Data-Driven 3D Reconstruction of Dressed Humans From Sparse Views

Pierre Zins, Yuanlu Xu, Edmond Boyer, Stefanie Wuhrer, Tony Tung

Data-Driven 3D Reconstruction of Dressed Humans From Sparse Views

Pierre Zins[1,2] Yuanlu Xu[2] Edmond Boyer[1] Stefanie Wuhrer[1] Tony Tung[2]
[1]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP [*], LJK, 38000 Grenoble, France
[2]Facebook Reality Labs, Sausalito, USA
name.surname@inria.fr, merayxu@gmail.com, tony.tung@fb.com

Abstract

Recently, data-driven single-view reconstruction methods have shown great progress in modeling 3D dressed humans. However, such methods suffer heavily from depth ambiguities and occlusions inherent to single view inputs. In this paper, we tackle this problem by considering a small set of input views and investigate the best strategy to suitably exploit information from these views. We propose a data-driven end-to-end approach that reconstructs an implicit 3D representation of dressed humans from sparse camera views. Specifically, we introduce three key components: first a spatially consistent reconstruction that allows for arbitrary placement of the person in the input views using a perspective camera model; second an attention-based fusion layer that learns to aggregate visual information from several viewpoints; and third a mechanism that encodes local 3D patterns under the multi-view context. In the experiments, we show the proposed approach outperforms the state of the art on standard data both quantitatively and qualitatively. To demonstrate the spatially consistent reconstruction, we apply our approach to dynamic scenes. Additionally, we apply our method on real data acquired with a multi-camera platform and demonstrate our approach can obtain results comparable to multi-view stereo with dramatically less views. Code is released at https://gitlab.inria.fr/pzins/data-driven-3d-reconstruction-of-dressed-humans-from-sparse-views/.

1. Introduction

The ability to produce accurate visual models of real humans in every-day context, in particular with their clothing and accessories, is useful in a wide range of applications that deal with captured human avatars, typically in the virtual and augmented reality or telepresence domains. Using images for that purpose has been an active field of research for decades, with issues that result, in part, from the high dimensionality of the space of human shapes and appearances, especially with dressed people. The challenge is accentuated when only few viewpoints are considered, a situation that is, on the other hand, common in many practical contexts, for instance with mobile devices. While model based strategies (*e.g.*, SMPL [32]) have shown impressive results in case of undressed bodies, they cannot easily generalize to generic humans with clothing and accessories. This paper investigates how to recover such 3D models by combining information from sparse calibrated views.

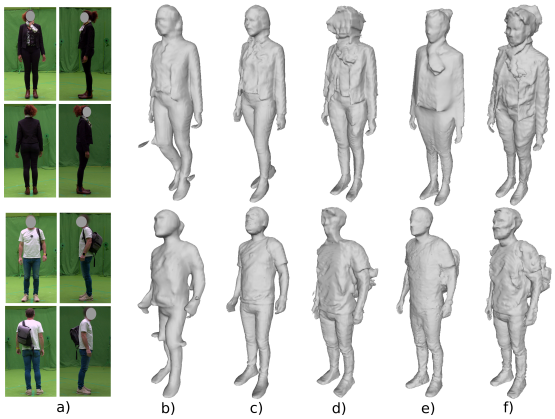

Figure 1. a) Real scene cropped images. b) PIFu [45] and c) PIFuHD [46] with a single frontal view. d) PIFu with 4 views. e) Multi-view stereo [30] reconstruction with 60 views. f) Our method with 4 views.

Acquiring 3D human models from images is a long standing research topic in computer vision. When images from several viewpoints are available, multi-view stereo approaches (*e.g.* [47, 15]), and their learning based extensions (*e.g.* [25, 30]), allow for highly detailed 3D reconstructions by combining multi-view information with photo-consistency criteria. This generative strategy builds on photo-metric redundancy among input images and tends to fail with only sparse viewpoints. Besides, data-driven reconstruction methods, that only require a single view,

[*]Institute of Engineering Univ. Grenoble Alpes

have been proposed. This includes methods based on low-dimensional parametric models (*e.g.* [40]) which are anyway limited with clothing and accessories; methods based on volumetric representations (*e.g.* [55]) with bounded level-of-details by construction; and methods based on implicitly defined continuous neural representations (*e.g.* [45]). These latter methods have demonstrated their ability to recover humans with clothing and accessories. Yet, the single-view reconstruction problem is highly ambiguous and results easily suffer from artifacts when the input scene differs substantially from the training set. To remedy this, methods accounting for multiple input views have been proposed *e.g.* [22, 45]. These extensions, however, merely combine single-view estimations with simple average pooling. Such ways of fusion do not fully exploit multi-view cues and are still plagued by single-view ambiguities.

In this paper, we adopt the widely approved implicit neural representations and focus on multi-view fusion. With respect to single-view estimation this task raises several issues. First, single-view reconstruction methods generally assume a person centered and scaled input image. This needs to be compensated for when dealing with sequences of moving humans and in order to obtain spatially consistent reconstruction with coherent localization and scales among the sequence frames. The second question is how to aggregate local information from viewpoints that can differ significantly, for instance front and side-views, and which can therefore predict different occupancy at a given spatial location. The third issue is how to account for local contexts, defined by image color cues around a 3D point, that gain in variability with increasing views but also allow to better differentiate local geometric patterns. To address these issues, we propose a data-driven end-to-end approach that reconstructs a 3D model of the dressed human from sparse camera views using an implicit representation. Specifically, our method has three key components:

- A spatially consistent 3D reconstruction framework that allows for arbitrary placement of the human in the scene that uses the perspective camera model, achieved by learning the model in a canonical coordinate system and by accounting for the transformation of each input view to this system.
- A learnable attention-based fusion layer that weighs view contributions. This layer implements a multi-head self-attention mechanism inspired by the transformer network [56].
- A local 3D context encoding layer that better generalizes over the local geometric configurations, which is implemented through randomized 3D local grids.

In the experiments, we evaluate our approach against the state of the art on public benchmarks. To demonstrate the value of the spatially consistent reconstruction, we apply our method to dynamic scenes with large displacements. Moreover we also contribute with results on new real data obtained with a multi-view platform. They demonstrate the feasibility of data-driven approaches in practical real-world capture scenarios, even trained solely on synthetic data.

2. Related Work

In this section, we focus on methods that reconstruct the 3D geometry of humans, possibly in clothing.

Monocular 3D reconstruction is an ill-posed problem, as a result of depth ambiguities and occlusions. Dimension reduction with parametric models is a strategy that has been extensively studied in the past two decades. Early achievements use a set of simple geometric primitives to track and reconstruct humans from monocular video *e.g.* [43, 51]. Statistical human body models learned from 3D scans allow to infer the naked body shape from monocular depth images [3] or color images [4, 20, 19, 9, 27, 41, 28, 59]. Some of these models are even sufficiently detailed to allow capturing facial expressions and hand gestures [40]. More recent techniques tend to directly regress parameters of human body models with deep neural networks [27, 41, 59]. Another line of work uses a 3D template mesh as input and trains a deep neural network to deform or regress the template vertices given a monocular image [40, 68]. All of these methods are limited to undressed human bodies, and cannot reconstruct clothing or accessories. Some methods allow nevertheless for clothing as offsets from a parametric body model based on an input monocular video [2, 7] or single image [67], or using physics-based simulation [65, 39]. Using parametric models, some approaches allow for real-time reconstruction of dynamic humans from a single depth camera [8, 63, 61, 64]. While parametric models allow for interesting solutions, the level of detail and variability of the reconstructed clothing and accessories remain inherently limited. To overcome this problem, alternative representations have been explored. Volumetric representations [55, 24] and methods that estimate novel silhouettes to enable visual hull reconstruction [36] offer the advantage of allowing for more clothing variety at the cost of requiring large memory. Methods that represent the reconstruction using few depth maps [16, 52] are less memory demanding, but cannot represent arbitrary clothing topology. Many recent approaches address the problems of memory efficiency and resolution with implicitly-defined continuous neural representations. A seminal work that uses this representation to reconstruct humans from monocular images is PIFu [45], which learns pixel-aligned implicit functions to locally align image pixels with the global location of the 3D human. Follow-up methods increase the image resolution for higher levels of detail [46], propose animatable reconstructions [23], combine PIFu with a volumetric representation or voxelized model to incorporate global 3D information [21, 66], and combine PIFu with a parametric model

to allow for coherent body reconstruction [6]. Alternative representations propose using tetrahedral truncated signed distance functions [38], and using periodic activation functions to better capture high frequencies [49].

Our work also builds on implicit neural representations for their ability to efficiently encode shape information. However, departing from the single view paradigm we focus on how to leverage several views to overcome some of the limitations of single-view inference.

Multiple View Reconstruction has been researched extensively, and a full review is beyond the scope of this paper. Classical stereo and multi-view stereo techniques reconstruct 3D geometry from a set of images under assumptions, especially photo-coherent Lambertian surfaces, *e.g.*, [47, 15]. More recent methods allow for improved results by learning some parts of the classical multi-view stereo pipelines like the photo-consistency [30] or the depth maps fusion [12]. These methods require short baselines and many views [14, 29], which lead to practical limitations. Methods based on Neural Radiance Field (NeRF) achieve photo-realistic rendering but also require numerous views or images (typically more than 50, up to hundreds) for learning an MLP that represents the scene. They are usually scene specific although some limitations have been explored in recent work [34, 62, 50, 42].

When focusing on humans, several previous methods use a template-based approach to reconstruct the 3D geometry from silhouette information [10, 57, 17, 11]. Some techniques take advantage of low-dimensional parametric models to reconstruct the 3D body from multiple RGB images [5, 26]. When multiple depth images are available, a full 3D human model can be reconstructed by fitting a parametric model to the scans [58] or by globally registering and merging the depth images [31].

Closer to our work, two methods [22, 45] propose the use of implicit representations to reconstruct 3D humans from multiple input images. Unlike our work, these methods combine the views by simply averaging their contributions. They do not handle visibility consistency among the different views, and in particular occlusions. In this paper, we propose a solution based on an implicit representation with a novel learnable attention-based fusion layer that efficiently weighs the available views and outputs a spatially consistent reconstruction.

3. Method

In this section we first give an overview of our method and explain the representation that is used. We then present our strategy to learn and infer humans in a large scene and our contributions with the spatially consistent reconstruction, the attention-based fusion layer and the local context learning.

3.1. Overview

Our pipeline is described in Fig. 2. High resolution images of a human and background masks are used as inputs to reconstruct a spatially consistent 3D model using an implicit representation. To allow for a spatially consistent reconstruction with proper scales and localization, we learn the model in a canonical 3D local coordinate system, and transform each observation to this space. This is achieved by localizing the 2D center of the human in each view, by triangulating to find the 3D position of the human center, and by defining a canonical 3D local coordinate system based on this information. This allows to create canonical crops of the input images and background masks so they can be fed to our deep neural network that learns to predict an implicit 3D reconstruction in a canonical space. The result, combined with the canonical 3D local coordinate system, allows to reconstruct a spatially consistent 3D model in the scene by placing the reconstruction in world coordinates.

Fig 3 gives an overview of our deep neural network for multi-view 3D reconstruction. Image features are first extracted using a standard multi-scale image encoder. Please refer to the supplementary materials for an ablation study on the image encoder. We then sample points by combining two strategies: random sampling in a 3D bounding box and importance sampling close to the surface with half of the points inside the mesh and the other half outside. We also construct a local 3D grid around each sample. Here we describe the method for a single sample but in practice a large number of points are processed in parallel. Using projection and bilinear interpolation, each point of the local grid is associated with a 2D feature, which is concatenated with the depth of the point. It is important to note that the previous steps are performed per-view and in the end a 3D local grid of features is obtained for each view. An attention-based module efficiently combines the information from the different views by merging the 3D local grids. A second fully connected fusion layer extracts a final 3D feature from the local grid. At inference time, we define a grid at the desired resolution, evaluate the occupancy function at every grid location, and apply the Marching Cubes algorithm [33] with a pre-defined threshold of 0.5 to recover a 3D mesh.

3.2. Multi-view Implicit Surface Representation

Following recent progresses in learning-based shape modeling, we use an implicit 3D surface representation for the reconstruction task. Implicit surface representation converts arbitrary mesh surfaces into a function defined on a volume and allows for geometric details to be represented at arbitrary resolution. Furthermore, the use of neural implicit representations is memory-efficient and solves the main issue in other volumetric representations. Similar to methods like [46, 45], our implicit function takes the combination of pixel-aligned features with depth values as input and pre-

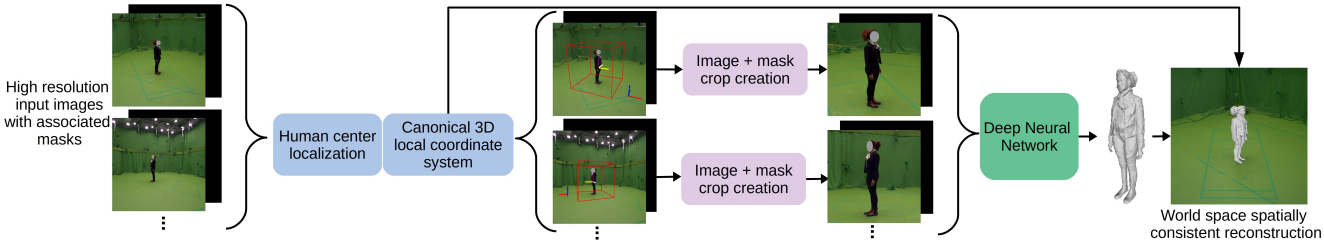

Figure 2. Overview of the proposed pipeline. Given a sparse set of input images with associated background masks and known calibration, our method reconstructs a spatially consistent 3D model.

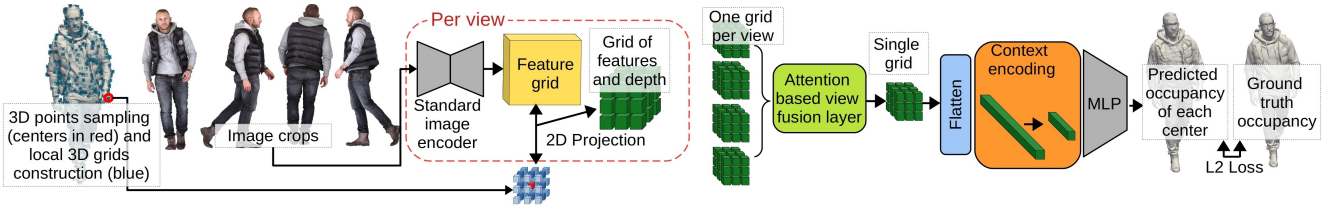

Figure 3. Overview of the deep neural network for multi-view training. Image features are extracted per view, and queried for a local grid around each sample. All views are integrated using an attention-based fusion layer, and a context encoding layer based on 3D convolution is applied before predicting occupancy.

dicts an occupancy probability $o \in [0, 1]$.

When reconstructing from a single image, the conditioning on the depth is necessary to differentiate points on the same camera ray as their appearance features are the same. In our case with multiple views, associations of features can discriminate points of the same view line but the conditioning on the depth is still helpful to capture details as the spatial resolution of the features is limited. To optimally benefit from this conditioning, training examples should all be aligned so that the network can learn a prior of the depth from the training set. Therefore even if we consider reconstruction in large scene, we work in a canonical local coordinate system during training and at inference. The origin of the coordinate system is defined at the center of each training mesh and its orientation is the same as the world coordinate system, so we have the following equation: $X^j_{local} = X_{world} + T_j$, where T_j is the translation between the world origin and the center of the j-th mesh. The exact definition of the center of a mesh is arbitrary but should be consistent for all the training examples. In practice, we use the median over all mesh vertices for the x and z coordinates and the mean between the highest and the lowest vertices for the vertical coordinate y. For each 3D point, the depth value given as input of the implicit function is its z-coordinate in the local coordinate system aligned with each of the cameras by applying the rotation R_i. The implicit function takes the form:

$$f(E_I(K_i\,[R|t]_i\,X_w), z(R_i X_{local}); \theta) = o, \quad [|E| \times \mathbb{R}] \mapsto [0, 1] \tag{1}$$

where X_w is the 3D point in world coordinates, K_i and $[R|t]_i$ are respectively the intrinsic and extrinsic parameters of the i-th camera, o is the occupancy probability at X_w, and $|E|$ the dimension of the 2D image feature. $E_I(...)$ is defined at any location in the image using bilinear interpolation of the values of E_I at pixel locations.

3.3. Spatially Consistent Reconstruction

Most existing works based on pixel-aligned features and implicit representation consider orthographic projection where the appearance of a subject is the same at any position in the scene. In single-view reconstruction, this simplified scenario removes the ambiguity between the size of the subject and its distance from the camera. On the contrary we deal with perspective projection like in real environments with the pinhole camera model. We consider the case where enough views are available to avoid the size versus distance ambiguity. To accommodate for perspective deformations, we augment the data during training by randomly placing the subjects in the scene. As we are learning an implicit representation in a canonical 3D local coordinate system, the reconstruction at inference is inconsistent with the world space. Previous work tackles this problem with a neural network that estimates the spatial transformation of humans from a single image [35]. In our context, we propose to take advantage of the multiple views and triangulate the 3D coordinates from multiple 2D detections of the center of the human as shown in Fig. 2. The 2D center positions are known at training time and predicted during inference using a convolutional deep neural network. The exact definition of the center of a human should be coherent with the point used to define the origin of the canonical coordinate systems. To supervise this network, we can use a similar dataset as in the remaining pipeline. Knowing the 3D center position, we can define a canonical 3D local coordinate system, perform the inference in that space

and replace the result in world coordinates. Note that the height of the subject is preserved as we do not apply any normalization on the size of the meshes during training.

3.4. Attention-based Fusion Layer

Image-based reconstruction benefits from multi-view cues, *e.g.*, stereo vision, which should be combined before the reconstruction is carried out in order to avoid premature single-view decisions and therefore limit ambiguities. Each view provides a feature and the question is how to aggregate them. Concatenating all the features, while simple, does not appear optimal because the fused features may become large when many images are considered, making it impossible to learn from an arbitrary numbers of views. Concatenation also imposes an order between views, which is undesirable in practice.

Besides concatenation, fusion approaches based on statistics, such as sum-pooling [13], average pooling [18] or max pooling [53] were proposed in the literature. The advantages are simplicity and invariance to both the order and the number of views. However, pooling operation loses information about individual view contributions. In particular, views in which a point is visible are considered equal to views in which the point is occluded and, more generally, erroneous information from an input view will contaminate the final prediction.

We propose to go one step further by learning the fusion and contextualising the information from different views. Previous work [60] proposes a simple learned fusion layer that computes a normalized score for each view, for each channel of a global feature. The main limitation is that the score of each view is computed individually without taking into account the information from the other views.

Inspired by recent progress in natural language processing to learn from sequences, we propose an architecture based on the transformer network [56] which implements a multi-head self-attention mechanism and is described in Fig. 4. One key component is the *scaled dot-product attention* which is a mapping function from a query along with a key / value pair to an output. The three vectors query $Q = M^q X$, key $K = M^k X$, and value $V = M^v X$ are the embedding of the original feature X parameterized by matrices M^q, M^k and M^v, respectively. The idea is to compute an attention score for each view based on a compatibility of a query with a corresponding key:

$$\text{Attention}(Q, K, V) = \mathbf{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (2)$$

where d_k the common dimension of K, Q and V.

To allow the network to attend to different geometric patterns, we propose to use multiple heads. For that, Q, K and V are linearly projected h times and processed in parallel through a scaled dot-product attention layer. The results

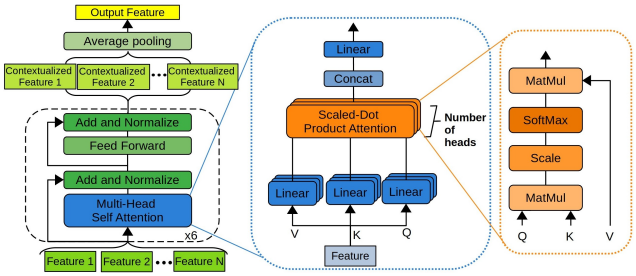

Figure 4. (Left) Our view fusion module. (Middle) Multi-Head Attention module. (Right) Scaled Dot-Product Attention.

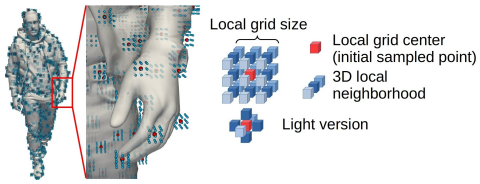

Figure 5. A local 3D grid is constructed around each sampled point (in red), and parameterized by a size and an orientation.

from the different heads are concatenated and finally projected once again to obtain the final output :

$$\begin{aligned} &\text{MultiHead}(Q, K, V) = \mathbf{concat}(H_1, ..., H_h)W^o \\ &\text{with } H_i = \text{Attention}(QW_i^q, KW_i^k, VW_i^v) \end{aligned} \quad (3)$$

where W_i^q, W_i^k, W_i^v are respectively the parameters of the linear mapping of Q, K and V, and W^o the parameters of the final projection.

The output of the attention modules is a set of features. Each of them contains the original information from the corresponding view that now takes into account the information from all the other available views. Finally we use the mean of these features as output of our view fusion module. Note also, that we do not use any positional encoding on the input feature sequence to remain invariant to the view order.

3.5. Local 3D Context Encoding

In the proposed framework, projection is used to associate 3D points with 2D image features for each available view. Then, the attention-based fusion layer weighs the contribution of each view in the fused feature. Finally, a Multi-Layer Perceptron (MLP) predicts an occupancy probability. However, such features do not take the 3D geometric context into consideration since the neighbourhood is only considered in 2D when features are extracted from the images. To include 3D context, we propose to build a local 3D grid around each sampled point and associate each point of the local 3D grid with 2D image features by projection.

The attention-based layer is applied individually on each point of the local grids, after which we add another context fusion module that combines the information coming from a 3D neighbourhood of a sampled point. This module is shown in orange in Fig. 3, and is implemented with a fully connected layer. Thanks to this additional layer, the neural

network is aware of the local 3D context of a point. In this way, we expect the network to better capture 3D geometric patterns and to increase robustness against nuisance factors (*e.g.*, texture, lighting).

As shown in Fig. 5, the local grid is parameterized by the size S and orientation R. Empirically, we found that fixing R during training strongly links the local grid to the global coordinate system and the orientation of the human body. To remain invariant to the orientation of the human, during training we randomly align R with one of the available views at each iteration.

The grid size S needs to be chosen based on the training data and the type of the targeted 3D patterns. Our goal is to learn local 3D patterns that typically contain points in the same or close-by body parts. As a full local grid can be expensive in computation time and memory, we propose a variant that uses only the cells along the three grid axes that traverse the center of the grid. In that case, three one-dimensional vectors are considered instead of one three-dimensional grid, which significantly decreases the number of grid points while still allowing to take into account local context along three directions. We call this version "light" and use it in all our experiments. We also provide an ablation study on the grid size in supplementary materials.

4. Experimental Results

In this section, we evaluate the proposed method and compare it with the state of the art. First, we give implementation details and introduce the training and testing datasets as well as the evaluation metrics. We then compare our approach quantitatively and qualitatively against the current state of the art and provide an ablation study to justify our contributions. Finally we show results of spatially consistent reconstruction and applications on real multi-view stereo data. Please refer to the supplementary materials for additional visual results and comparisons.

4.1. Implementation Details

Our human center localization network is implemented with the standard VGG16 [48] architecture. The image encoder or our reconstruction network is a Stacked Hourglass Network, with intermediate supervision, composed of 4 hourglass modules each of depth 2. The size of the output features is $128 \times 128 \times 256$. Since we trained the network with a small batch size, we also introduced group normalization instead of batch normalization. Our view fusion layer is composed of 6 modules based on multi-head self-attention with 6 heads. The local 3D context fusion maps features from a $3 \times 3 \times 3$ grid into a single feature of size 256. The Multi Layer Perceptron (MLP) is composed of 6 layers of dimensions $256, 1024, 512, 256, 128, 1$ with skip connections between the first layer and all the other layers except the last one. We optimized our network during 100 epochs using the root mean square propagation algorithm with a learning rate of 1×10^{-4} that is divided by 10 at iterations 60 and 80. More details are available in the supplementary materials.

Methods	CD (cm) ↓		Occ L1 ↓		Norm Cosine ↓		Norm L2 ↓	
	mean	median	mean	median	mean	median	mean	median
PaMIR [66]	0.554	0.508	1.977	1.754	0.097	0.090	0.361	0.343
PIFu [45]	0.592	0.510	2.079	1.773	0.103	0.093	0.376	0.358
PIFuHD [46]	2.008	1.624	5.837	4.543	0.181	0.162	0.544	0.503
Ours	**0.367**	**0.316**	**1.538**	**1.323**	**0.089**	**0.083**	**0.350**	**0.337**

Table 1. Quantitative results and comparisons with PaMIR [66], PIFu [45] and PIFuHD [46] on Renderpeople dataset. PaMIR, PIFu and ours use 4 views as input (see Fig. 6) and PIFuHD uses a single frontal view. Best scores are in **bold.**

Variants	CD (cm) ↓		Occ L1 ↓		Norm Cosine ↓		Norm L2 ↓	
	mean	median	mean	median	mean	median	mean	median
w./o. fusion	0.553	0.478	2.013	1.755	0.101	0.093	0.373	0.353
w./o. context	0.413	0.363	1.622	1.399	0.091	0.087	0.353	0.342
Ours full	**0.367**	**0.316**	**1.538**	**1.323**	**0.089**	**0.083**	**0.350**	**0.337**

Table 2. Ablation studies on the effectiveness of different components. We evaluate our method when deactivating the view fusion module and the local 3d context encoding, respectively. Best scores are in **bold.**

Variants	CD (cm) ↓		Occ L1 ↓		Norm Cosine ↓		Norm L2 ↓	
	mean	median	mean	median	mean	median	mean	median
2 views	0.870	0.753	2.909	2.474	0.121	0.114	0.407	0.392
4 views	0.367	0.316	1.538	1.323	0.089	0.083	0.350	0.337
6 views	**0.279**	**0.245**	**1.383**	**1.215**	**0.082**	**0.079**	**0.337**	**0.327**

Table 3. Ablation studies on using different number of views as input. Best scores are in **bold.**

4.2. Settings

We create our synthetic dataset with Renderpeople [1], a public commercial dataset that provides highly detailed meshes obtained from 3D scans and corrected by artists. Its main advantage is the very high quality of the geometry which is essential to learn geometric details, especially with clothing. The humans from this set are in relatively standard poses and often hold accessories such as bags, cups or other objects. In total we have 1026 meshes, split into 800 meshes for training, 100 for validation and 126 for testing.

To evaluate quantitatively the reconstructed human meshes, we first compute the Chamfer Distance (**CD**) between the ground truth mesh and the reconstructed mesh. By considering average distances between meshes, this metric tends to measure the global quality of the reconstructions. To focus more on local details, we also consider surface normal of the reconstructed and ground truth meshes and compute the $\mathbb{L}_2$ and cosine distances between them (**Norm Cosine** and **Norm L2**, respectively). Finally, in order to evaluate accurately the raw predictions of our network before the Marching Cubes post-processing that transforms the occupancy probability grid into a mesh, we compute the average $\mathbb{L}_1$ distance ($\times 10^3$) between predicted and ground truth occupancy (**Occ L1**).

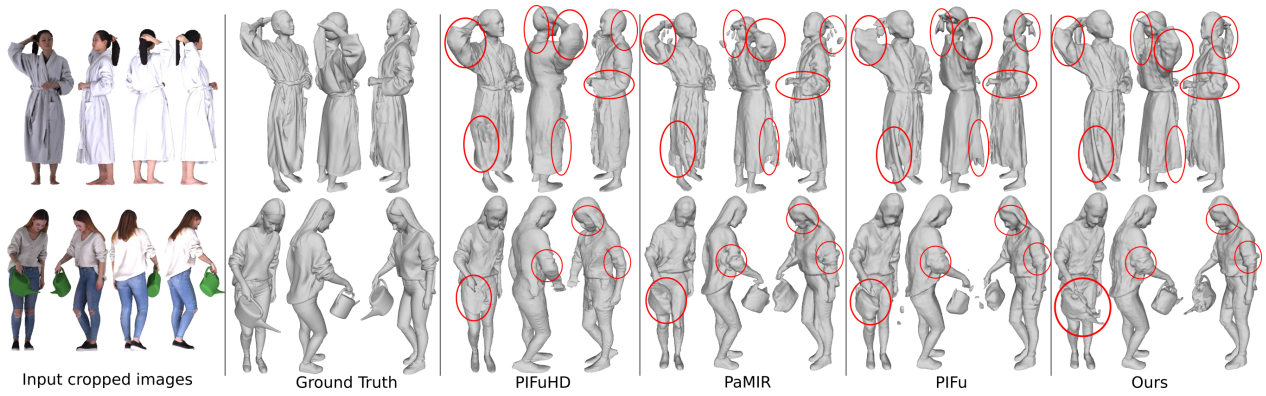

Figure 6. Qualitative results and comparisons with multi-view PIFu [45], multi-view PaMIR [66] and PIFuHD [46]. The 4 input images are rendered with the rotations around the vertical axis : 10°, 110°, 150°, 300°. PIFuHD uses a single frontal view as input.

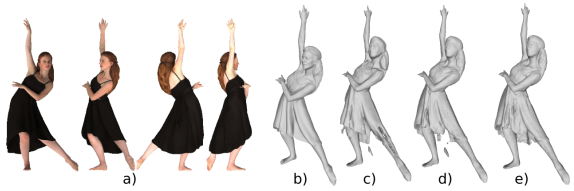

Figure 7. Ablation studies of our approach: a) Input cropped images. The 4 input images are rendered with the rotations around the vertical axis : 10°, 110°, 150°, 300°. b) Ground truth models. c) Ours without the attention-based view fusion module. d) Ours without the local 3D context encoding. e) Our full method.

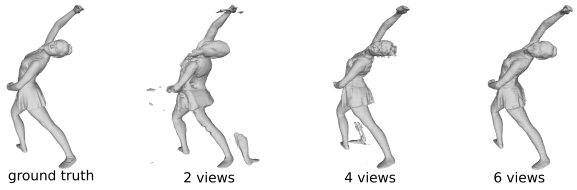

Figure 8. Ablation on different number of input views. As more views are added, the reconstruction with our method are improved.

4.3. Comparisons

In the context of 3D reconstruction of dressed humans from a few sparse views, PIFu [45] demonstrated state-of-the-art results so we consider it as the baseline result. For the comparison we trained it on our training dataset. This method has proven its benefit against model-based reconstructions and we do not provide comparisons with the latter. PIFuHD [46] extends PIFu to high resolution images and shows impressive single view reconstructions of details for the visible parts. No training code is available, so we use the published pre-trained model for the comparison. PAMIR [66] combines the implicit representation with a parametric body model and shows improved single-view and multi-view reconstructions. The released code and pre-trained model are only for single-view reconstruction, so we implemented the missing parts ourselves and trained a multi-view model on our training dataset. We do not provide direct comparisons between our method and multi-view stereo (MVS) methods applied on the exact same input data since MVS methods fail when only few images are available. PIFu, PIFuHD and PaMIR use orthographic images in which the human is at the center and cannot address the spatial consistency in world space. For a fair evaluation, we create a corresponding training / validation / test dataset composed of meshes from Renderpeople and evaluate all four methods on this data.

Qualitative results on synthetic data are presented in Fig. 6. PIFu and PaMIR achieve promising reconstructions but fail on some parts like the hair and the arm in the first row, or the watering can and clothing wrinkles in the second row. Our method appears clearly more robust and captures more geometric detail as can be seen on faces and clothing wrinkles. PIFuHD achieves detailed reconstructions for the visible parts like the face but, unlike for our method, the quality decreases significantly for the hidden parts and the global shape is not respected like the head on both rows. This is inherent to single view reconstruction methods and emphasizes the utility of using multiple views.

This intuition is verified by the associated quantitative results in Tab. 1 that confirm the benefit of our method on three aspects. First, the global quality of the reconstructions is improved by a large margin with the Chamfer distance. Second, metrics on surface normal are also in line and show that local geometric details are better captured. Third, our method achieves better results on the raw values of the implicit function.

4.4. Ablation Studies

To evaluate the impact of our contributions, namely the multi-head self-attention fusion layer and the local 3D context encoding, we conducted qualitative and quantitative ablation studies. To isolate these contributions from eventual human center detection errors, we place here the human person at the center of the scene. For the first contribution, we replaced the view fusion module by a simple average pooling strategy and for the second, individual sample points

were considered in place of the proposed local 3D grid.

Quantitatively, disabling the view fusion or the context encoding module both affect the reconstruction performance. From the results shown in Fig. 7 and Tab. 2, we clearly see that the multi-head self-attention view fusion module is crucial for both the global quality and the local geometric details. On the other hand, the local 3D context encoding is not sufficient by itself but when combined with the view fusion module helps the global reconstruction quality and avoids holes or missing parts.

To evaluate the scalability of our method, we compare reconstructions with different numbers of input views. Visual results in Fig. 8 show that the global quality of the shape (noise and missing parts) as well as geometric details (face and skirt) are improved as more views are used. Visual results are confirmed by the quantitative evaluation in Tab. 3. In particular, we observe a stronger improvement when using 4 views instead of 2 compared to 6 views instead of 4. This observation seems reasonable since the views used here are distributed evenly around the person and 4 views are sufficient to observe every side.

4.5. Spatially Consistent Reconstruction

To demonstrate the spatial consistency of the reconstructions we consider two scenarios, using data from Renderpeople [1]. First, we apply our method to dynamic input, namely to four synchronized video sequences showing a human walking in a scene. We reconstruct the sequence frame-by-frame, and Fig. 9(a) shows that the reconstructions contain details (ears, clothing wrinkles) and are spatially consistent with the ground truth. A better visualization is provided in the supplementary video.

As a second scenario, we consider a static scene containing multiple persons at different positions and render high resolution images with 4 cameras. Note that this evaluation focuses on spatially consistent reconstructions and not occlusions between persons. Hence, we render each person individually while the other persons are hidden. Fig. 9(b) shows that the reconstructions are spatially consistent with the ground truth and we can also note that the heights of the persons are correctly reconstructed.

4.6. Application to Real-world Data

To demonstrate the generalization of our method, we show 3D reconstructions of clothed humans with real images obtained with a 60 camera multi-view capture system. We compare with PIFu and PIFuHD when reconstructing with the front view only, to PIFu when reconstructing with 4 views, and to a multi-view stereo method [30] on the same scenes but with 60 images. For all methods to be applicable, we consider the person centered in the middle of the scene. It is important to note that the networks were trained purely on synthetic data while tested on images from a real acquisition scenario. Fig. 1 shows that single view reconstructions suffer from an inherent depth ambiguity: some parts are missing (hair and backpack) and the pose is incorrect. Our method performs better than PIFu when 4 views are available, with more realistic global shapes and more detailed local geometries. More importantly, the comparisons with the multi-view stereo method applied to 60 images demonstrate the potential of data-driven strategies in the multi-view reconstruction domain.

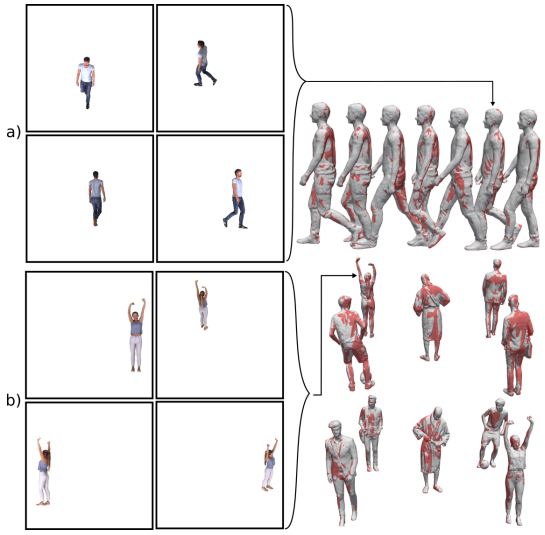

Figure 9. Spatially consistent reconstructions. a) Frame-by-frame reconstruction of a sequence from synchronized videos. Left: high resolution images for one example frame. Right: our result with the ground truth superimposed in red. b) Person-by-person reconstruction of a scene with multiple people. Left: high resolution images for one example person. Right: our result with the ground truth superimposed in red. a) and b) For both, the camera rotations around the vertical axis are 10°, 110°, 200°and 300°with an random elevation angle between 0°and 40°.

5. Conclusion

In this paper, we build on recent progress on implicit representations of 3D data and propose a method for 3D reconstruction of clothed humans from a few sparse views. We introduce three key components: 1) a spatially consistent reconstruction that allows for arbitrary placement of the person in the input views using a perspective camera mode; 2) a fusion layer based on an attention mechanism that learns to efficiently combine the information from all available views; 3) a mechanism that encodes local 3D patterns in the multi-view context. Our experiments show that the proposed method outperforms the state of the art in terms of details and global quality of the reconstructions on synthetic data. We also demonstrate a better generalization of our method on real data acquired with a multi-view platform. Additionally, we show that our approach can approximate multi-view stereo results with dramatically less views.

6. Acknowledgements

We thank Laurence Boissieux and Julien Pansiot from the Kinovis platform at Inria Grenoble and our volunteer subjects for help with the 3D data acquisition.

References

[1] Renderpeople, 2018. https://renderpeople.com/3d-people/. 4326, 4328, 4332, 4334

[2] Thiemo Alldieck, Marcus A. Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 4322

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. *ACM Transactions on Graphics*, 24(3):408–416, 2005. 4322

[4] Alexandru O. Balan and Michael J. Black. The naked truth: Estimating body shape under clothing. In *European Conference on Computer Vision*, pages 15–29, 2008. 4322

[5] Alexandru O. Balan, Leonid Sigal, Michael J. Black, James E. Davis, and Horst W. Haussecker. Detailed human shape and pose from images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 4323

[6] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision*, pages 311–329, 2020. 4323

[7] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE/CVF International Conference on Computer Vision*, pages 5419–5429, 2019. 4322

[8] Federica Bogo, Michael J. Black, Matthew Loper, and Javier Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *IEEE/CVF International Conference on Computer Vision*, pages 2300–2308, 2015. 4322

[9] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578, 2016. 4322

[10] Edilson de Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. *ACM Transactions on Graphics*, 27(3):98, 2008. 4323

[11] Endri Dibra, Himanshu Jain, A. Cengiz Öztireli, Remo Ziegler, and Markus H. Gross. Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In *International Conference on 3D Vision*, pages 108–117, 2016. 4323

[12] Simon Donné and Andreas Geiger. Learning non-volumetric depth fusion using successive reprojections. In *IEEE IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7634–7643. Computer Vision Foundation / IEEE, 2019. 4323

[13] S M Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, David P Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz, Helen King, Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018. 4325

[14] Jean-Sébastien Franco and Edmond Boyer. Fusion of multiview silhouette cues using a space occupancy grid. In *IEEE/CVF International Conference on Computer Vision*, pages 1747–1753, 2005. 4323

[15] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multi-view stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010. 4321, 4323

[16] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Grégory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *IEEE/CVF International Conference on Computer Vision*, pages 2232–2241, 2019. 4322

[17] Juergen Gall, Carsten Stoll, Edilson de Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1746–1753, 2009. 4323

[18] Andrew Gardner, Jinko Kanno, Christian A. Duncan, and Rastko R. Selmic. Classifying unordered feature sets with convolutional deep averaging networks. *CoRR*, abs/1709.03019, 2017. 4325

[19] Peng Guan, Alexander Weiss, Alexandru O. Balan, and Michael J. Black. Estimating human shape and pose from a single image. In *IEEE/CVF International Conference on Computer Vision*, pages 1381–1388, 2009. 4322

[20] Nils Hasler, Hanno Ackermann, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1823–1830, 2010. 4322

[21] Tong He, John P. Collomosse, Hailin Jin, and Stefano Soatto. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Annual Conference on Neural Information Processing Systems*, pages 9276–9287, 2020. 4322

[22] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. Deep volumetric video from very sparse multi-view performance capture. In *European Conference on Computer Vision*, pages 351–369, 2018. 4322, 4323

[23] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: animatable reconstruction of clothed humans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3090–3099, 2020. 4322

[24] Aaron S. Jackson, Chris Manafas, and Georgios Tzimiropoulos. 3d human body reconstruction from a single image via

volumetric regression. In *ECCV Workshop*, pages 64–77, 2018. 4322

[25] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2307–2315, 2017. 4321

[26] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 4323

[27] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 4322

[28] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019. 4322

[29] Kiriakos N. Kutulakos and Steven M. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3):199–218, 2000. 4323

[30] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Shape reconstruction using volume sweeping and learned photoconsistency. In *European Conference on Computer Vision*, pages 796–811, 2018. 4321, 4323, 4328, 4333, 4336

[31] Zhenbao Liu, Hongliang Qin, Shuhui Bu, Meng Yan, Jinxin Huang, Xiaojun Tang, and Junwei Han. 3d real human reconstruction via multiple low-cost depth cameras. *Signal Processing*, 112:162–179, 2015. 4323

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015. 4321

[33] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *SIGGRAPH*, pages 163–169, 1987. 4323

[34] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421, 2020. 4323

[35] Armin Mustafa, Akin Caliskan, Lourdes Agapito, and Adrian Hilton. Multi-person implicit reconstruction from a single image. *CoRR*, abs/2104.09283, 2021. 4324

[36] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4480–4490, 2019. 4322

[37] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499, 2016. 4334, 4339

[38] Hayato Onizuka, Zehra Hayirci, Diego Thomas, Akihiro Sugimoto, Hideaki Uchiyama, and Rin-Ichiro Taniguchi. Tetratsdf: 3d human reconstruction from a single image with a tetrahedral outer shell. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6010–6019, 2020. 4323

[39] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7363–7373, 2020. 4322

[40] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 4322

[41] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018. 4322

[42] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page to appear, 2021. 4323

[43] Ralf Plänkers and Pascal Fua. Tracking and modeling people in video sequences. *Computer Vision and Image Understanding*, 81(3):285–302, 2001. 4322

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015. 4334, 4339

[45] Shunsuke Saito, , Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 4321, 4322, 4323, 4326, 4327, 4332, 4333, 4335, 4336, 4338

[46] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 81–90, 2020. 4321, 4322, 4323, 4326, 4327, 4332, 4333, 4335, 4336

[47] Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 519–528, 2006. 4321, 4323

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 4326, 4332, 4333

[49] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Conference on Neural Information Processing Systems*, pages 7462–7473, 2020. 4323

[50] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Conference on Neural Information Processing Systems*, pages 1119–1130, 2019. 4323

[51] Cristian Sminchisescu and Bill Triggs. Estimating articulated human motion with covariance scaled sampling. *International Journal of Robotics Research*, 22(6):371–392, 2003. 4322

[52] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. FACSIMILE: fast and accurate scans from an image in less than a second. In *IEEE/CVF International Conference on Computer Vision*, pages 5329–5338, 2019. 4322

[53] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 945–953, 2015. 4325

[54] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *CoRR*, abs/1902.09212, 2019. 4334, 4339

[55] Gül Varol, Duygu Ceylan, Bryan C. Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *European Conference on Computer Vision*, pages 20–38, 2018. 4322

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 4322, 4325

[57] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popovic. Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics*, 27(3):97, 2008. 4323

[58] Alexander Weiss, David Hirshberg, and Michael J. Black. Home 3d body scans from noisy image and range data. In *IEEE/CVF International Conference on Computer Vision*, 2011. 4323

[59] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *IEEE/CVF International Conference on Computer Vision*, pages 7759–7769, 2019. 4322

[60] Bo Yang, Sen Wang, Andrew Markham, and Niki Trigoni. Robust attentional aggregation of deep feature sets for multi-view 3d reconstruction. *International Journal on Computer Vision*, 128(1):53–73, 2020. 4325

[61] Mao Ye, Yang Shen, Chao Du, Zhigeng Pan, and Ruigang Yang. Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1517–1532, 2016. 4322

[62] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page to appear, 2021. 4323

[63] Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *IEEE/CVF International Conference on Computer Vision*, pages 910–919, 2017. 4322

[64] Tao Yu, Jianhui Zhao, Zerong Zheng, Kaiwen Guo, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2523–2539, 2020. 4322

[65] Tao Yu, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Qionghai Dai, Gerard Pons-Moll, and Yebin Liu. Simulcap : Single-view human performance capture with cloth simulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5504–5514, 2019. 4322

[66] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transsactions on Pattern Analysis and Machine Intelligence*, to appear, 2021. 4322, 4326, 4327

[67] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *IEEE/CVF International Conference on Computer Vision*, pages 7738–7748, 2019. 4322

[68] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4491–4500, 2019. 4322

Supplementary Materials

In the following, we first provide more details about our implementation. We give then additional qualitative results and ablation tests. Note that a supplementary video is also provided to better visualize the reconstruction results.

7. Implementation

7.1. Training Views

To train our deep neural network, we created a synthetic model view set by rendering 3D models from Renderpeople [1] using 360 cameras located around them as explained below. In contrast to Multi-View Stereo methods, only a few of these views are considered at inference (between 2 and 6 in our experiments). The views used at inference should be ideally evenly distributed around the person in order to increase its visibility. At inference, results are most of the time better for parts of the surface that are observed than hidden ones for which the reconstruction relies solely on the prior learned from the training set. To build such image sets for the training we sample the synthetic views of a 3D model and create several model view subsets with few images.

To define the position of our cameras when creating such a subset, we use a rotation angle around the up-axis and an elevation angle, as described in Figure 10. For the orientation, we assume that the cameras are always looking at the center of the scene.

In practice, at each training iteration we choose N angles around the up axis that are evenly distributed among $[0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°]$ and add a random offset between $-20°$ and $20°$. The elevation angles are selected randomly between $0°$ and $45°$. Note that we trained our model with a fixed elevation angle when comparing with other methods (*i.e.* PIFu [45] and PIFuHD[46]) that consider a similar scenario.

7.2. Human Center Localization

In Figure 11 we show the architecture of our deep neural network based on VGG16 [48] to detect the human center on each of the view. These 2D detections are then used to triangulate the 3D position of the person in the scene. The center of the person is arbitrarily defined but should be coherent with the origin of the canonical coordinate systems used at training. In practice, we defined it as :

$$\begin{bmatrix} median(vertices.x) \\ 0.5 * (max(vertices.y) - min(vertices.y)) \\ median(vertices.z) \end{bmatrix}$$

where y is the up-axis. We do not use the median for the up-axis to account for cases where numerous vertices are grouped at the top or the bottom. Such cases are worth considering since a human is less symmetric with respect to the horizontal plane. In Table 4, we compute the L_2 distance between the 2D detections and 2D ground truth as well as the 3D positions triangulated from the 2D detection and the 3D ground truth. Here we used 4 views evenly distributed around the person with a random elevation axis between $0°$ and $45°$.

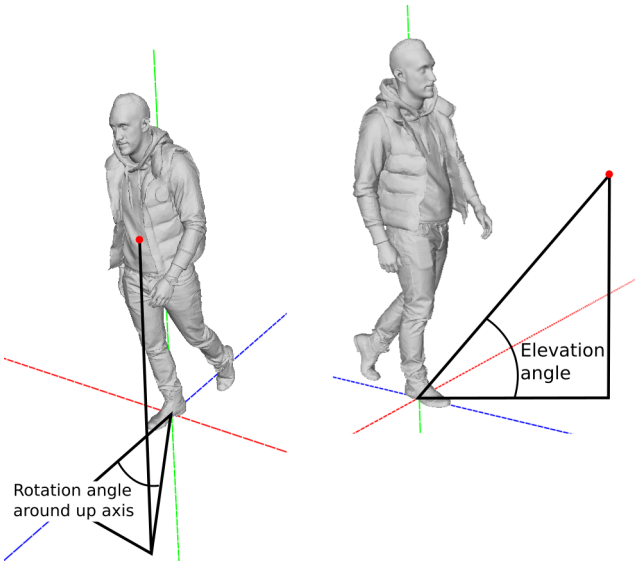

Figure 10. View selection angles.

L2 - 2D (pixels)		L2 - 3D (cm)	
mean	median	mean	median
9.795	8.944	4.398	4.291

Table 4. Evaluation of the human center detection on images and the 3D triangulated position of the center. Both are evaluated on test images.

As shown in Table 4, the average Euclidean distance between the ground truth and triangulated 3D human center is around $4.4cm$. We compute these metrics on test data (360 groups of 4 views for 50 persons) and follows the strategy explained in Section 7.1 to select the 4 views. Additionally, we show in Figure 12 an example of reconstruction with manually specified errors on the human center location. We see that the reconstruction quality is not affected too much up to 5cm. Noise starts being visible with an error of 10 cm and the reconstruction fails with larger error like 20 cm.

8. Attention scores

We provide in Figure 13 a visualization of the attention scores of our view fusion module. We use 4 input views, evaluate our deep neural network in a 3D grid of resolution 256 and save the attention score of the first self-attention layer. Note that we use a single head for this experiment. Points that are predicted close to the surface inside or outside are visualized and the intensity of the red channel rep-

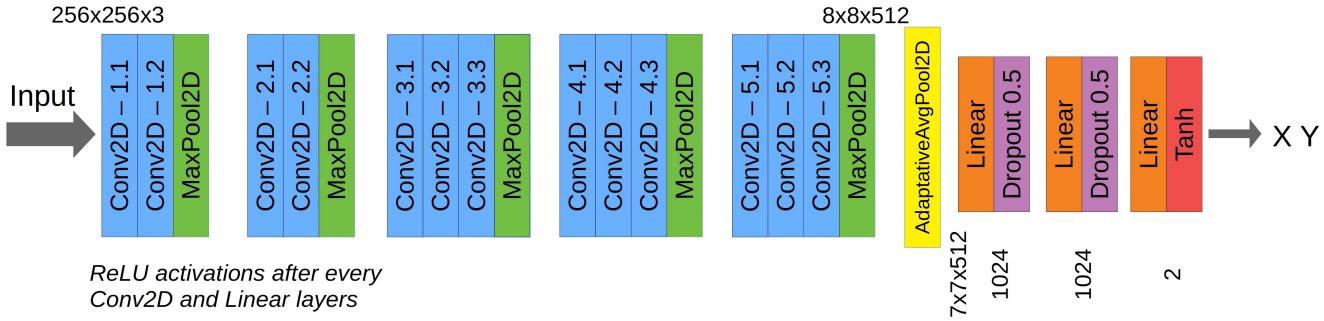

Figure 11. Human Center Detection network based on VGG16 [48].

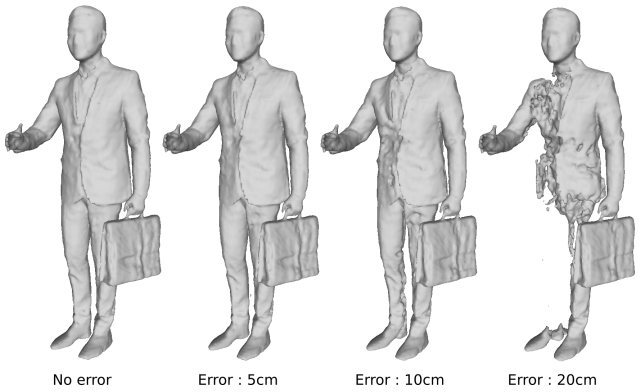

Figure 12. Reconstructions from 4 views that show the impact of a 3D human center localization error on the reconstruction.

resents how much the considered view contributed for each point. We clearly see that each point attend more to views in which they are visible.

9. Additional visual results

9.1. Comparison with the state of the art

Qualitative visual comparisons between our proposed method, the considered baseline [45] as well as state-of-the-art single-view reconstruction method PIFuHD [46] are presented in Figure 14. In particular, we note the improved global quality of the recovered accessories and the reduced level of noise in the reconstructions using our method. Moreover, sharper details on the faces and wrinkles on the clothes are recovered by our approach. Two difficult cases with less usual pose and thin structures are shown in the last two rows. Although our reconstructions contain some noise and missing parts, we can see a significant improvement compared to the other two methods.

9.2. Application to real-world data

A crucial aspect of our work is the applicability to real-world data. In Figure 15 we provide additional comparisons between our method, the considered baseline [45], state-of-the-art single-view reconstruction method [46], and a 60-view reconstruction obtained with a Multi-view stereo strategy [30]. In this real context, we observe that our method behaves better than the baseline and single-view reconstruction, especially with complex scenes, *e.g.* with accessories. The improvement is less obvious, yet there, with persons in standard poses and without accessories (*e.g.* columns 3 in Figure 15). In this case, the strategy from [45] already provides good results. Another interesting comparison in this figure is with multi-view stereo (row b). While the MVS strategy provides robust and accurate estimations of the global shapes, our data-driven strategy yields more local details.

Variants	CD (cm) ↓		Occ L1 ↓		Norm Cosine ↓		Norm L2 ↓	
	mean	median	mean	median	mean	median	mean	median
PIFu 2 v.	1.386	1.233	3.206	2.861	0.136	0.130	0.444	0.432
PIFu 4 v.	0.592	0.510	2.079	1.773	0.103	0.093	0.376	0.358
PIFu 6 v.	0.331	0.313	1.499	1.402	0.088	0.083	0.345	0.331
Ours 2 v.	0.870	0.753	2.909	2.474	0.121	0.114	0.407	0.392
Ours 4 v.	0.367	0.316	1.538	1.323	0.089	0.083	0.350	0.337
Ours 6 v.	**0.279**	**0.245**	**1.383**	**1.215**	**0.082**	**0.079**	**0.337**	**0.327**

Table 5. Ablation studies on using different number of views as input. Best scores are in **bold.**

9.3. Ablation visual results

Here, we show additional visual results of our ablation to evaluate the impact of our contributions. Quantitatively, disabling the view fusion or the context encoding module both affects the reconstruction performance. From the results shown in Fig. 16, we clearly see that the multi-head self-attention view fusion module is crucial for both the global quality and the local geometric details. On the other hand, the local 3D context encoding impacts more the global quality of the reconstruction and helps avoiding holes or missing parts.

9.4. Number of input views

Figure 17 shows results of our method with 2, 4 and 6 views as input. It demonstrates that adding views effectively decreases depth ambiguities and occlusions with a clear improvement in the reconstructions. It also shows the superiority of our proposed method compared to the baseline.

10. Additional ablations

10.1. Encoders

In our work, 2D features are extracted using the Stacked Hourglass encoder [37] that stacks multiple pooling and upsampling networks. It allows the extraction of information at multiple scales and accounts therefore for both local and global contexts. Intermediate supervision is also applied to the output of each module while training our network. Of course numerous alternative encoders exist and could be used in our architecture in place of the Stacked Hourglass encoder. We provide in this section a comparison with 2 popular options: U-Net [44] and HRNet [54]. Results are shown in Figure 18 and in Table 6. The U-Net [44], a fully convolutional network based on a contractive and an expansive part, gives results which are visually close to those obtained with the Stacked Hourglass encoder, with however significantly more noise as confirmed by the metrics in Table 6. On the other hand, the more recent work HRNet [54] fails to provide similar results in this context.

Encoders	CD (cm) ↓		OCC L1 ↓		Norm Cosine ↓		Norm L2 ↓	
	mean	median	mean	median	mean	median	mean	median
SHG	**0.385**	**0.322**	**1.602**	**1.380**	**0.087**	**0.081**	**0.343**	**0.326**
U-Net [44]	0.572	0.482	1.984	1.688	0.108	0.101	0.389	0.369
HRNet [54]	1.092	1.075	3.682	3.547	0.181	0.178	0.565	0.553

Table 6. Quantitative results obtained by our approach, on Renderpeople data [1], with 3 different image encoders (see text in Sec. 10.1 for comments). Best scores are in bold.

10.2. Local grid size

A key point of our method is the encoding of the local context of each sampled 3D point. To this purpose, we use a local 3D grid around each sampled point and in the pipeline, each original sampled point is associated with the additional points from their 3D local neighboorhood. At each training iteration, the local grids are aligned randomly with one of the camera used and the grid size is constant and defined before training.

Here we provide the results obtained with different grid sizes defined in world coordinates: small (2 cm), medium (10 cm) and large (20 cm) grids.

Table 7 shows that the best results were obtained with the medium-sized local grid, which is the one that was used for the other results in this paper. This result is confirmed visually on Figure 19, where the medium grid shows better reconstructions with more details and less noise. This experiment demonstrates that the size of the local grid is important as it defines the neighborhood considered to predict the occupancy probability of the grid center. With a small grid, all grid points tend to be projected on the same 2D feature which prevents the 3D context to be encoded. On the other hand, with large grids, points can be far from each other, even on different body parts. In that case, the neighborhood considered is too large and not informative when predicting occupancies.

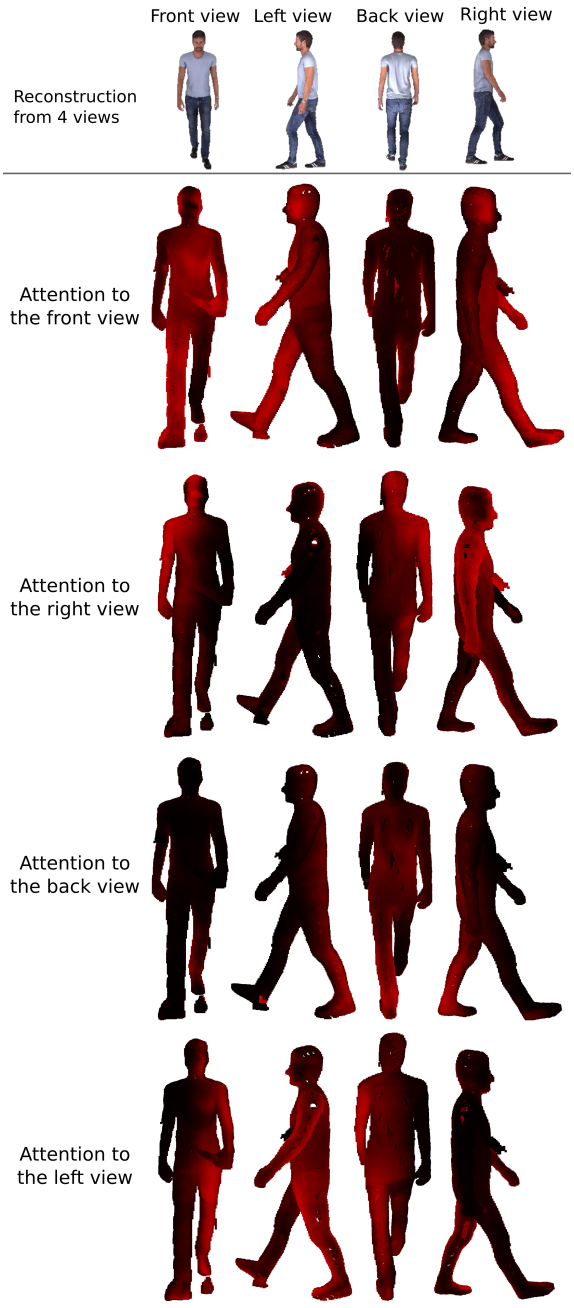

Figure 13. Attention scores: points predicted as close to the surface (in or out) are visualized. The intensity of the red channel represents the contribution of the considered view.

Grid size	CD (cm) ↓		OCC L1 ↓		Norm Cosine ↓		Norm L2 ↓	
	mean	median	mean	median	mean	median	mean	median
small (2 cm)	0.422	0.413	1.668	1.566	0.089	0.087	**0.342**	0.336
medium (10 cm)	**0.385**	**0.322**	**1.602**	**1.380**	**0.087**	**0.081**	0.343	**0.326**
large (20 cm)	0.441	0.421	1.677	1.592	0.091	0.089	0.350	0.341

Table 7. Quantitative results and comparisons with 3 local grid sizes on Renderpeople data [1]. Best scores are in bold.

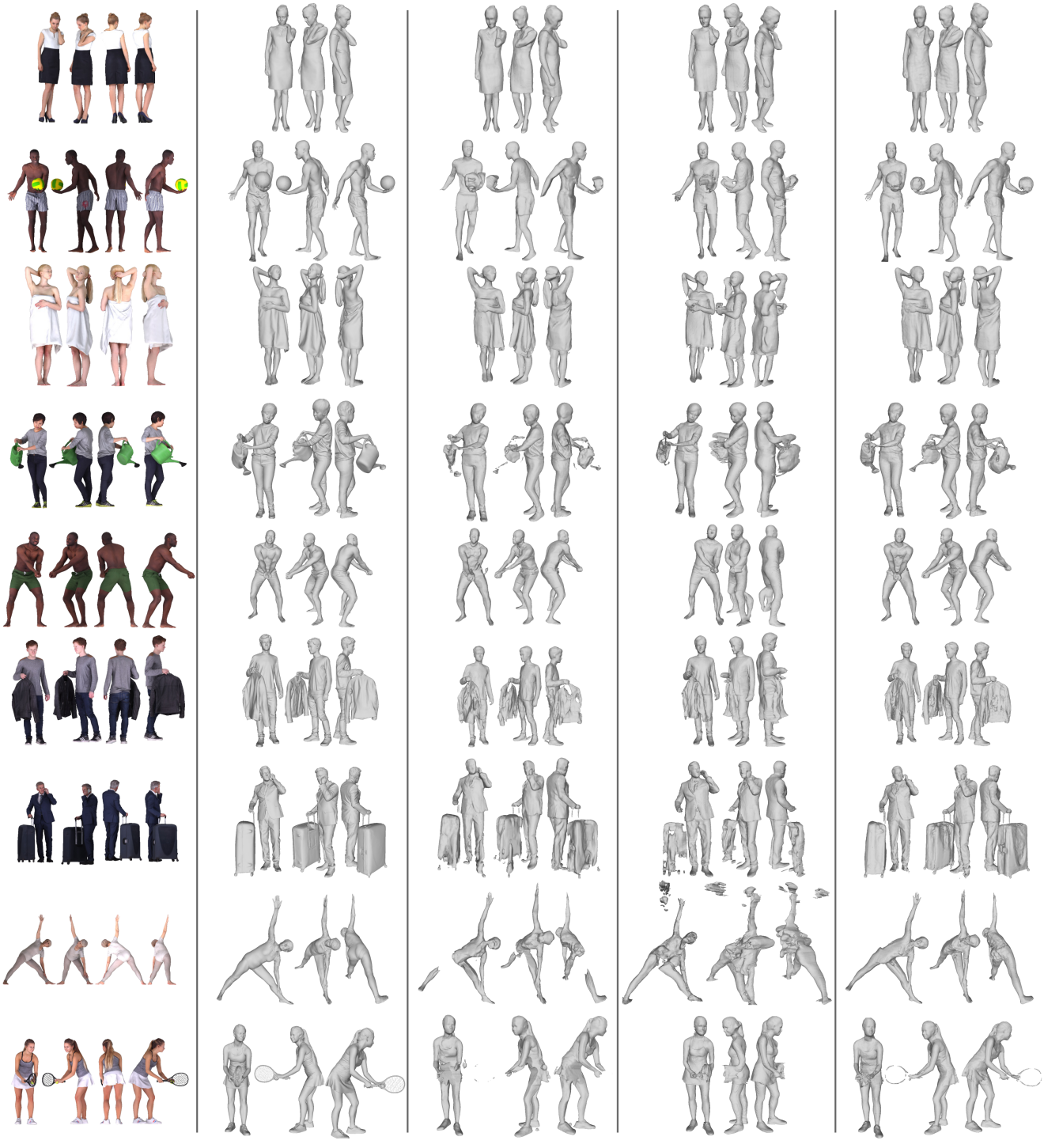

Figure 14. Qualitative results and comparisons. PIFu [45] and our method take as input the 4 cropped images, whereas PIFuHD [46] receives only the frontal view. The 4 input images are rendered with the rotations around the vertical axis : 10°, 110°, 150°, 300°.

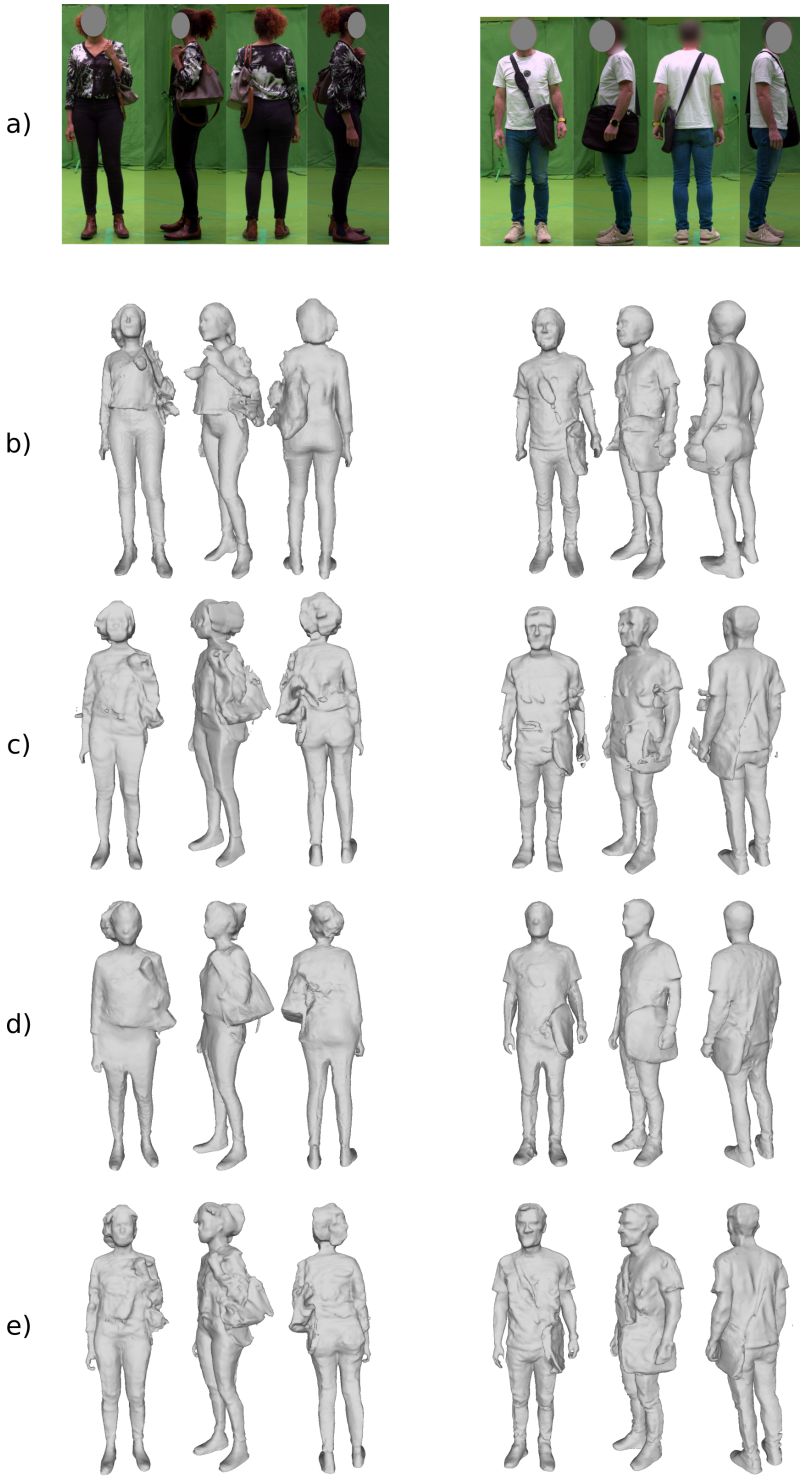

Figure 15. Qualitative results and comparisons with a real capture apparatus: a) real RGB images. b) single frontal view reconstruction using PIFuHD [46]. c) 4-view reconstructions using PIFu [45]. d) 60-view reconstructions using a multi-view stereo approach [30]. e) 4-view reconstructions using our method.

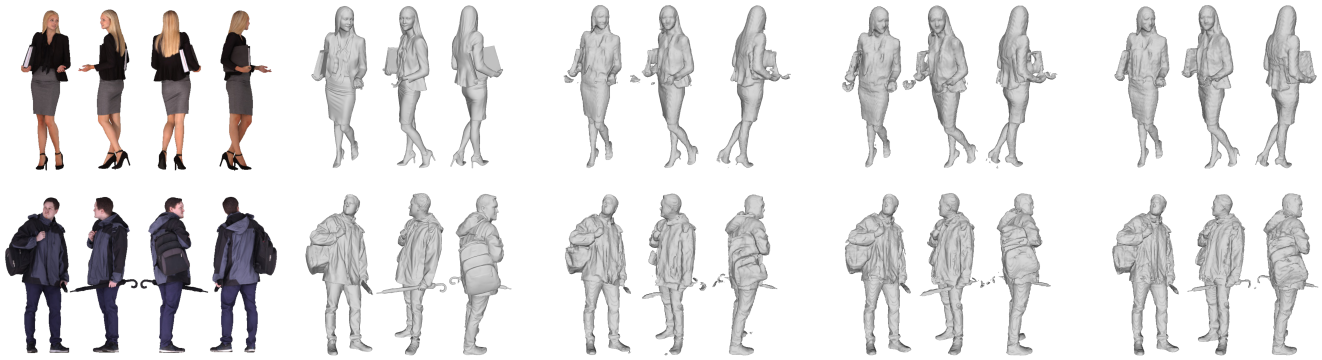

Figure 16. Ablation studies of our approach: a) Input cropped images. The 4 input images are rendered with the rotations around the vertical axis : 0°, 90°, 180°, 270°. b) Ground truth models. c) Ours without the attention-based view fusion module. d) Ours without the local 3D context encoding. e) Our full method.

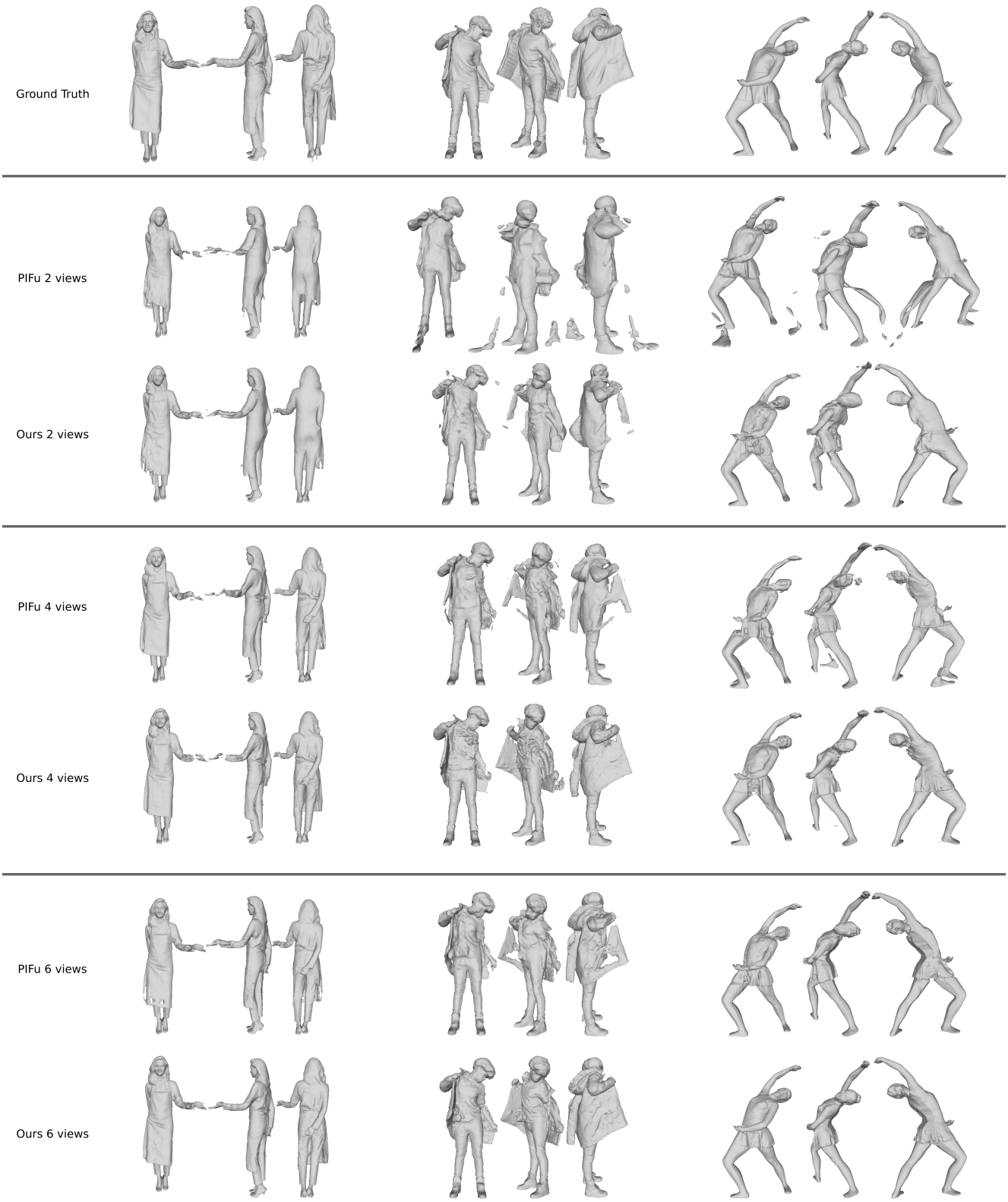

Figure 17. Impact of the number of input views on the reconstruction quality and comparison with PIFu [45].

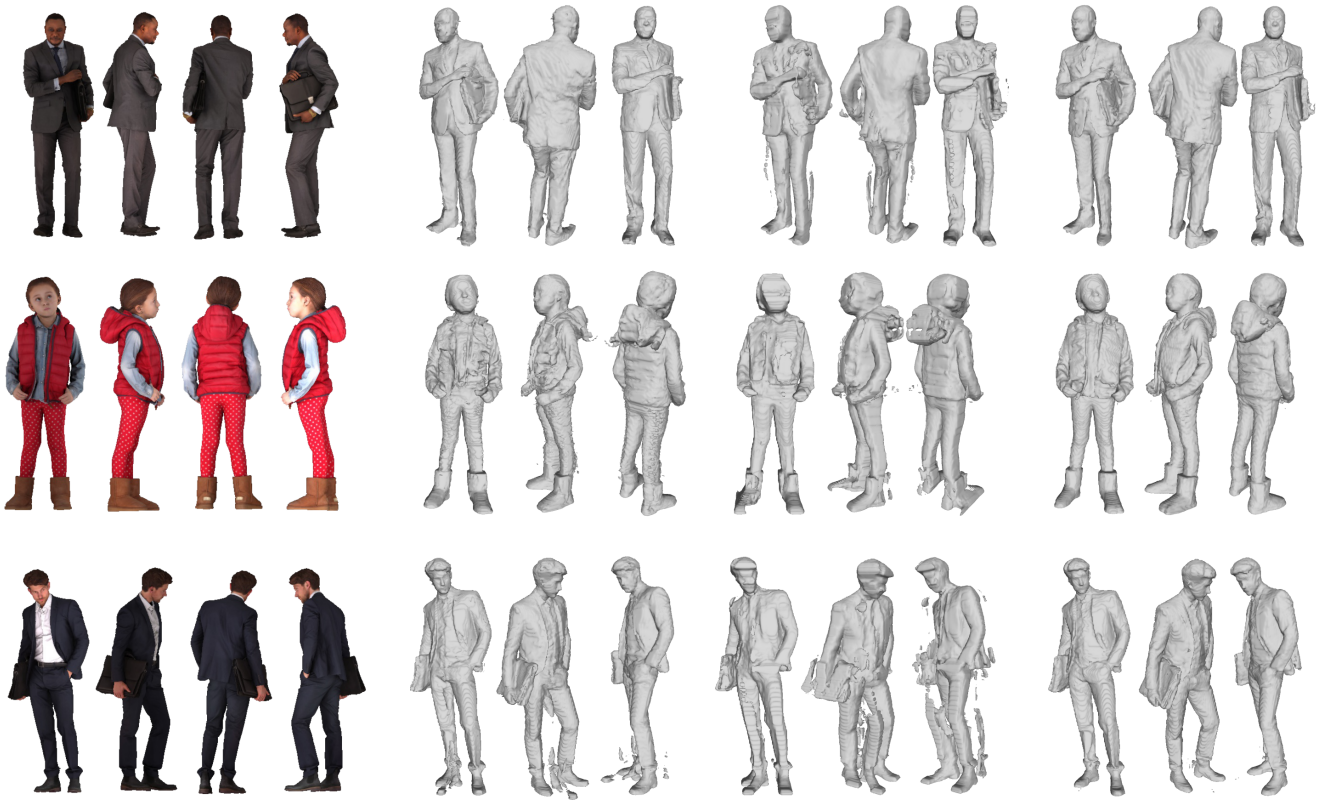

Figure 18. Comparative reconstruction results with our approach applied using 3 different image encoders. a) Input RGB images. b) U-Net [44]. c) HRNet [54]. d) Stacked Hourglass [37].

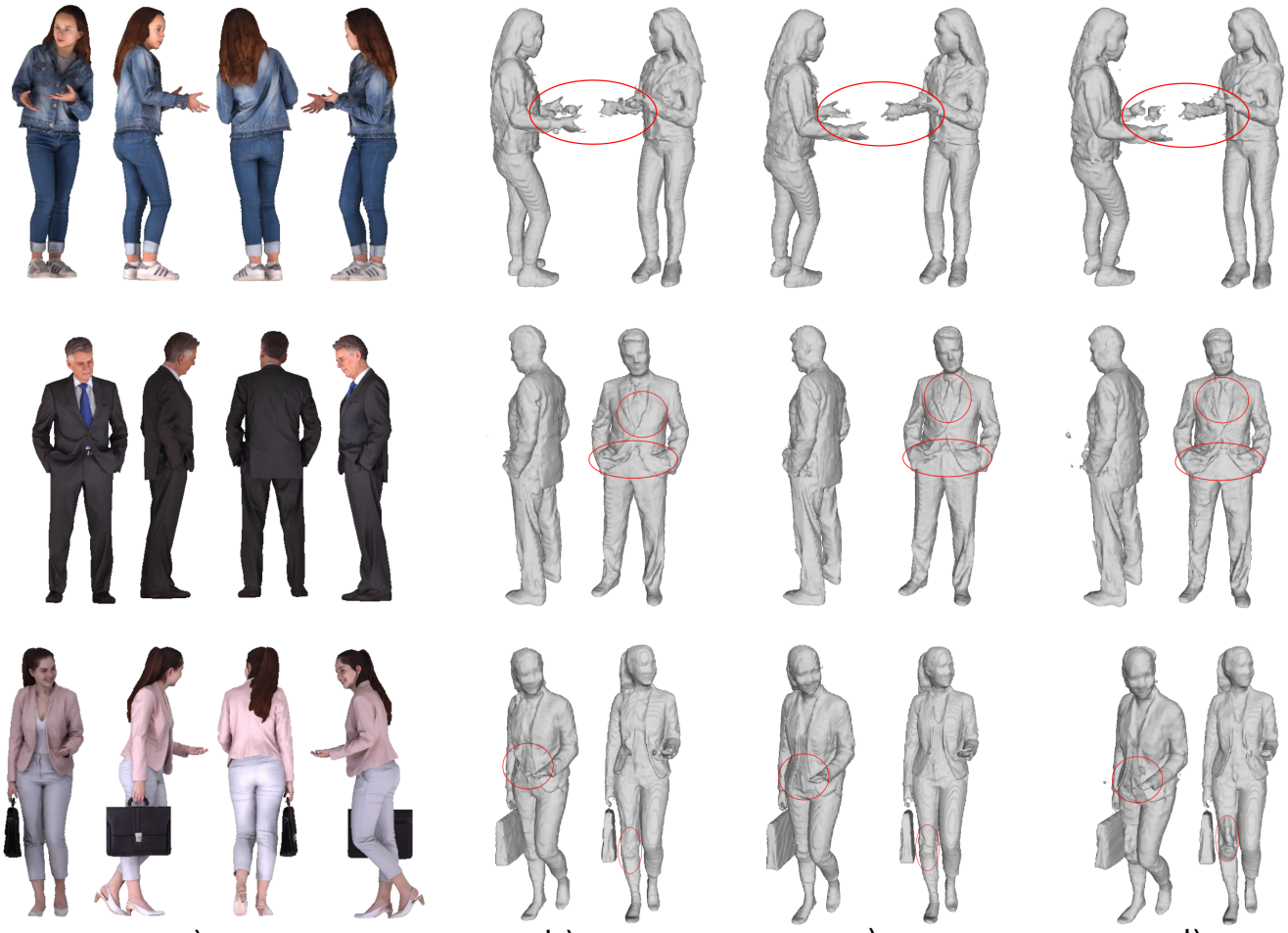

Figure 19. Qualitative comparison of the reconstructions using our method with 3 different local grid sizes. a) Input RGB images. b) Small grid (2 cm). c) Medium grid (10 cm). d) Large grid (20cm).