Human shape reconstruction with loose clothes from partially observed data by pose specific deformation

Akihiko Sayo[1], Hayato Onizuka[1], Diego Thomas[1], Yuta Nakashima[2], Hiroshi Kawasaki[1], and Katsushi Ikeuchi[3]

[1] Kyushu University, Japan
{a.sayo.004,h.onizuka.392}@s.kyushu-u.ac.jp,
{thomas,kawasaki}@ait.kyushu-u.ac.jp
[2] Osaka University, Japan n-yuta@ids.osaka-u.ac.jp
[3] Microsoft Corp, USA katsuike@microsoft.com

Abstract. Reconstructing the entire body of moving human in a computer is important for various applications, such as tele-presence, virtual try-on, etc. For the purpose, realistic representation of loose clothes or non-rigid body deformation is a challenging and important task. Recent approaches for full-body reconstruction use a statistical shape model, which is built upon accurate full-body scans of people in skin-tight clothes. Such a model can be fitted to a point cloud of a person wearing loose clothes, however, it cannot represent the detailed shape of loose clothes, such as wrinkles and/or folds. In this paper, we propose a method that reconstructs 3D model of full-body human with loose clothes by reproducing the deformations as displacements from the skin-tight body mesh. We take advantage of a statistical shape model as base shape of full-body human mesh, and then, obtain displacements from the base mesh by non-rigid registration. To efficiently represent such displacements, we use lower dimensional embeddings of the deformations. This enables us to regress the coefficients corresponding to the small number of bases. We also propose a method to reconstruct shape only from a single 3D scanner, which is realized by shape fitting to only visible meshes as well as intra-frame shape interpolation. Our experiments with both unknown scene and partial body scans confirm the reconstruction ability of our proposed method.

Keywords: Non-rigid registration · eigen-deformation · neural network · human shape reconstruction.

1 Introduction

A full-body human shape reconstruction with realistic clothes deformations is important for various applications, such as tele-presence, virtual try-on, etc. in order to achieve immersive feelings and realistic sensations. Unlike skin-tight clothes, loose clothes can be deformed dynamically, and thus, it is an open problem to capture and represent them. There are two main approaches to solve this problem: non-prior and prior based methods. Non-prior methods can reconstruct any shape and the resolution of the reconstructed model varies depending on the resolution of the camera [6,14,21,25,26]. Some recent approaches [6,14] fused non-rigid deforming surfaces into a volumetric model by using a single-RGBD camera and could achieve real-time reconstruction. These approaches can reconstruct any shape as they are not specified to a human body. Therefore, they cannot also describe deformations of clothes efficiently that are caused by human motion and pose. Although recent approaches can reconstruct humans who wear any type of clothes [25,26], they may generate holes, wrong connections in the meshes, unwanted shapes at unobserved regions.

On the other hand, prior based methods can reconstruct full-body human with a consistent mesh without any hole. Recently, statistical shape models [1, 2,4,11] are often used as full-body human priors. Statistical shape models can represent pose-dependent deformations, such as bending arms or deformed muscles, with only a few shapes and pose parameters. There are some methods that can reconstruct full-body 3D human shape only from a single image using a statistical shape model [3,8,15,17,20,23]. One severe drawback of these methods is that the model can only be used with people wearing skin-tight clothes; however, in real situation, people rarely wear such clothes.

Although statistical shape models have several limitations as mentioned above, their ability on shape representation by shape and pose parameters is still promising. In this paper, we propose a non-rigid full-body human shape reconstruction from point clouds measurements, even when the person is wearing loose clothes and captured from a single viewpoint. To this end, we use a statistical shape model as a base for full-body human modeling and regress the displacements between the base mesh and the clothed mesh by training a neural network. This allows us to take advantage of the efficient representation of the model and represent the clothed human mesh in lower-dimension. Similar to our work, Yang *et al.* [24] regresses coefficients corresponding to eigenvectors which are obtained through PCA of the displacements and reconstruct full-body human with the same resolution as the target mesh. However, they assume that the target mesh is consistent through the process, whereas such limitation does not exist in our method. Kimura *et al.* [9] also regress the coefficients corresponding to eigenvectors obtained by PCA of target displacements. In contrast, we not only regress the coefficients but also train eigenvectors as weight of the last layer of a neural network which is used as a regressor. The main contributions of this paper are summarized as follows:

1. Eigenvectors and coefficients are obtained by trianing a neural network to achieve accurate full-body reconstruction with using a lower dimensional space compared with previous method.
2. Direct regression of displacement values by neural network is proposed to enable reconstruction of full-body human shape even if parts of the body were not observed for training data.

2 Related Work

The 3D reconstruction problem is extremely challenging for the case of non-rigid objects/scenes because it requires a non-rigid 3D registration.

In the case of human body, deeper prior knowledge has been used. For example, Malleson *et al.* [12] proposed a method that assumes that each body part is rigid. A human body is then described as a combination of these body parts under this assumption, and a voxel carving-based approach is used for reconstructing each body part with texture. Due to the rigidity assumption, however, this method sometimes suffers from discontinuity at body part joints.

Another interesting approach is to use a statistical human body shape model, such as SCAPE [1] or SMPL [11]. A statistical shape model for the human body is usually trained with many full-body measurements of different body shapes and poses, and these variations are parameterized in a low dimensional space. Therefore, aligning the human body model to a depth measurement is much easier. Bogo *et al.* presented a high-quality method for full-body reconstruction upon a statisticcal shape model [2] from a single RGB-D camera. Recently, several methods perform full-body reconstruction from a single RGB image [3,8,17,20].

One major criticism on such methods is their insensitivity to deformations that are not described by the statistical shape model. In most cases, the full-body measurements used for training the model are in skin-tight clothes. Thus, such methods are usually tested with skin-tight clothes.

Recently, several methods have used statistical shape models to reconstruct not only the human body but also the clothes [5, 7, 9, 18, 24, 26]. Pons-moll *et al.* [18] proposed to use 4D scan data of human with clothes as input and fit a model. By segmenting the body part and clothes part, they can retarget the clothes to a novel model. Yu *et al.* [26] used SMPL model as a semantic and real non-rigid human motion prior. They pre-defined an on-body node graph on the SMPL model and a method ran similar to DynamicFusion [14], which means that the outer surface is parameterized by such nodes and the deformations of the surface can be represented by a rigid transformation of the nodes. They also define the points far from the SMPL body as far-body nodes. This enables their method to be robust against noise. Nevertheless, their results are still noisy and cannot describe pose-dependent clothes deformations. Joo *et al.* [7] presented the Adam model in their work. Adam model can represent not only the body pose and shape but also facial expressions, finger motion, head and clothes. The reconstruction results are impressive but their model is too smooth to represent clothes deformations and thus cannot represent the deformations depending on the human pose. Habermann *et al.* [5] proposed a model-based full-body reconstruction method via a low cost single RGB camera. They built subject and cloth specific models in a pre-process and reconstruct full-body 3D model using pose optimization and non-rigid deformation. Their results are attractive but due to the final non-rigid deformation method, clothes deformations of their model are not realistic.

The methods proposed by Kimura *et al.* [9] and Yang *et al.* [24] are similar to our work. Both of them represent clothes deformations as displacements from the

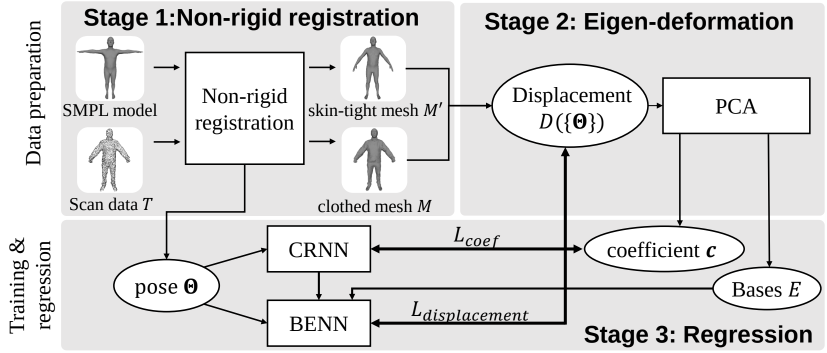

Fig. 1. Overview of our system. We first calculate the displacements between human shape with clothes and a template model. Then, these displacements are represented by small number of bases and coefficients. The displacements and coefficients are estimated by regressing the human pose parameters.

skin-tight model, which is reconstructed from shape and pose parameter of statistical shape model. They both represent such deformations in lower-dimensional space and deal with the deformations of the clothes caused by the pose of the person wearing it. Based on this assumption they regress the coefficients corresponding to the eigenvector of lower-dimensional space. Although Yang *et al.* [24] assume that the input data has topologically consistent mesh, Kimura *et al.* [9] does not. This means that their method can be applied to any point cloud data. However, Kimura *et al.* [9] only estimate the coefficients, that is, they cannot estimate if the input data is partially observed.

In our proposed method, deformations that cannot be described by a statistical shape model are obtained by fitting a 3D mesh to a depth measurement. The deformations in each triangle in the mesh are then embedded in a space parameterized by body part rotations, which we call eigen-deformations, assuming that the deformations are dependent only on them. For the back side of the measurements, we regress the deformations based on the eigen-deformation. Note that our registration process is modular: we introduce our own implementation, but Bogo *et al.* [2], for example, can be used instead.

3 Method

3.1 Overview

The deformations of clothes are in general due to the specific pose of the person wearing them. Therefore, our proposed method regresses displacements between loose and skin-tight clothes from the pose. Fig. 1 shows an overview of our proposed method. In the data preparation, which consists of non-rigid registration (Sec. 3.2) and eigen-deformation (Sec. 3.3), we use a sequence of point cloud data T as input, the input sequence can contain data about either the full body or only a part of it and obtain full-body human meshes with skin-tight and loose clothes (M' and M) as output. In the regression stage(Sec. 3.4), we input a human pose and regress displacements D. We describe how to train a neural network to regress the displacements with partially observed data in Sec. 3.5.

We use a statistical shape model to represent the full-body human mesh and deform it to represent loose clothes. Specifically, we use the SMPL model [11] as our base model, which is parameterized with the body shape parameter $\boldsymbol{\beta}$ and body pose parameter $\boldsymbol{\theta}$.

First, we fit the SMPL model to the scanned point cloud data frame by frame. This is done step by step. Through this stage, we obtain skin-tight mesh M', which has the optimized joint angles $\boldsymbol{\theta}$ and body shape $\boldsymbol{\beta}$ of the SMPL model, and clothed mesh M, which represents clothes deformations. Second, we calculate displacements D between M and M' and perform PCA on it to obtain bases of the displacements and the coefficients corresponding to the bases. Third, we train two types of neural network (NN): a coefficient regression neural network (called CRNN) and a bases estimation neural network (called BENN). CRNN is a network that regresses the coefficients corresponding to the bases obtained in the previous stage. BENN is a network that has the bases of the displacements as a part of weight parameters and directly regresses displacements D($\{\boldsymbol{\Theta}\}$). When the training data is partially occluded, we use a visibility map to ignore the occluded parts.

To summarize these notations, the mesh $\bar{M}$ obtained by regression is represented as:

$$\bar{M} = M'(\boldsymbol{\beta}, \boldsymbol{\theta}) + D(\{\boldsymbol{\Theta}\}), \tag{1}$$

Our goal is to regress unobserved pose-dependent clothes deformations from the pose of the human body.

3.2 Non-rigid registration of the SMPL model

First, Our system fits the SMPL model [11] to the input sequence of point clouds, each of which is a human body scan. This facilitates compression (and efficient representation) of human bodies with loose clothes, because the SMPL model gives us an efficient parametrization of the human body by body shape parameters $\boldsymbol{\beta}$ and joint angles $\boldsymbol{\theta}$.

First, our algorithm fits the SMPL model to the target point cloud T in the input sequence in the coarsest level, which means optimizing only the SMPL pose and shape parameters. The following energy function is minimized over body shape parameters $\boldsymbol{\beta}$ and joint angles $\boldsymbol{\theta}$.

$$\begin{aligned} E(\boldsymbol{\beta}, \boldsymbol{\theta}) = & \sum_{i \in C_{\text{anc}}} \|x'_i(\boldsymbol{\beta}, \boldsymbol{\theta}) - y_i\|^2 \\ & + P(\boldsymbol{\theta}) + R(\boldsymbol{\theta}) + S(\boldsymbol{\beta}) + Q(\boldsymbol{\beta}, \boldsymbol{\theta}) \end{aligned} \tag{2}$$

where $x'_i(\boldsymbol{\beta}, \boldsymbol{\theta})$ and y_i are the i-th vertex in $M'(\boldsymbol{\beta}, \boldsymbol{\Theta})$ and its corresponding point in T. P is a pose prior built as logarithm of as sum of mixture of Gaussians, which keeps $\boldsymbol{\theta}$ in probable pose space. R is the penalties for unusual joint angles, S for large $\boldsymbol{\beta}$ values that correspond to inplausible shapes, Q for inter-penetration of body parts. More details on these penalty terms can be found in [3].

Second, the skin-tight 3D mesh M' is inflated to roughly minimize the distance between the mesh of M' and the target cloud points T according to their silhouette points. The silhouette points are identified in 2D reprojected images. When using a single camera, we project both M' and T to the image plane of the camera to find the silhouette points. This step introduces more flexibility in vertex positions so that M can be closer to T, which makes the finest step's fitting more stable.

In this step, we restrict the possible displacement of the i-th vertex in M to

$$x_i(\boldsymbol{\beta}, \boldsymbol{\theta}, \{n, l\}) = x'_i(\boldsymbol{\beta}, \boldsymbol{\theta}) + n_i l_i, \tag{3}$$

for all i, where $x_i \in M$, $x'_i \in M'$, n_i is the direction of the movement on M' at x', and l_i is the amplitude of the displacement. Here, Kimura *et al.* [9] allowed each vertex to move only in their normal direction, while we allow for more freedom in the displacement direction. This new formulation prevents each mesh to interpenetrate (*e.g.* around the crotch when a person stand). The energy function to be minimized according to the silhouette correspondences C_{sil} during the fitting process is:

$$\begin{aligned} E(\{n, l\}) = & \sum_{i,j \in C_{\text{sil}}} \|x_i - y_j\|^2 + \lambda_l \sum_{i,j \in A_{ver}} (l_i - l_j)^2 \\ & + \lambda_{normal} \sum_{i,j \in A_{face}} |N_i^{face} - N_j^{face}| \\ & + \lambda_{rot} \sum |n_i - n'_i| + \lambda_{laplacian} \sum |x_i - \frac{1}{z_i} \sum_{j \in \mathscr{N}_i} \boldsymbol{x}_j|, \end{aligned} \tag{4}$$

where A_{ver} and A_{face} are adjacency of vertices and faces of the SMPL model respectively. N_i^{face} is the normal of the i-th face. n'_i is the normal vectors of the i-th vertex of M'. $\mathscr{N}_i$ is the set of adjacent indices of the i-th vertex and z_i is the number of the adjacent vertices around the i-th vertex. λ_l, λ_{normal}, λ_{rot}, and $\lambda_{laplacian}$ are weights to control the contributions of the associated terms. The second term is a regularizer to keep the displacement of the adjacent vertices similar. The third term regularizes the direction of the adjacent mesh normals to avoid too flat surfaces. The fourth term keeps the vertex direction of the movement as rigid as possible. The last term is the Laplacian mesh regularizer inspired from [13, 19], which enforces smoothness.

At the finest step, we use all vertices and points instead of the silhouettes to make correspondences. Also, the displacements are not restricted by (Eq. (3)), *i.e.*,

$$x_i(\boldsymbol{\beta}, \boldsymbol{\theta}, \{d\}) = x'_i(\boldsymbol{\beta}, \boldsymbol{\theta}) + d_i. \tag{5}$$

With these modifications, the energy function to be minimized is:

$$\begin{aligned} E(\{d\}) = & \sum_{i \in C_{\text{all}}} \|x_i - y_i\|^2 + \lambda_{normal} \sum_{i,j \in A_{face}} |N_i^{face} - N_j^{face}| \\ & + \lambda_{laplacian} \sum |x_i - \frac{1}{d_i} \sum_{j \in \mathscr{N}_i} \boldsymbol{x}_j|, \end{aligned} \tag{6}$$

where C_{all} are the correspondences between M and T. In this step, the nearest neighbors from T to M are updated in an iterative manner. At the final iteration, if the target point cloud T is a full-body human data, we identify correspondences not only from T to M but also from M to T. This allows us to fill gaps, which sometimes occur around the armpit and the crotch. Owing to the Laplacian regularizer term, vertices in the SMPL model which do not have any correspondence also move by being dragged by the other vertices.

After this non-rigid registration step, we obtain the body shape parameters $\boldsymbol{\beta}$, the joint angles $\boldsymbol{\theta}$, and a fully-registered 3D mesh M for each point cloud in the input sequence. The hands and feet in M may be collapsed because hands' configuration can be different (the SMPL model assumes that the hands are open) and the scanned person may wear shoes (the SMPL model assumes no shoes). We consider that the appearance of the resulting 3D mesh is important, so we do not reconstruct the hands and feet through this non-rigid registration.

3.3 Sparse representation of human shape deformation

The deformations of clothes (*e.g.* clothing wrinkles and folds) are difficult to be reconstructed using only statistical shape models and pose parameters. We employ the eigen-deformation method [9] to deal with such deformations. We reason that the deformations can be represented by displacements from the skin-tight model to the model with loose clothes and that the displacements can be compressed to a low dimensional space.

If the angles of joints are the same but the orientation of the human body is different, the values of the displacements will change. Therefore, we perform the eigen-deformation method for each body part independently in each body part local coordinate system. We compute the displacement vector d_{lk} of the k-th vertex of the l-th body part in mesh M', which is represented by $d_{lk} = H_l(x_{lk} - x'_{lk})$, where x_{lk} and x'_{lk} are the k-th vertices in the l-th body parts in M and M'. As mentioned above, every vertex is transformed to the local coordinates of each body part using each rigid transformation matrix H_l. On each body part, we concatenate these displacement vectors to make the column vector $d_l^\top = (d_{l1}^\top \; d_{l2}^\top \; \dots)$. We perform these calculations to each frame and then aggregate these displacement vectors over all frames to obtain the matrix $D \in \mathbb{R}^{(3\times K_l)\times F}$. K_l is the number of vertices that belong to l-th body part. F is the number of frames used to perform PCA calculation. With the eigen-deformation, we obtain bases (eigenvectors) $E_l \in \mathbb{R}^{(3\times K_l)\times L}$ and coefficients $c_l \in \mathbb{R}^L$ for each body part respectively. L is the number of bases. So, the displacements in the low dimensional space is represented by

$$\tilde{d_{lk}} = E_l c_l + \overline{d_l}, \tag{7}$$

that is, a linear combination. This means that if we can regress the low dimensional coefficients, unobserved deformations can be estimated.

When reconstructing $\bar{M}$, we firstly recover M' from $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, and then recover $\tilde{d_{lk}}$ using a small number of eigen vectors. Finally $H_l^{-1}\tilde{d_{lk}}$ is calculated for each vertex of each body part and added to M'.

3.4 Shape deformation bases and coefficients estimation by neural network

We train two types of neural networks to estimate the detailed shape of the clothed human from a given pose of the SMPL model: a coefficient regressor network (called CRNN) [9], and a bases estimation network (called BENN). BENN is composed of the CRNN with an additional output layer. In the layer, the bases and the means calculated in the previous section are used as weight and bias parameters and updated in training. This allows the network to output more accurate displacements to regress the displacements directly.

We train one CRNN for each body part using the coefficients obtained from the eigen-deformation of the displacement vectors as training data and the poses $\boldsymbol{\Theta}$, which consist of the combination of the SMPL pose parameters $\boldsymbol{\theta}$ and the angle between each body part and the gravity as input. We reason that the displacements of the vertices of a body part are affected not only by the motion of adjacent body parts, but also by the motion of non adjacent body parts that affect the deformation of the clothes. For example, when a person raises his arms, deformations occur not only on the shoulders but also on the torso of the clothes he wears. Therefore, for each body part we use the pose of multiple joints as input for both the CRNN and the BENN. For each body part, the CRNN is composed of four fully connected layers. As mentioned above, at any time, the body pose is represented using a set of quaternions, which is calculated from $\boldsymbol{\theta}$ and an angle formed by the vertically downward $\Theta \in \mathbb{R}^{5 \times G}$, where G is the number of joints that used for each body part. Once the CRNN is trained, we obtain the coefficients $\tilde{c} \in \mathbb{R}^{L}$ for any body pose Θ.

For each body part we also train a BENN, whose weights and bias are initialized with those of the trained CRNN. The bases E and mean shape $\overline{d}$ obtained at the end of the eigen-deformation stage are used as the initial weights and bias of the additional output layer. The inputs of the BENN are the same as those of the CRNN. The output of the BENN are the displacements $\tilde{d} \in \mathbb{R}^{(3 \times K_l)}$ of all vertices that belong to the body part. The bases are determined by the output of the BENN and the means of all bases are equal to the bias of the BENN(Fig. 2). Finally, the displacement values are computed as:

$$\tilde{d}(\Theta) = W\tilde{c}(\Theta) + b, \tag{8}$$

where W is in $\mathbb{R}^{(3 \times K_l) \times L}$ and b is in $\mathbb{R}^{(3 \times K_l)}$.

Here, when we train the BENN with partially observed data, the previous step will be omitted. However the previous step is still necessary to get a good initial value to converge quickly.

In our case, both network (CRNN and BENN) are implemented in PyTorch [16]. We choose the mean squared error as loss function and Adam algorithm [10], which learning rate is set to 0.001 as optimizer. Examples of the reconstructed shapes are shown in Fig. 4. The results of estimating bases is better than only regress coefficient. Here, we do not perform eigen-deformation and regression for not clothed body part (*e.g.*face and hands) because such parts must not deform.

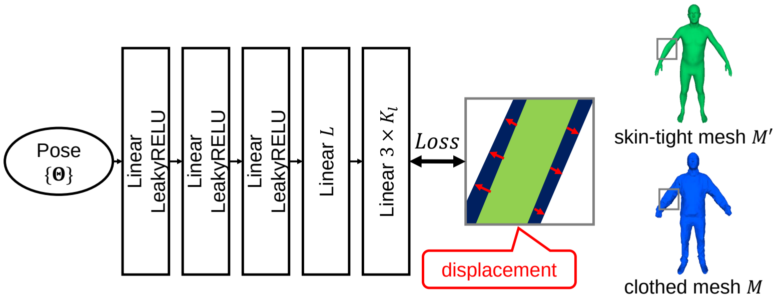

Fig. 2. Network architecture of BENN. When the target data are partially observed, we use visibility map not to calculate the gradient of unobserved body part.

3.5 Handling partially observed data

Our technique can efficiently handle partially observed data by training the BENN using a visibility map to mask the loss of invisible body parts. The visibility map represents whether each vertex of the mesh in the training dataset is visible or not. The map can be created by detecting collisions between the ray from each vertex to the principal point and each mesh. This allows us to back propagate only from the unit of the output layer corresponding to the visible body vertex. To utilize this mask based method, subjects have to show all body parts at least once.

4 Experiment

To evaluate our method, we used the public dataset [22] and compared the results obtained with our proposed method with the results obtained with previous methods [24] [9] quantitatively and qualitatively. We also demonstrate the ability of our proposed method to cope with partially observed data, which we created from the public dataset [22]. We also captured a sequence of RGB-D images of a moving person with three depth cameras and compared qualitatively the results obtained with our proposed method with those obtained with the method proposed in [26].

4.1 Non-rigid Registration using SMPL model

We performed non-rigid registration as explained in Sec. 3.2, to the meshes publicly available from [22], which consists of people with loose clothes, and point clouds captured by Kinect V2. Results are shown in Fig. 3, where the top row is the target scan data and bottom row is the result of our non-rigid registration. From the results, we can confirm that most parts of our fully registered mesh (clothed mesh) are visually close to the original mesh. However, some differences are still observable (*e.g.* there are some gap around crotch in Fig. 3 (c)). Although this problem occurs only about 2% of the entire sequence, it can interfere with the training of NN. Therefore, at this moment, problematic frames are manually adjusted. We will investigate how to solve this problem in our future work.

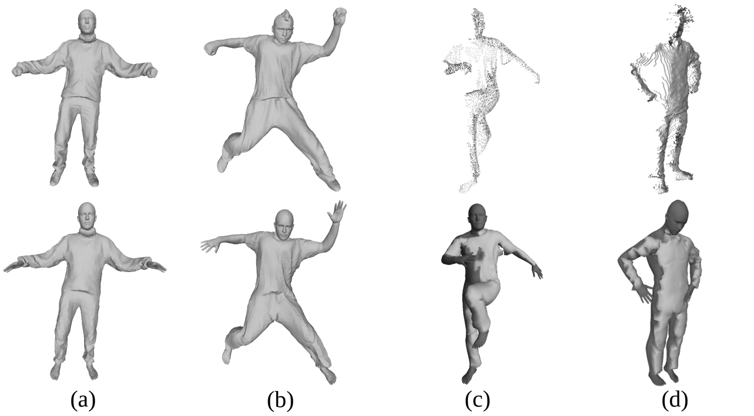

Fig. 3. The result of non-rigid registration (Sec. 3.2). Top row shows target scan data and bottom shows the result of non-rigid registration. (a) and (b) shows the result of full-body target data(For the visibility, the target scans are showed with mesh which is not used). (c) and (d) shows the result of partially observed data (dark gray means invisible body parts and light gray means visible body parts).

Table 1. Average RMSE of each vertex position in *mm*. Ours means the result of BENN. (The value of first row is cited from [24].)

Seq	Bounc.	Hand.	Crane	Jump.	Mar. 1	Mar. 2	Squ. 1	Squ. 2
Yang *et al.* [24]	10.27	-	**4.27**	-	3.93	-	4.31	-
Kimura *et al.* [9] (CRNN)	5.45	3.45	7.61	**9.24**	3.06	8.25	3.20	7.36
Ours (BENN)	**5.28**	**3.03**	7.70	9.52	**2.87**	**8.13**	**2.87**	**7.29**

4.2 Inter-frame interpolation

In this section, we compare the reconstruction results obtained with our proposed method with those obtained with the previous methods [9] [24] using the publicly available dataset [22], which consists of human mesh with loose clothes. On one hand, Kimura *et al.* [9] have the same strategy as our method, which means that they also firstly fit the model to the target scan data and regress clothes deformations. On the other hand, Yang *et al.* [24] calculate the displacements between models and target mesh without fitting strategy. Instead, they assume that the input scan data has mesh consistency.

Following Yang *et al.* [24], we also use 80% of each sequence for training and others for testing. Although Yang *et al.* used 40 bases for regression, we and Kimura *et al.* [9] (CRNN) use only 30 bases. The quantitative comparison result is shown in Table 1. In this table, **Ours** means the result of BENN. The table reports average of RMSE (root mean squared error) in *mm* of each vertex. Here, because the density of vertices is different between the target point clouds and our outputs, we could not simply calculate the error using one-to-one correspondence of points between them. Therefore, we calculate the error between clothed meshes obtained at non-rigid registration stage (Sec. 3.2) and

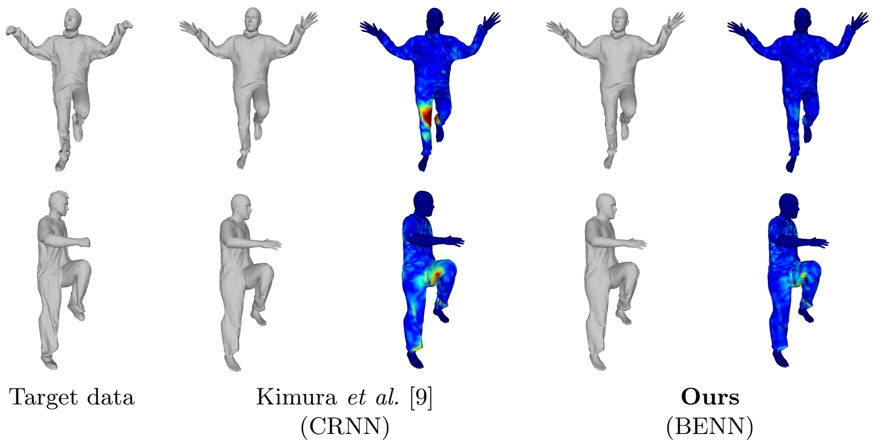

Fig. 4. Qualitative comparison using the public dataset. The left most column shows target data, next two columns show the results of Kimura *et al.* [9] (CRNN) and last two column show the results of the our proposed method (BENN). The second column of each results shows error in color.(Blue = $0mm$ and red >= $50mm$)

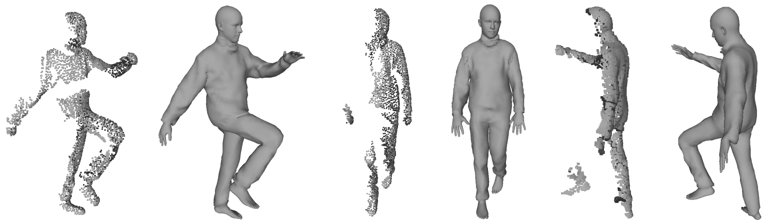

Fig. 5. Each left column shows the target scan data of non-rigid registration and right shows the results of intra-frame interpolation. Even in the invisible body part the clothes deformations are reconstructed well.

the results of our proposed method. It is reasonable to calculate the error in this way because we use clothed meshes as training data for BENN. Our method outperforms previous methods in most sequences while using less bases than Yang *et al.* and the same amount of bases as Kimura *et al.*. The qualitative comparison is shown in Fig. 4. The first column shows the target data of non-rigid registration. Next two columns show the results obtained with the method of Kimura *et al.* (CRNN) and the right two columns show the results obtained with our proposed method. The first column of each set of two columns shows the results of regressed deformation and the second shows the error with pseudo-color (blue = $0mm$ and red >= $50mm$). From these results, we can also confirm that our proposed BENN works well qualitatively.

4.3 Intra-frame interpolation

Owing to our BENN implementation, we can ignore the invisible parts of the scanned data. This allows us to use partially observed data to train our network.

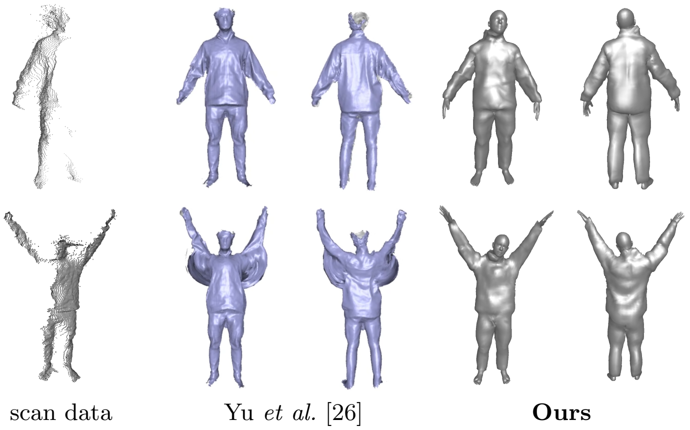

Fig. 6. Qualitative comparison with [26] using Kinect V2. Left column of each result shows front view and right shows back-side view. Our result represents pose-dependent deformation (*e.g.*clothes back). DoubleFusion has some artifacts (*e.g.*armpit), whereas no artifacts in ours.

In this section, we estimate clothes deformations of such an invisible body part using BENN trained with only the visible body part data. We performed intra-frame interpolation to the partially observed data that we generated from the public dataset [22]. We also captured a sequence of RGB-D images of human in motion using Kinect V2 and performed intra-frame interpolation to compare the results obtained with our proposed method with those obtained with DoubleFusion [26], which is a real-time system that can reconstruct not only the human inner body (optimized parameters of the SMPL model) but also the clothes that the subject is wearing, with using a single depth camera (Kinect V2). The results are shown in Fig. 5 and Fig. 6. To obtain enough data for training, we used the data which is visible from the front view and the data which is visible from the back-side view as training data separately. Note that, we only use the displacements of visible body parts to train BENN. Invisible body parts are reconstructed through the regression step. These results show the effectiveness of our proposed method to handle partially observed data. In Fig. 6, the first column shows target data of non-rigid registration for our proposed method and input data for DoubleFusion. Next two columns show the results obtained by [26] and the right two columns show the results obtained by our proposed method. Here, to obtain enough data for training, we used three Kinect V2 cameras, which are placed in circle centered on the subject, to capture the sequence. Again, we use the data which is visible from each view as training data separately. In the case of DoubleFusion, the results were sometimes collapsed (*e.g.*head and armpit). This means that this method deforms the node-graph and cannot describe the pose-dependent deformations. Although the results obtained with DoubleFusion can represent fine details of the clothes, they cannot describe the pose-dependent deformations because this technique only deforms the node-graph. In contrast, the results obtained with our proposed method could represent the pose-dependent deformations (wrinkles around back side in Fig. 6). Again, we do not use the displacements of the occluded body parts from the front view at the network training step.

5 Conclusion

In this paper, we present a full-body human with loose clothes reconstruction method from point clouds measurements using neural network based regressors, even when the target data has largely occluded. Our fitting method and neural network based regression outperform previous methods, which is proved by numerical evaluations. Moreover, our BENN can also reconstruct full-body even if the training data have only partially observed data. We visually confirm that the regression ability is sufficient and better than previous method. We believe that this method have a impact to many applications, such as AR/VR, virtual try-on, avatar making, etc.

References

1. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. In: ACM transactions on graphics (TOG). vol. 24, pp. 408–416. ACM (2005)
2. Bogo, F., Black, M.J., Loper, M., Romero, J.: Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2300–2308 (2015)
3. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Springer International Publishing (Oct 2016)
4. Chen, Y., Liu, Z., Zhang, Z.: Tensor-based human body modeling. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). pp. 105–112 (2013)
5. Habermann, M., Xu, W., Zollhöfer, M., Pons-Moll, G., Theobalt, C.: Livecap: Real-time human performance capture from monocular video. ACM Transactions on Graphics (TOG) **38**(2), 14 (2019)
6. Innmann, M., Zollhöfer, M., Nießner, M., Theobalt, C., Stamminger, M.: Volumedeform: Real-time volumetric non-rigid reconstruction. In: European Conference on Computer Vision. pp. 362–379. Springer (2016)
7. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8320–8329 (2018)
8. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Regognition (CVPR) (2018)
9. Kimura, R., Sayo, A., Dayrit, F.L., Nakashima, Y., Kawasaki, H., Blanco, A., Ikeuchi, K.: Representing a partially observed non-rigid 3d human using eigen-texture and eigen-deformation. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 1043–1048. IEEE (2018)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

11. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
12. Malleson, C., Klaudiny, M., Hilton, A., Guillemaut, J.Y.: Single-view RGBD-based reconstruction of dynamic human geometry. In: IEEE Int. Conf, Computer Vision Workshops. pp. 307–314 (2013)
13. Nealen, A., Igarashi, T., Sorkine, O., Alexa, M.: Laplacian mesh optimization. In: Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia. pp. 381–389. ACM (2006)
14. Newcombe, R.A., Fox, D., Seitz, S.M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 343–352 (2015)
15. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P.V., Schiele, B.: Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. Verona, Italy (2018)
16. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
17. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2019)
18. Pons-Moll, G., Pujades, S., Hu, S., Black, M.: ClothCap: Seamless 4D clothing capture and retargeting. ACM Transactions on Graphics, (Proc. SIGGRAPH) **36**(4) (2017), http://dx.doi.org/10.1145/3072959.3073711, two first authors contributed equally
19. Sorkine, O., Cohen-Or, D., Lipman, Y., Alexa, M., Rössl, C., Seidel, H.P.: Laplacian surface editing. In: Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing. pp. 175–184. ACM (2004)
20. Tan, V., Budvytis, I., Cipolla, R.: Indirect deep structured learning for 3d human body shape and pose prediction (2018)
21. Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., Schmid, C.: BodyNet: Volumetric inference of 3D human body shapes. In: ECCV (2018)
22. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. ACM Trans. Graph. **27**(3), 97:1–97:9 (Aug 2008). https://doi.org/10.1145/1360612.1360696, http://doi.acm.org/10.1145/1360612.1360696
23. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild (2018)
24. Yang, J., Franco, J.S., Hétroy-Wheeler, F., Wuhrer, S.: Analyzing clothing layer deformation statistics of 3d human motions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 237–253 (2018)
25. Yu, T., Guo, K., Xu, F., Dong, Y., Su, Z., Zhao, J., Li, J., Dai, Q., Liu, Y.: Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 910–919 (2017)
26. Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., Liu, Y.: Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7287–7296 (2018)