rigid misalignment. • Global fusion may introduce local drift and visible seam gaps. • Clothing wrinkles can cause missing data (holes). • Varying scanning distances during hand-held movement result in uneven point-cloud accuracy.

ground points, no denoising, registration, or hole filling was applied. Despite these low-cost scanning conditions and the absence of professional data refinement, MeasureXpert is still able to produce reasonably reliable anthropometric measurements using unprocessed, directly captured data. Furthermore, as mentioned before, we designed two

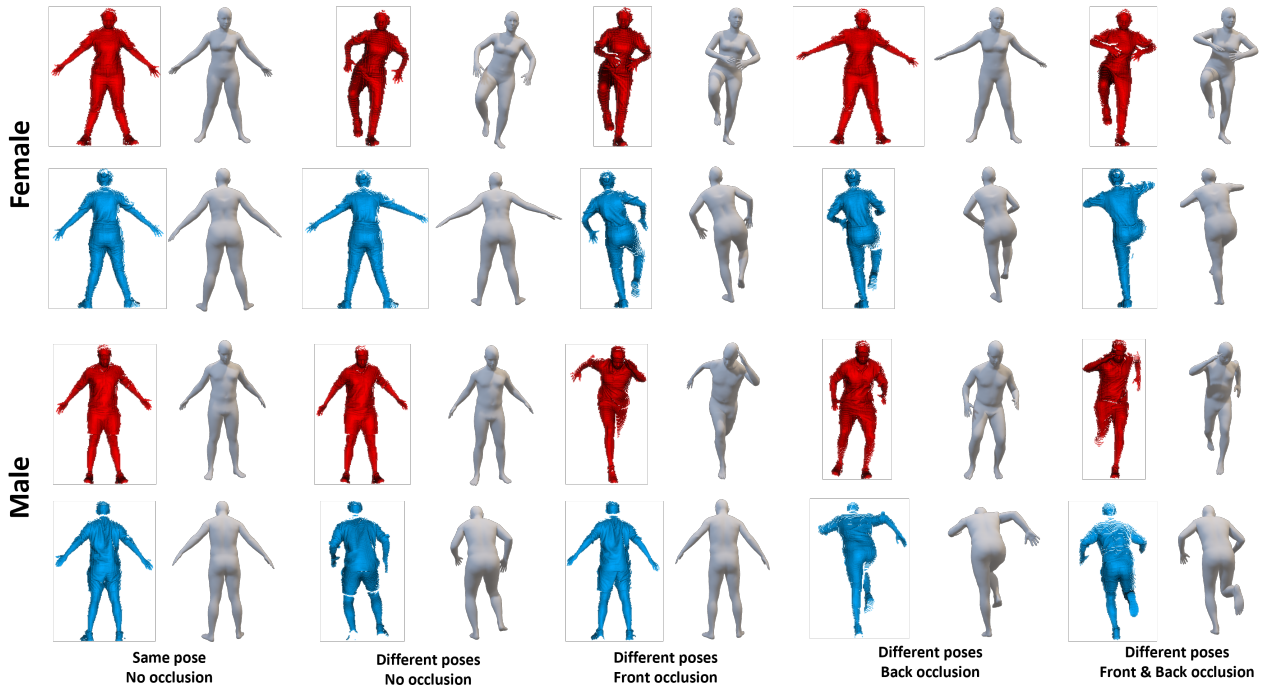

Figure 12. Pose estimation results of TrioNet based on 4D-Dress dataset: the red point clouds are captured from the front-view of bodies in the 4D-Dress dataset, while the blue point clouds are the back-view captures. The meshes following the point clouds are the posed bodies TrioNet predicted.

Table 8. Measurement errors across devices (unit: cm)

	bust	underbust	hip	waist	thigh	knee	calf	arm	elbow	wrist	relbow	lelbow	rwrist	lwrist
Astra2	2.15	2.58	4.17	4.14	3.49	0.99	1.76	1.72	1.83	1.06	1.64	2.12	1.28	0.84
Otter	2.58	2.25	3.60	2.20	1.72	0.69	0.96	1.01	1.97	1.17	1.74	2.21	1.45	0.88

Figure 13. The upper eight point clouds were captured with the Orbbec Astra 2, and the lower eight with the CR-Scan Otter 3D Scanner. The red rectangles indicate designed postures.

occlusion-specific poses in our newly collected dataset, where the same region (e.g., forearm or waist) is occluded in both views. We compared performance under occlusion and no-occlusion conditions. For the waist and single wrist, the average errors under occlusion were $4.69cm$ and $1.28cm$, respectively, while the errors without occlusion were $3.64cm$ and $1.12cm$, respectively.

8.5. Ablation studies

In this section, we will present more experimental evidence through ablation studies.

8.5.1. Input design

Initially, we discuss whether a single partial point cloud as input is enough for pose and shape prediction. We use the architecture of PoseNormNet as our baseline and input a partial dressed body scan. We refer to the PoseNormNet configuration with the front partial point cloud as Base-F and the one with the back partial point cloud as Base-B. We compare the performance of Base-F, Base-B, and TrioNet to discuss i) which partial point cloud is more suitable for pose and shape estimation, and ii) if only one partial point cloud is enough for shape estimation.

Table 9. Loss function design

	TrioNet	TrioNet-C	TrioNet-D	TrioNet-E	TrioNet-F
Shape	✓	×	✓	✓	✓
Body	L2	L2	L2	L2	L2
Landmarks	CD	CD	L2	L2+$\mathcal{L}_{Norm}$	CD+$\mathcal{L}_{Norm}$

8.5.2. TrioNet loss functions design

Regarding the decoder architecture, we first discuss the two input branches. We designed TrioNet-A and TrioNet-B, using three fully-connected layers to decode $\mathcal{F}^i$ directly to $\widetilde{\mathcal{S}}^i$ with L2 loss and CD, respectively, to discuss which loss function is better for posed body point cloud reconstruction. **T**rioNet with **S**ingle **D**ecoder in two input branches (TSD) in Section 4.3 is the better one. We chose the better loss function to design the multi-decoder TrioNet and compared it with TSD to discuss if a multi-decoder is necessary in Section 4.3.

We then discuss the loss function of the shape branch. The loss function should help supervise predicted T-posed body and landmarks. Here, we do not discuss the choice of body loss because we want to predict ordered points as vertices of the SMPL surface, which can maintain the topology of the SMPL model. Therefore, L2 is the best choice to supervise bodies, and no need to discuss. We concentrate solely on the loss of landmarks prediction. We first discuss CD or L2 loss, which is the best loss function for landmarks prediction. In addition, when we focus on landmarks, we draw inspiration from A-Net, a successful shape prediction and measurement neural network. A-Net introduced conducted experiments to prove that one-level landmark points should be constrained on one plane, which is important for measurement. We take the landmark constrain loss $\mathcal{L}_{Norm}$ (Eq. 36) into consideration, which is similar to A-Net [29] to constrain landmark points on the same planar:

$$\mathcal{L}_{Norm} = \sum_{i=1}^{21} \frac{1}{|l_i'|} \sum_{j=1}^{|l_i'|} \sum_{k=1}^{|l_i'|} |l_{ij}' - l_{ik}'| \tag{36}$$

Here, l_i' is one level of the landmark point set. There are two different types of l_i': the plane of l_i' is parallel to the ground or vertical. If the plane is parallel to the ground, l_{ij}' and l_{ik}' represent $y-$axis values of points. Otherwise, l_{ij}' and l_{ik}' represent $x-$axis values of points.

Based on these functions, we define the variants TrioNet-D through TrioNet-F. Additionally, we reintroduce TNS, as mentioned in Section 4.3, under the name TrioNet-C to facilitate a more detailed comparison. The loss functions for TrioNet, TrioNet-C, TrioNet-D, TrioNet-E, and TrioNet-F are presented in Table 9.

8.5.3. OR-Net design

In Section 4.3, we demonstrated that OR-Net significantly improves prediction performance on both T-posed meshes and measurement values. In this section, we aim to further investigate the network architecture and its associated loss functions to achieve more accurate measurement values.

We compare the following designs: i) We split offset learning and regression into two tasks, where regression directly operates on the landmarks predicted by TrioNet to obtain measurement values. We refer to this structure as Regression-only for value prediction; ii) We adopt the OR-Net structure but focus on the design of the loss functions. Since the landmarks predicted by TrioNet are not strictly constrained to the ground truth measurement-level planes by the current loss function, and we need to predict values based on these landmarks, we revisit the conclusion drawn from A-Net: ensuring that all landmark points at the same level are constrained to a single plane is critical for accurate measurements. We design the OR-Net with the loss function as:

$$\mathcal{L}_{op} = \mathcal{L}_{value} + \mathcal{L}_{lmk} + \mathcal{L}_{tpose} + \gamma \mathcal{L}_{Norm} \tag{37}$$

and conduct four experiments with $\gamma = 0$, $\gamma = 1/3$, $\gamma = 2/3$, and $\gamma = 1$, which we denote as 0-Norm experiment (OR-Net), $1/3$-Norm experiment, $2/3$-Norm experiment, and 1-Norm experiment, respectively.

8.5.4. Results and discussions

To evaluate the performance of TrioNet, Base-F, Base-B, and TrioNet variants A to F, we used the Chamfer distance (CD), mean squared error (MSE), and mean absolute error (MAE) to calculate vertex-to-vertex errors between predicted and ground-truth meshes. The results in 0.1mm unit are presented in Fig. 14.

We first compare among Base-F Base-B and TrioNet. The performance of shape prediction is comparable for single partial inputs; however, for posed body prediction, the front view outperforms the back view. We attribute this to the fact that the back view typically experiences greater occlusion, leading to the loss of hand or forearm posture information. However, TrioNet, taking both front and back view as input, significantly improves performance in both pose estimation and shape estimation. This indicates that one partial point cloud is insufficient for accurate body estimation, whereas two partial point clouds provide adequate accuracy for this task.

When comparing TrioNet-A with TrioNet-B, the L2 loss shows better performance in estimating the front and back posed bodies. Furthermore, comparing TrioNet with TrioNet-B reveals that the multi-decoder architecture outperforms the single-decoder in posed body estimation. This improved pose estimation leads to more accurate shape estimation. Therefore, for the pose branches, the combination of a multi-decoder architecture with L2 loss and boundary constraints proves to be the most effective design.

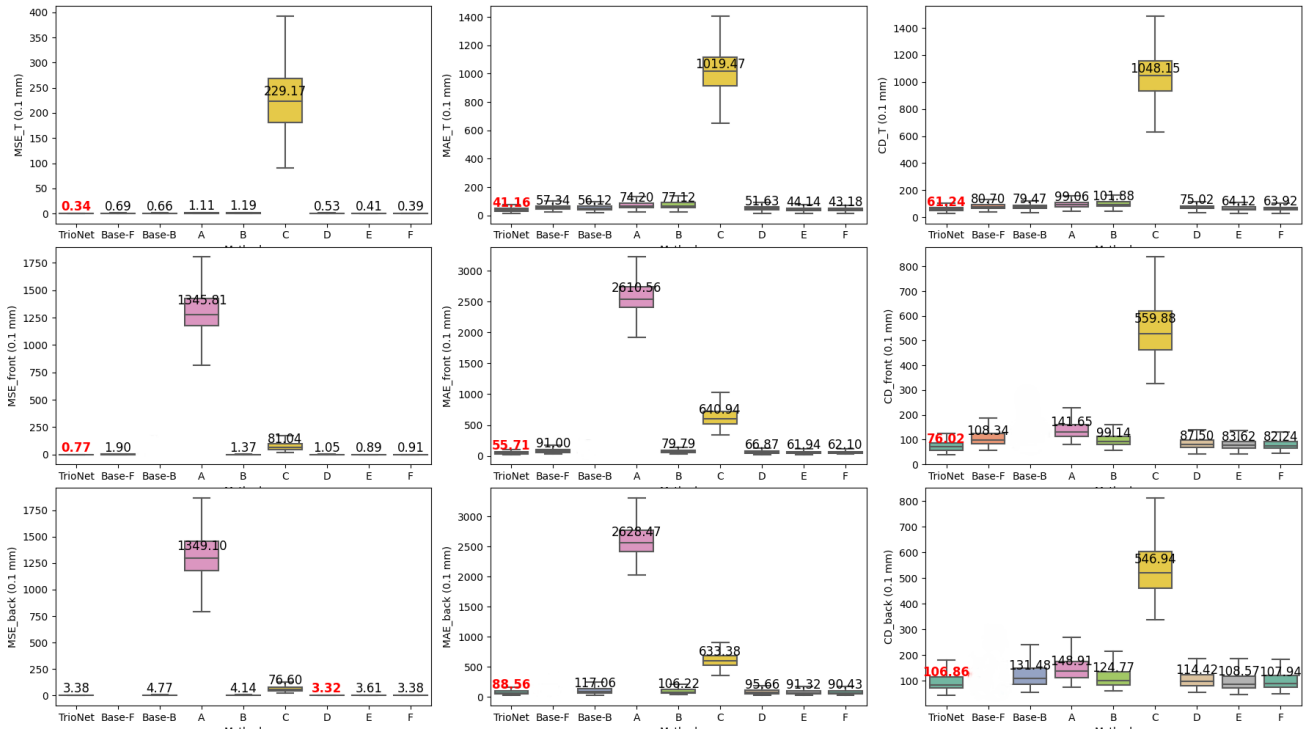

Figure 14. The figure presents a comparative analysis of different model configurations across three error metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Chamfer distance(CD) with 0.1mm as their units. Each box plot illustrates the error distribution for configurations labeled TrioNet, Base-F, Base-B, and A through F, evaluated under T-posed body prediction, front-posed body prediction and back-posed body prediction conditions.

When focusing on the loss design of shape estimation, we observed that TrioNet-C performed the worst across all metrics, highlighting the necessity and validity of the $\mathcal{L}_{shape}$ design. Comparing TrioNet-D and TrioNet-E, we found that using L2 loss to learn landmarks, combined with normalization constraints, significantly improved the overall prediction accuracy. Under the constraint of normalization, the performance difference between using L2 and CD as loss functions was minimal. Interestingly, when CD was used alone for landmark learning without normalization constraints, the results for both pose estimation and shape estimation were better. We think that $\mathcal{L}_{Norm}$ distracts from learning shape and pose. Therefore, we designed TrioNet with shape loss, L2 loss as body loss, and CD as landmark loss without $\mathcal{L}_{Norm}$.

Table 10 presents the means and standard deviations (in centimeters) for 21 different measurement values across the 0-Norm, $1/3$-Norm, $2/3$-Norm, 1-Norm, and Regression-only experiments. Overall, the 0-Norm experiment (OR-Net) demonstrates the best performance among all designs, suggesting that whether the landmarks are strictly constrained to a single plane has little impact on the regression outcomes. When the landmarks are not rigidly confined to a single plane, and instead the focus is shifted toward optimizing both values and point positions, the results are comparatively better. Furthermore, a comparison between OR-Net and Regression-only indicates that splitting optimization and regression into two separate tasks is unnecessary. Therefore, OR-Net is the best design for our task.

Measurement Levels	0-Norm		1/3-Norm		2/3-Norm		1-Norm		Regression-only	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
bust	**1.41**	**1.14**	1.45	**1.14**	1.41	1.16	1.43	1.14	1.41	1.14
under bust	**1.50**	**1.24**	1.57	1.31	1.53	1.28	1.53	1.27	**1.50**	1.24
hip	**1.23**	**1.05**	1.24	1.13	1.28	1.10	1.26	1.11	**1.23**	1.09
waist1	**1.79**	**1.55**	1.86	1.67	1.83	1.59	**1.79**	1.62	1.83	1.58
waist2	**1.83**	**1.56**	1.89	1.68	1.87	1.63	1.84	1.64	1.84	1.59
waist3	**1.87**	**1.61**	1.94	1.68	1.90	1.66	1.91	1.66	1.89	1.62
waist4	**1.88**	1.64	1.96	1.69	1.89	1.67	1.91	1.65	1.89	**1.63**
waist5	**1.82**	1.57	1.89	1.64	1.84	1.60	1.87	1.62	1.84	**1.56**
waist6	**1.73**	1.47	1.82	1.51	1.74	1.49	1.75	1.50	1.73	**1.45**
right middle thigh	**0.95**	**0.85**	0.98	0.88	0.99	0.86	1.00	0.89	0.97	0.86
left middle thigh	**0.97**	**0.87**	1.00	0.90	1.00	0.89	1.03	0.91	1.00	0.88
right knee	0.41	**0.34**	0.41	0.35	**0.40**	0.37	0.43	0.36	0.41	0.36
left knee	0.39	**0.33**	0.40	0.35	**0.38**	0.35	0.42	0.35	0.39	0.35
right calf	**0.47**	**0.42**	0.47	0.43	0.48	0.43	0.52	0.44	0.47	0.42
left calf	**0.51**	**0.46**	0.52	0.47	0.53	0.48	0.53	0.49	0.52	0.46
right upper arm	**0.62**	**0.51**	0.63	0.54	0.64	0.53	0.63	0.52	0.63	0.51
left upper arm	**0.68**	**0.56**	0.71	0.61	0.70	0.56	0.70	0.58	0.70	0.57
right elbow	**0.33**	0.30	0.33	0.29	0.33	**0.28**	0.33	0.28	0.33	0.28
left elbow	**0.32**	0.29	0.33	0.31	0.33	0.30	0.33	0.29	0.33	**0.28**
right wrist	**0.21**	**0.17**	0.22	0.18	0.21	0.17	0.22	0.17	0.21	0.17
left wrist	**0.17**	0.15	0.17	**0.14**	0.17	0.14	0.17	0.14	0.17	0.14

Table 10. Comparison of measurements across different OR-Net designs (Unit: cm).