Coupling dense point cloud correspondence and template model fitting for 3D human pose and shape reconstruction from a single depth image

Xiaofang Wang, Adnane Boukhayma, Stéphanie Prévost, Eric Desjardin, Céline Loscos, Franck Multon

Coupling dense point cloud correspondence and template model fitting for 3D human pose and shape reconstruction from a single depth image

Wang Xiaofang
LICIIS
University of Reims Champagne Ardenne
Reims, France
xiaofang.wang@univ-reims.fr

Boukhayma Adnane
Inria, Univ. Rennes, CNRS, IRISA, M2S
Rennes, France
adnane.boukhayma@inria.fr

Prevost Stéphanie
LICIIS
University of Reims Champagne Ardenne
Reims, France
stephanie.prevost@univ-reims.fr

Desjardin Eric
CReSTIC
University of Reims Champagne Ardenne
Reims, France
eric.desjardin@univ-reims.fr

Loscos Céline
LICIIS
University of Reims Champagne Ardenne
Reims, France
celine.loscos@univ-reims.fr

Multon Franck
Inria, Univ. Rennes, CNRS, IRISA, M2S
City, Country
fmulton@irisa.fr

***Abstract*—In this paper, we address the problem of capturing both the shape and the pose of a character using a single depth sensor. Some previous works proposed to fit a parametric generic human template in the depth image, while others developed deep learning (DL) approaches to find the correspondence between depth pixels and vertices of the template. In this paper, we explore the possibility of combining these two approaches to benefit from their respective advantages. The hypothesis is that DL dense correspondence should provide more accurate information to template model fitting, compared to previous approaches which only use estimated joint position only. Thus, we stacked a state-of-the-art DL dense correspondence method (namely double U-Net) and parametric model fitting (namely Simplify-X). The experiments on the SURREAL [1], DFAUST datasets [2] and a subset of AMASS [3], show that this hybrid approach enables us to enhance pose and shape estimation compared to using DL or model fitting separately. This result opens new perspectives in pose and shape estimation in many applications where complex or invasive motion capture set-ups are impossible, such as sports, dance, ergonomic assessment, etc.**

***Index Terms*—Human motion capture, shape reconstruction, deep learning, computer vision, depth sensor**

I. Introduction

With the dissemination of cheap depth sensors in the consumer market, such as the Microsoft Kinect, reconstructing the body shape and pose from depth images and point clouds has become a very active field of research. Most previously proposed methods focused on pose and shape reconstruction from human part [4], [5], skeleton joints [6], [7], or dense correspondence, but remain not enough precise [8]. Indeed, depth images provide incomplete and noisy information, mainly due to occlusions, which makes it challenging to build a complete and accurate body surface [9]–[13].

Funded by ANR-JPCH "SCHEDAR" project (ANR-17-JPCH-0004). Special thanks to the Centre Image at URCA for their computing resources.

Other approaches proposed to reconstruct the 3D shape and pose using a single RGB image, by fitting parametric models (such as SMPL [14], SMPLify [6]) to the RGB information, or by directly learning parameters of parametric models [15], [16]. Recent work demonstrated that DL was promising to learn the correspondence between the image domain and the SMPL parameter space [17]. However, this correspondence problem is complex because of the high variation in human poses, shapes, and camera viewpoints [17]–[20].

We assume that adapting these RGB-based approaches to depth images is promising. Indeed, using depth instead of RGB images helps to resolve ambiguity from 2D to 3D [7], [21]. We also suppose that combining the advantages of DL-based dense correspondence estimation, with a parametric model fitting for the fine tuning of the shape and the pose, would enhance the results compared to using them separately. Hence, the main hypotheses we wish to validate in this paper are:

- H1 Depth image segmentation before dense correspondence should provide better results. Like [19], [20], as a first step, we establish dense correspondences via mapping 3D vertices to the color domain. We use a DoubleUnet network [22] to obtain this color embedding for each depth pixel. A first U-Net aims at segmenting the depth images into 15 classes (body parts and background), which should help a second U-Net to regress color embedding for each pixel.
- H2 Using dense correspondence as an input of the model fitting algorithm should improve the performance of pose and shape reconstruction. Most of the previous works based on this model fitting used joint position estimation as an input of the optimization, which makes the approach very sensitive to noise and inac-

curacies. We assume that using thousands of pixel-to-vertex correspondences instead of 15 joint positions would increase the accuracy of the reconstruction.

For dense correspondence estimation, we trained a neural network to map depth pixels to a low dimensional canonical template geometry representation (geometry embedding). This representation entails normalized spatial coordinates of the T-pose human SMPL template vertices, in addition to body part segmentation labels. Based on the success of previous works [17], [20], we regress this representation in an image-to-image manner. This pixel-to-vertex correspondence is next used to optimize the shape and pose parameters of SMPL, inspired by previous works on hands [21], [23].

We compared our method to state-of-the-art competition that solves for both monocular RGB and depth inputs on standard human shape in motion datasets following the experimental setting of [12], using synthetic (SURREAL), pseudo-real (DanseDB), and real (DFAUST) data. Our experiments validate the hypothesis H2: by leveraging the combination of deep dense correspondences and parametric model optimization we obtain state-of-the-art performances. But, these results remain less competitive than recent works introducing temporal information [12]. We also provide an in-depth ablative analysis of the various components involved in our method. This ablative analysis supports hypothesis H1: using segmentation into the geometry embedding improves the results compared to using dense correspondence only.

In the following, we first review previous work most related to our approach in section II. Our two-step approach is presented in section III. We present an extensive evaluation study in section IV before concluding.

II. Related Works

Human 3D shape reconstruction and pose estimation have generated vast literature. We refer the reader to [24]–[26] for more extensive overviews.

a) 3D Human Shape and Pose from Depth Images: Previous 3D human body modeling from depth images can be roughly categorized into template-based, template-free capture and hybrid methods. The template-based methods utilize template priors for the 3D body model recovery, such as embedded skeletons [5], [27], template models [9], [11], [28], [29], or parametric models [10], [12]. With the evolution of depth sensing, range data acquired by commodity depth sensors such as the Microsoft Kinect, can be used as prior information. An improved SCAPE model can also be fitted to the range data [8], [30]. Researchers also proposed several different cues from a depth sensor to estimate pose and shape via a silhouette, depth or color data [8], [27], [29], [30]. Bashirov et al. [7] proposed a neural network that used the 3D joints position delivered by the Kinect API to infer SMPL pose parameters. The DoubleFusion approach [10] starts with a pre-constructed 3D template mesh and uses a template-free method (i.e., DynamicFusion [31]) to update the current mesh in combination with the SMPL parametric model to construct an inner human body. Although it shows very promising performances, the initial configuration and subject pre-scanning are not trivial inputs. Recently, DL-based methods have shown impressive performance improvement. Most of these leaning methods rely on 3D human models, such as SMPL. Zhang et al. [32] trained a weakly supervised network from depth or point cloud to learn 3D joints from annotated 2D joints, but they did not recover human shape information. Wang et al. [12] proposed to regress 3D coordinates of mesh vertices at different resolutions from the latent features of point clouds. Jiang et al. [13] also proposed a deep network that takes 3D point cloud as input and learns to predict SMPL shape and pose parameters.

b) 3D Human Shape and Pose from RGB Images: With the development of deep neural networks, capturing a 3D human shape and pose from a single color image has become possible through several diverse approaches [25]. A family of works leveraged 2D joint information in predicting 3D human pose [33] and shape [34]. Bogo et al. [6], when proposing SMPLify, applied a CNN-based method to predict 2D joint locations and then fitted a 3D body SMPL model to estimate 3D body shape and pose. Other methods used regression of 3D human model parameters. They used deep neural networks as encoders to estimate the pose and shape parameters directly from images. For instance, Kolotouros et al. [15] proposed a deep network to infer SMPL parameters through iterating between learnable regression and the optimization-based approach SMPLify [6]. ExPose [35] is a deep neural network predicting the whole set of SMPL-X [36] parameters to overcome the problem of lacking training data for the human body model. Kanazawa et al. [16] employed adversarial learning by using a generator to predict parameters of SMPL, and a discriminator to distinguish the real mesh instances and the predicted ones. Other deep learning-based methods [37]–[40] inferred the 3D body shape or mesh directly from color images using convolutional networks. Graph CNN method [37] first attached the extracted features from an input color image to the 3D vertex coordinates of a template mesh, and then predicted the vertex coordinates of the 3D body meshes using a convolutional mesh regression. Moon et al. [39] proposed a new heat-map representation, called "lixel", to recover 3D human meshes. [40] used image convolutional features and Transformers [41] to estimate a human mesh from a single RGB image.

c) 3D Human Shape and Pose via Dense Correspondences: Several methods enable to compute correspondences between human shapes in arbitrary poses [19], [20], [42]. Finding correspondences across images or point clouds is a fundamental building block for many 2D/3D computer vision tasks, such as reconstruction or tracking. Bogo et al. [8] proposed to optimize parameters of 3D human body model fitted through point cloud corresponded vertices. However, the correspondences were computed by a nearest-neighbor algorithm based on Euclidean distance, which demands a good initialization. Other works relied on an underlying parametric model of a human, such as SMPL, and directly performed a correspondence regression. A strong benefit of this was that the 3D model shares the same topology across different

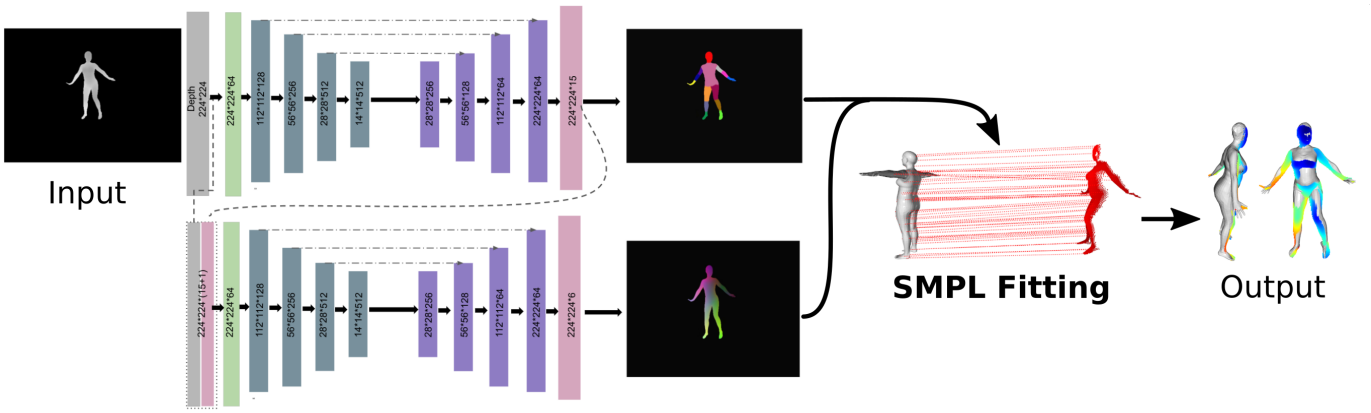

Fig. 1. Overview of the proposed framework. Our method can predict 3D human shape and pose from an input depth image. A double U-Net network is applied to predict body part segmentation and to regress normalized canonical vertex coordinates. These outputs are used to compute dense correspondence between the input depth pixels and the template geometry via nearest neighbor in a low dimensional embedding. We then fit the SMPL model to the input depth by minimizing the distances between vertices and their corresponding depth pixels. The final output is shown on the right hand side from two viewpoints with the overlaid input depth point cloud.

people. DensePose [17] showed that this can be learnt via gathering dense correspondences between the SMPL and body data using the COCO dataset, including simulated data [43]. Another popular type of method consisted in learning feature descriptors attached to RGB, depth, or points cloud. While early works used hand-crafted shape descriptors to identify geometric features [44], Huang et al. [42] used deep learning. They applied PointNet++ [45] to learn a representation of each point cloud, and further enforce local smoothness to compute dense correspondences across full or partial human shapes. They used a depth image as input and learned a descriptor for each pixel. Tan et al. [20] learned an embedding from RGB images that follows the geodesic properties of an underlying 3D surface, which enabled the inference of human correspondences. In this paper, we aim at demonstrating that combining this type of approach with SMPL model fitting should enhance the accuracy of the pose and shape reconstruction.

III. Proposed Approach

Given an input depth image containing a minimally clothed person, our method predicts a mesh representing the corresponding 3D human posed shape in the input camera coordinate frame. This is achieved through the two-stage method depicted in Fig.1. The formalization of the 3D model used in our approach is described in section III-A. In the first step, a convolutional network (see section III-B) first predicts a segmentation of the input depth image into several human body parts, along with a 2D correspondence map associating pixels to a template mesh vertices. Then, pixel-to-vertex correspondences are then established using the segmentation and correspondence maps. In a second step, a parametric shape model is fitted to the depth image using the resulting correspondence maps (see section III-C).

A. 3D Human Model

To model the 3D shape and pose of the character, we used the SMPL model. The shape of a human body is defined by a parametric deformable mesh $\mathcal{M}(\beta, \theta, \gamma)$. Shape parameters β are coefficients of low-dimensional shape space. γ is the global translation. The pose of the body is defined by a skeleton rig with 23 joints; pose parameters θ represent the relative rotation between parts.The model generates a triangle mesh $\mathcal{M}$ with 6890 vertices:

$$\begin{aligned} \mathcal{M}(\beta, \theta, \gamma) &= W(\mathcal{T}_p(\beta, \theta), J(\beta), \theta, \mathcal{W}) + \gamma, \\ \mathcal{T}_p(\beta, \theta) &= \mathcal{T} + B_S(\beta) + B_P(\theta), \end{aligned} \tag{1}$$

where W is a linear blend skinning function with vertex-joint assignment weights $\mathcal{W}$, and $J(\beta)$ is a joint location regressions function. The pose parameters θ encode the global rotation and the rotation angles of each skeleton joint, while the shape parameters β contain the coefficients of the ten most significant PCA components of the human shape learned from registered real human scans. $\mathcal{T}_p(\beta, \theta)$ is the deformed template mesh in a default body pose. It is expressed as the sum of a template mesh $\mathcal{T}$ with the shape and pose blendshape functions B_S and B_P, which add shape and pose dependent vertex-wise corrective displacements to the skinning template in order to reduce the artifacts of linear blend skinning.

B. Dense Correspondences

This section introduces the DL dense correspondence stage of our method: given a depth image and a template geometry mesh, a convolutional neural network predicts a dense mapping between the depth pixels and the SMPL template vertices. Mapping is obtained through the combination of a body part segmentation map and a pixel-to-vertex correspondence map.

a) Template Geometry Embedding: Our goal is to establish a mapping function $c : \Gamma \rightarrow \mathcal{T}$ putting pixels in the depth image domain $i \in \Gamma$ in correspondence with vertices in the

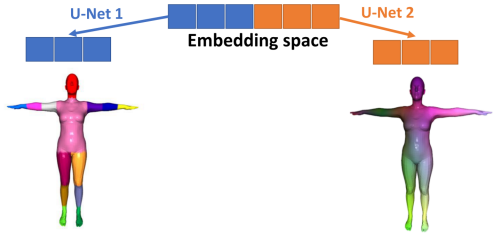

Fig. 2. Visualization of the six-dimensional template geometry embedding that we use to build pixel-to-vertex correspondences. It is a combination of a fifteen-part segmentation labelled with three-dimensional normalized color (left), and three normalized canonical vertex coordinates (right).

template mesh $j \in \mathcal{T}$ using a deep neural network. As it is computationally expensive to learn a mapping from pixels to the entire span of the 6890 template vertices, we embedded the template geometry in a low dimensional space. We note this fixed embedding $E : \mathcal{T} \rightarrow [\![0, 1]\!]^6$. Building a reliable embedding E is crucial for our method, especially because neural networks can infer erroneous correspondences: switching body limbs due to the inherent symmetry of the human body, or confusing pixels belonging to adjacent body parts. E aims at mapping pixels to the template geometry, but images are 2D projections of the 3D world, this embedding must capture the underlying 3D geometry of the human shape under arbitrary poses and viewing angles. We started by defining the first 3 components of the embedding as the normalized 3 spatial coordinates of the template mesh in the canonical T-pose (see Fig. 2), by mapping the 3 normalized vertex coordinates to RGB values. As we found this representation insufficient to distinguish vertices in our experiments, we added 3 extra dimensions to our embedding to help us distinguish body parts more robustly. Hence, we divided the template geometry into 15 parts, including a background class, as shown in Fig. 2. We picked 15 distinctive RGB colors for each class, which represents the extra 3 coordinates of the embedding E. Next, we trained a deep neural network slm to map pixels to the template geometry embedding space: $slm : \Gamma \rightarrow E(\mathcal{T})$.

b) Neural Network: We stacked two U-Net [22] networks (image-to-image architecture) as illustrated in Fig. 1 to predict body part segmentation, and to regress normalized mesh colors. These two outputs are concatenated to generate the pixel embedding values $slm(i) \in E(\mathcal{T})$. The first U-Net aims at segmenting the input depth image into 15 classes. This segmentation label corresponds to the last three components of the embedding for each depth pixel. This embedding is then concatenated with the depth image and fed to the second U-Net to predict a 3-channel image corresponding to the 3 first components of the embedding. The network was trained using the combination of a cross-entropy loss on the output of the segmentation branch, and an L_2 loss on the output of the normalized color regression branch.

c) Pixel-to-Vertex Correspondence: To obtain correspondences v_c for a given depth image to the template geometry, we first map the image pixels to the low dimensional embedding using our convolutional neural network inference slm. The vertex j matching pixel i is then defined as the nearest template vertex in the embedding space, which writes:

$$v_c(i) = \arg\min_{j \in \mathcal{T}} ||slm(i) - E(j)||_2^2 \quad (2)$$

C. Model Fitting

In this section, we introduce the model fitting stage of our method. Given an input depth image and pixel-to-vertex correspondences obtained from the previous stage, we fit the SMPL model to the depth image to recover the human shape and pose parameters of the adapted template mesh. To this end, we minimize the following objective function:

$$E(\theta, \beta, \gamma) = \lambda_D E_D(\theta, \beta, \gamma) + \lambda_\theta E_\theta(\theta) + \lambda_\beta E_\beta(\beta). \quad (3)$$

where E_D is the data term. The data term stands for minimizing a L_2 distance between pixel i's 3D point p_i (obtained using the intrinsic matrix and the pixel's depth value), and the corresponding vertex $v_c(i)$. This distance is summed over all pixels that belong to the body region $\Omega \subset \Gamma$ in the segmentation map:

$$E_D(\theta, \beta, \gamma) = \frac{1}{|\Omega|} \sum_{p_i \in \Omega} \rho(||p_i - v_c(i)||_2^2), \quad (4)$$

where $|\Omega|$ is the total number of pixels in Ω. Following previous work [6], [36], we use a robust differential Geman-McClure penalty function ρ to deal with noisy estimates.

E_θ represents the body pose prior $E_\theta(\theta) = \sum exp(\theta_i)$ which penalizes joints that bend unnaturally. The shape prior E_β implements an L_2 regularization on the shape parameters $E_\beta(\beta) = ||\beta||^2$. Hyper parameters $\lambda_D, \lambda_\theta, \lambda_\beta$ are trade-off weights of the objective function terms.

IV. Evaluation

In this section, we describe the experimental setting used to evaluate this approach in subsection IV-A. We also provide implementation details in subsection IV-B and explain the evaluation metrics in subsection IV-C. This evaluation aims at providing quantitative results compared to state-of-the-art competitors (see subsection IV-D) and qualitative results of the reconstructed shape and pose character (see subsection IV-F). To evaluate the two hypotheses $H1$ (prior segmentation enhances dense correspondence) and $H2$ (combining dense correspondence and model fitting enhance the quality of pose and shape reconstruction), we carried out a specific ablation study, described in subsection IV-E.

A. Datasets

We conducted experiments on standard datasets of 3D minimally clothed human shape in motion. Following protocols introduced in previous works [12], [13], we used the SURREAL [1], DFAUST [2] and also a subset of the AMASS [3] dataset entitled DanseDB[1]. We rendered the 3D models contained in these datasets to simulate depth images of the same resolutions

[1] http://dancedb.eu/

but different viewpoints. Consequently, we obtained depth images with the corresponding ground truth character shape and pose, initially used in the rendering process.

The SURREAL data consists of motion sequences of synthetic human 3D body models. It contains $55,001$ training clips and $12,528$ testing clips, each is roughly 100 frames long. The DFAUST data consists of motion sequences of registered real people scans. It contains around $40,000$ 3D models. The DanseDB data consists of dancing sequences of synthetic human models fitted to real motion capture data. It contains 20 subjects and 153 sequences totaling roughly 3.5 hours of dancing motion. We uniformly sampled $50,000$ training frames and $10,000$ testing frames from 4 different random viewpoints, for each of these datasets.

B. Implementation Details

We trained our neural network with a batchsize of 12 using the RMSprop optimizer and a learning rate of $6.14e^{-4}$ found through the learning rate range test described in [46]. To solve the optimization in Eq.3, similar to [36], we used the Pytorch implementation of the Limited-memory BFGS optimizer, with strong Wolfe line search with a learning rate of 0.6. We ran the optimization for 20 iterations. At test time, the neural network inference stage took about 35ms (may change if size of the input depth image changes) and the optimization stage took 6.43s on a NVIDIA 1080Ti GPU.

C. Evaluation Metric

We quantified the reconstruction quality with the Mean Average Vertex Error in millimeter (mm), averaged subsequently over all testing frames:

$$\epsilon = \frac{1}{N}\sum_{i=1}^{N}\sqrt{||v_i - \hat{v}_i||_2^2}, \tag{5}$$

where vertices $\{v_i\}$ are the prediction, $\{\hat{v}_i\}$ are the ground-truth, and N is the total number of vertices.

D. Comparison to the State-of-the-art

Methods	Input	SURREAL	DFAUST	DanseDB
Model fitting [14]	D	140.6	110.1	-
Lassner at al. [38]	RGB	155.5	-	-
Bogo et al. [6]	RGB	56.1	57.5	-
Wei et al. [19]	D	58.6	62.2	-
Kanaz. et al. [16]	RGB	54.3	58.1	-
Kanaz. et al. [47]	RGB	52.7	56.1	-
Kolot. et al. [37]	RGB	49.5	52.2	-
Wang et al. [12]	D	18.2	19.7	-
Jiang et al. [13]	D	15.5	8.1	-
Ours	D	**49.6**	**53.6**	**55.0**

TABLE I
COMPARISON TO METHODS PREDICTING THE MESH OF A 3D MINIMALLY-CLOTHED HUMAN BODY FROM A MONOCULAR RGB OR DEPTH (D) IMAGE IN TERMS OF RECONSTRUCTION ERRORS (MM). FOR CLARITY, THE ERROR RATES CLOSE, ABOVE AND BELOW 10% OF OUR SOLUTION HAVE BEEN WRITTEN RESPECTIVELY IN BLUE, ORANGE AND GREEN.

We compared our proposed approach to state-of-the-art methods that predict human shape and pose from a single RGB or depth image, in Table I. We reported the performance of methods [6], [14], [16], [19], [37], [47] as they are reported in [12], because we used similar evaluation protocol. The model-fitting method [14] deformed the SMPL model to the depth using naive correspondences between the template and the input depth. Kanazawa et al. [6] first detected 2D body joints from an RGB image and then fitted the SMPL model to the detected joints. Lassner et al. [38] and Kanazawa et al. [16] inferred SMPL parameters directly from RGB images. Wei et al. [19] built point correspondences by matching learned feature descriptors for pixels in depth images. The 3D models were then generated by fitting the SMPL model to point correspondences. The rest of the methods in Table I directly inferred 3D meshes from RGB [37] or depth images [12], [13]. These two last methods based on depth images also used temporal information, which helps to reconstruct missing information and to obtain consistent shape along time, and continuity of the motion.

Our method has similar results to most previous works, except for the two recent papers using temporal information [12], [13]. The superiority of our method compared to the RGB deep learning-based ones (e.g. [6], [16], [37], [38], [47]) could be partly explained by using 3D cues instead of 2D RGB information. Model fitting approaches [14] struggle to obtain good results due to the difficulty of getting a good initialization for the optimization using merely naive initialization heuristics.

E. Ablative Analysis

In this paper, we wish to evaluate two main hypotheses $H1$ and $H2$. To this end, ablative analysis is required, aiming at evaluating the impact of each choice in the approach: the interest of introducing segmentation prior to dense correspondence ($H1$) and using or not dense correspondence before model-fitting ($H2$). For this ablative analysis (see Table II), we focused on the SURREAL dataset. We first tested an approach in which the 3D joint coordinates are known and used as inputs to the model-fitting algorithm. This is the reference model fitting approach. Note that the 3D joints are ground truth values, leading to better results than previous works based on 2D-3D joint estimation [14], [38]. Because 3D joint estimation is a very active field of research, we considered it as a fair comparison with our approach, to use the ultimate quality of joint reconstruction, i.e., using ground truth.

We then tested using DL dense correspondence only with a single UNet before model-fitting. The results showed that the pose and shape reconstruction was enhanced compared to the reference model fitting method (based on ground truth joints). We then added the prior segmentation stage to evaluate hypothesis $H1$. Compared to using dense correspondence only, we decreased the error from $59.7mm$ to $49.6mm$, which means that it avoids some mismatch in the dense correspondence algorithm, supporting $H1$.

To evaluate the limit of this approach and perform a fair comparison to the reference model-fitting approach, we tested another method using the ground truth dense correspondence.

Using this knowledge before model fitting, we decreased the error down to $44.3mm$, which is far below the error obtained with the reference model fitting test ($80.1mm$), thus supporting hypothesis $H2$.

Method	SURREAL
GT 3D joints + optimization	80.1
Dense correspondence only + optim.	59.7
Our method: Segmentation & correspondence + optim.	49.6
GT dense correspondence + optim.	44.3

TABLE II
RECONSTRUCTION ERRORS (MM) OF DIFFERENT ABLATION STUDIES PERFORMED ON THE SURREAL DATASET.

F. Qualitative Results

Fig. 3 shows our results on test sets from SURREAL, DFAUST and DanseDB datasets. From left to right, one can see the input depth image, the segmentation, the color embedding, the dense correspondence between depth pixels and template vertices, and two different viewpoints of the resulting shape. We also display an overlay of the point cloud on the fitted mesh. These visualizations illustrate the robustness of our method to changes in body poses, shapes, self-occlusions, and viewpoints.

Figure 4 depicts the vertex-wise color-coded reconstruction errors of our method on SURREAL, DFAUST and DanseDB datasets. It can be seen that in most cases, our reconstruction is extremely precise. This figure also illustrates the remaining limits not yet solved by our approach. The mesh-born spatial distribution of errors shows that reconstruction is most challenging for non-frontal views, body part extremities, and self-occluded body areas.

V. CONCLUSION

The main contribution of this paper is to explore the interest of combining DL dense correspondence between depth pixels and human template vertices, and model fitting. Our results demonstrated that this combination enhances the accuracy of the human pose and shape reconstruction using a single depth image. Indeed, the thousands of correspondences used as inputs to the model fitting stage offer richer information than simply using joint positions, as traditional methods did. This supports our second hypothesis $H2$.

Dense correspondence may contain errors due to the complexity of human pose, shape, symmetry, and camera viewpoint. Our results support hypothesis $H1$ and suggest that prior depth image segmentation in body parts helps to enhance the dense correspondence estimation. However, this prior segmentation knowledge did not enable us to reach the ideal performance obtained with ground truth correspondence. It means that future work is needed to make this dense correspondence become more robust and accurate.

Although we demonstrated the interest of combining DL and model fitting, compared to DL only, recent works using temporal information obtained very impressive improvements. It seems to show that working on a unique RGB or depth image is limited, and future work should continue to explore using temporal information. It would also be interesting to test if combining DL and model fitting is still an advantage.

In this paper, as in many previous works, we tested our approach on simulated depth images, whereas Kolotouros et al. [15] showed to perform well on 'in-the-wild' data sets. Real captured depth images may exhibit low quality, which may impact the quality of the pose and shape reconstruction. Future work should evaluate these methods on real segmented depth images, to address real-world challenges.

REFERENCES

[1] 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.

[2] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black, "Dynamic faust: Registering human bodies in motion," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6233–6242.

[3] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "AMASS: Archive of motion capture as surface shapes," in *International Conference on Computer Vision*, Oct. 2019, pp. 5442–5451.

[4] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman *et al.*, "Efficient human pose estimation from single depth images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2821–2840, 2012.

[5] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *CVPR 2011*, pp. 1297–1304, 2011.

[6] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *European conference on computer vision*. Springer, 2016, pp. 561–578.

[7] R. Bashirov, A. Ianina, K. Iskakov, Y. Kononenko, V. Strizhkova, V. Lempitsky, and A. Vakhitov, "Real-time rgbd-based extended body pose estimation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2807–2816.

[8] F. Bogo, M. J. Black, M. Loper, and J. Romero, "Detailed full-body reconstructions of moving people from monocular RGB-D sequences," in *International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 2300–2308.

[9] A. Baak, M. Müller, G. Bharaj, H.-P. Seidel, and C. Theobalt, "A data-driven approach for real-time full body pose reconstruction from a depth camera," in *2011 International Conference on Computer Vision*, 2011, pp. 1092–1099.

[10] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu, "Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor," in *The IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*. IEEE, June 2018.

[11] K. Guo, F. Xu, Y. Wang, Y. Liu, and Q. Dai, "Robust non-rigid motion tracking and surface reconstruction using l0 regularization," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3083–3091.

[12] K. Wang, J. Xie, G. Zhang, L. Liu, and J. Yang, "Sequential 3d human pose and shape estimation from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7275–7284.

[13] H. Jiang, J. Cai, and J. Zheng, "Skeleton-aware 3d human shape reconstruction from point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5431–5441.

[14] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[15] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *ICCV*, 2019.

[16] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

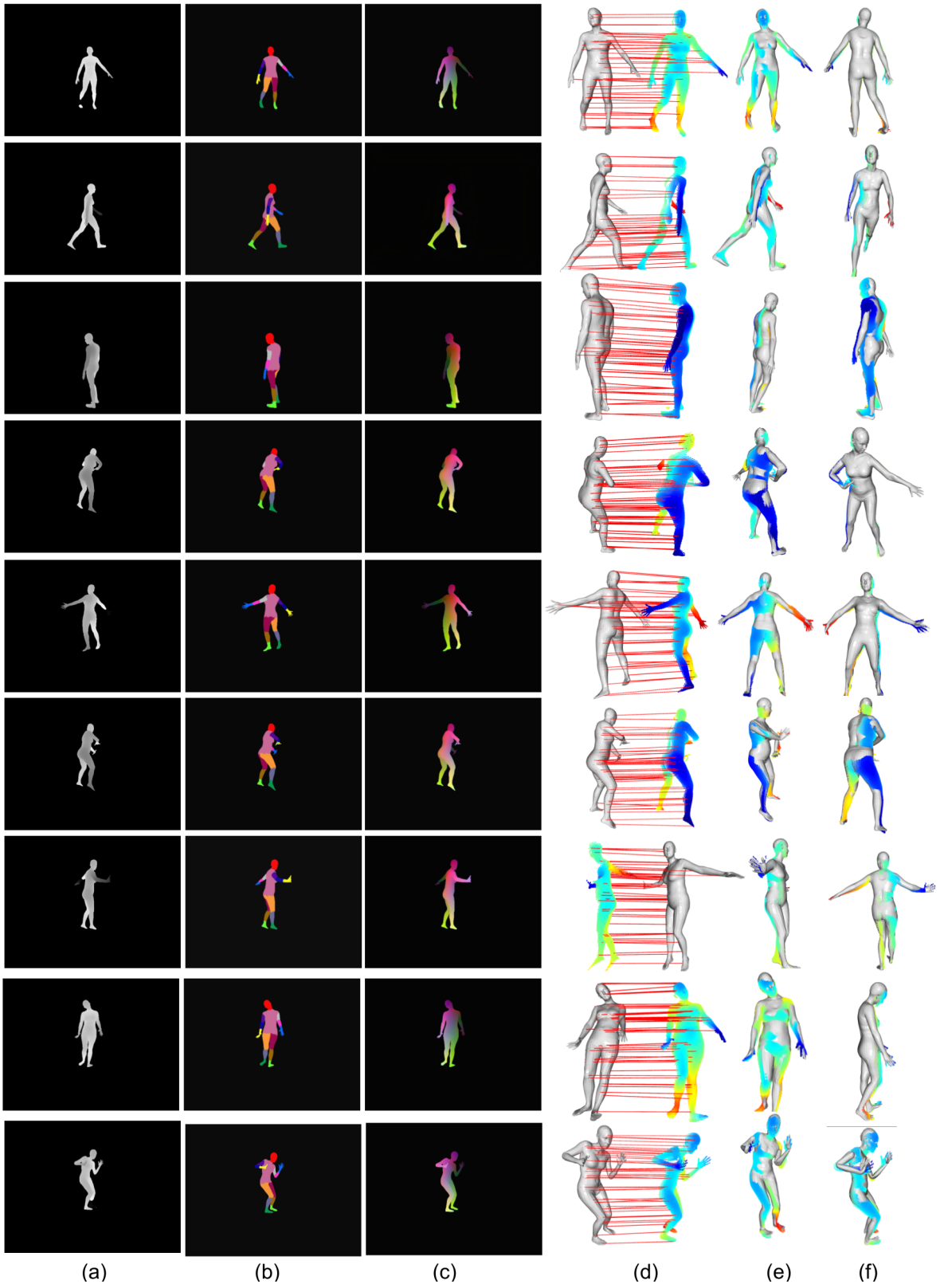

Fig. 3. Visualization of our results on SURREAL (rows 1-3), DFAUST (rows 4-6) and DanseDB (rows 7-9). (a) Input depth image; (b) Output human part segmentation; (c) Regressed template vertex color; (d) Correspondences between the depth point cloud and the fitted mesh; (e) & (f) Output fitted mesh visualized from 2 different viewpoints. The point cloud, overlaid on the fitted mesh, is colored according to the depth values.

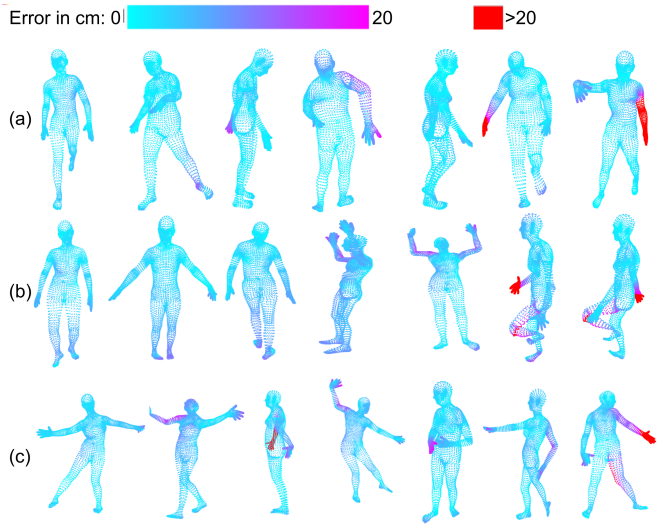

Fig. 4. Visualization of spatial distribution of reconstruction errors on (a) SURREAL, (b) DFAUST and (c) DanseDB. Vertex errors under 20cm are shown using the top left colormap, those larger than 20cm are shown in red. While most errors are close to few millimeters (light blue), large errors remain in challenging cases, like side views and self-occluded areas.

[17] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7297–7306.

[18] L. Wei, Q. Huang, D. Ceylan, E. Vouga, and H. Li, "Dense human body correspondences using convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1544–1553.

[19] ——, "Dense human body correspondences using convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1544–1553.

[20] F. Tan, D. Tang, M. Dou, K. Guo, R. Pandey, C. Keskin, R. Du, D. Sun, S. Bouaziz, S. Fanello *et al.*, "Humangps: Geodesic preserving feature for dense human correspondences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1820–1830.

[21] F. Mueller, M. Davis, F. Bernard, O. Sotnychenko, M. Verschoor, M. A. Otaduy, D. Casas, and C. Theobalt, "Real-time Pose and Shape Reconstruction of Two Interacting Hands With a Single Depth Camera," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, 2019.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[23] J. Wang, F. Mueller, F. Bernard, S. Sorli, O. Sotnychenko, N. Qian, M. A. Otaduy, D. Casas, and C. Theobalt, "Rgb2hands: Real-time tracking of 3d hand interactions from monocular rgb video," vol. 39, no. 6, 2020.

[24] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer vision and image understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.

[25] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Computer Vision and Image Understanding*, vol. 192, p. 102897, 2020.

[26] I. Jegham, A. B. Khalifa, I. Alouani, and M. A. Mahjoub, "Vision-based human action recognition: An overview and real world challenges," *Forensic Science International: Digital Investigation*, vol. 32, p. 200901, 2020.

[27] M. Ye, Y. Shen, C. Du, Z. Pan, and R. Yang, "Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 1517–1532, 2016.

[28] G. Mishra, S. Saini, K. Varanasi, and P. J. Narayanan, "Human shape capture and tracking at home," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 390–399.

[29] Q. Zhang, B. Fu, M. Ye, and R. Yang, "Quality dynamic human body modeling using a single low-cost depth camera," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 676–683.

[30] A. Weiss, D. Hirshberg, and M. J. Black, "Home 3d body scans from noisy image and range data," in *2011 International Conference on Computer Vision*, 2011, pp. 1951–1958.

[31] R. A. Newcombe, D. Fox, and S. M. Seitz, "Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 343–352.

[32] Z. Zhang, L. Hu, X. Deng, and S. Xia, "Weakly supervised adversarial learning for 3d human pose estimation from point clouds," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 5, pp. 1851–1859, 2020.

[33] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, and C. Theobalt, "XNect: Real-time multi-person 3D motion capture with a single RGB camera," vol. 39, no. 4, July 2020. [Online]. Available: http://gvv.mpi-inf.mpg.de/projects/XNect/

[34] H. Choi, G. Moon, and K. M. Lee, "Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose," in *European Conference on Computer Vision (ECCV)*, 2020.

[35] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black, "Monocular expressive body regression through body-driven attention," in *European Conference on Computer Vision*. Springer, 2020, pp. 20–40.

[36] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3d hands, face, and body from a single image," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[37] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *CVPR*, 2019.

[38] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler, "Unite the people: Closing the loop between 3d and 2d human representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Jul. 2017. [Online]. Available: http://up.is.tuebingen.mpg.de

[39] G. Moon and K. M. Lee, "I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image," in *European Conference on Computer Vision (ECCV)*, 2020.

[40] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1954–1963.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[42] X. Huang, H. Yang, E. Vouga, and Q. Huang, "Dense correspondences between human bodies via learning transformation synchronization on graphs," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[43] T. L. Zhu, P. Karlsson, and C. Bregler, "Simpose: Effectively learning densepose and surface normals of people from simulated data," in *ECCV*, 2020.

[44] R. Litman and A. Bronstein, "Learning spectral descriptors for deformable shape correspondence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 171–180, 2014.

[45] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 5105–5114.

[46] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2017, pp. 464–472.

[47] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3d human dynamics from video," in *Computer Vision and Pattern Recognition (CVPR)*, 2019.