

Super-resolution 3D Human Shape from a Single Low-Resolution Image

Marco Pesavento¹, Marco Volino¹, and Adrian Hilton¹

Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey,
Guildford, UK {m.pesavento,m.volino,a.hilton}@surrey.ac.uk

Abstract. We propose a novel framework to reconstruct super-resolution human shape from a single low-resolution input image. The approach overcomes limitations of existing approaches that reconstruct 3D human shape from a single image, which require high-resolution images together with auxiliary data such as surface normal or a parametric model to reconstruct high-detail shape. The proposed framework represents the reconstructed shape with a high-detail implicit function. Analogous to the objective of 2D image super-resolution, the approach learns the mapping from a low-resolution shape to its high-resolution counterpart and it is applied to reconstruct 3D shape detail from low-resolution images. The approach is trained end-to-end employing a novel loss function which estimates the information lost between a low and high-resolution representation of the same 3D surface shape. Evaluation for single image reconstruction of clothed people demonstrates that our method achieves high-detail surface reconstruction from low-resolution images without auxiliary data. Extensive experiments show that the proposed approach can estimate super-resolution human geometries with a significantly higher level of detail than that obtained with previous approaches when applied to low-resolution images. <https://marcopesavento.github.io/SuRS/>

1 Introduction

The demand for user-friendly frameworks to create virtual 3D content is increasing at the same pace as the rise in the application of immersive technologies. This has led research to a focus on finding practical solutions to replace existing sophisticated multi-view systems for 3D reconstruction, which are impractical and expensive to build for the general community. With the recent advance of 3D deep learning, several approaches have been proposed to estimate the 3D shape directly from a single RGB image for specific object classes such as human bodies, faces or man made objects. Although consumer cameras are nowadays able to capture high-resolution (HR) images, there are several scenarios where the image of a single person is not at the full camera image resolution but at a relatively low-resolution (LR) sub-image. Reconstruction of high quality shape from LR images of people is therefore important for example for images of multiple people, scenes requiring a large capture volume or when people are distant from the camera (Fig. 1). Since the LR image contains little detail information,



Fig. 1. Person capture in a large space in a HR image [22]. Our approach is able to represent significant higher level of fine details compared to PIFuHD [34] even if these are not clear in the LR input image.

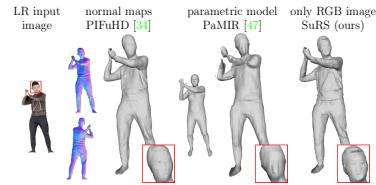


Fig. 2. 3D human digitization of a low-resolution (256×256) image. Our approach result represents significant higher level of details compared to related approaches that leverage auxiliary data.

state-of-the-art approaches cannot represent fine details in the reconstructed shape. The introduction of pixel-aligned implicit function (PIFu [33]) improved the quality of human shapes reconstructed from a single image. However, it is highly dependent in the level of details represented on the input image and it fails to reconstruct HR shape from a LR image that does not contain information on fine details. Recent works leverage auxiliary information such as displacement or normal maps together with HR images (1024×1024) to retrieve fine details [34,4]. Other methods combine the implicit function with 3D information by using voxel alignment and 3D convolution [11,47,20]. However, none of these works can reconstruct high quality shapes by only using a single LR RGB image.

To tackle this problem, we introduce a new framework that generates Super-Resolution Shape (SuRS) via a high-detail implicit function which learns the mapping from low to high-resolution shape. If a low-resolution shape S_{LR} is compared with its high-resolution counterpart S_{HR} , the primary difference is in the fine details that represent the information lost between S_{LR} and S_{HR} . This information is retrieved by learning the mapping from low to high-resolution shape such that corresponding fine details can be reproduced in the super-resolution shape S_{SR} . We apply SuRS to the task of 3D human digitization to estimate high-detail 3D human shape from a single LR RGB image (256×256) without assisting the training with auxiliary data such as normal maps or parametric models. Our approach learns the missing information of S_{LR} in order to estimate fine shape details from a LR image even if these details are not clearly visible in the input image. The final high-detail shape is represented with an implicit representation, which is learned by a novel end-to-end trainable framework with a new loss that estimates the information missing in the low-resolution shape S_{LR} by computing the difference between the implicit representation of high-resolution shape S_{HR} and the estimated implicit representation of the reconstructed super-resolution shape S_{SR} . This loss facilitates the learning of the map as shown in Section 4.2. Low and high-resolution features are extracted from the LR input image and then given as input to two multi-layer perceptrons (MLPs), which estimate the probability field over the 3D space. The pixel-alignment feature used in previous approaches [33,34] is adopted by SuRS to preserve local details and

infer plausible shape detail in unseen regions. Extensive experiments show that SuRS reconstructs significantly higher resolution human shape detail compared to previous single image human reconstruction approaches when LR images are given as input (Fig. 1). The level of details on the reconstructed surfaces is also much higher to that achieved by state-of-the-art approaches that leverage auxiliary data in the reconstruction process (Fig. 2).

The main contributions of this work are:

- An end-to-end trainable approach for implicit function learning of super-resolution shape for 3D human digitization from single low-resolution image.
- Introduction of a novel loss to learn the mapping between low and high resolution shapes based on the information loss between their representations.
- Improved quantitative and perceptual performance for human shape reconstruction from a single LR image compared with state-of-the-art approaches.

2 Related Works

3D super resolution: The concept of super resolution (SR) is well established in the 2D domain. On the other hand, there is considerably less work related to super resolution applied to the 3D domain. Most approaches focus on super-resolving 2.5D data such as texture maps [19,30,28,29], depth maps [40,25,38,35], or light field images [32,46]. All of these methods apply similar network architectures to the ones used to super-resolve 2D images. They consider the 2.5D data as a RGB image and leverage a modified convolutional neural network (CNN) architecture with additional data such as normal maps or edge images as input. Sinha et al. [37] and Chen et al. [7] first retrieve a geometry image that embeds geometric information and then super-resolve it with a standard SR CNN, reconstructing the final HR mesh. In this case, the super-resolution task consists of creating the HR mesh by subdividing the LR mesh through remeshing and super-resolving the extracted geometry images. In other methods [42,9], the super-resolution task is formulated for point cloud data as increasing the number of points in the final model, achieved by applying graph networks. Li et al. [20] first retrieve a coarse estimation of human shape from multi-view images and then refine the reconstruction by voxelizing the shape and applying 3D convolution to super-resolve the voxel. The SR voxel is obtained by subdividing each voxel grid in order to increase their total number in the final voxel. We introduce a new concept for super-resolution of shape represented via high-detail implicit function. In this paper, super resolution is defined as the reproduction of high quality details in the reconstructed shape even if they are not embedded in the input. To overcome the resolution limitation of previous methods, we propose a novel approach that learns to map a LR shape to its HR counterpart.

3D Human digitization: Early 3D human digitization strategies such as fitting the input image by a parametric model [6,16,27,17,43], using a 3D CNN to regress volumetric representation [39,48] or adding a displacement on each vertex of the mesh [3,4,2], mainly aim to recover the naked body’s shape with only approximate representation of shape details such as clothing and hairs.

With the ground-breaking introduction of implicit functions for 3D reconstruction [23, 26, 8, 14, 12], deep learning approaches that estimate 3D human shape, started to adopt this form of representation because it can represent fine-scale details on the reconstructed surfaces. The first work to adopt the implicit representation for 3D human digitization from a single image was PIFu [33], which exploits pixel-aligned image features rather than global features. The local details present in the input image are preserved since the occupancy of any 3D point is predicted.

There are two main problems with this approach that are addressed from other works: it may generate incorrect body structures and struggle to recover fine-scale geometric details. With regards to the first problem, several frameworks combine the implicit function with either a parametric model [14, 47] or geometry-aligned shape feature [11] to improve the stability of prediction since they serve as a shape prior for the reconstruction. These works focus on improving the parts of the reconstructed geometry that are not seen in the input image, deliberately neglecting the representation of fine details. To increase the quality of the reconstructed shape, PIFuHD [34] leverages HR images ($> 1k$ resolution) as well as front and back normal maps in the reconstruction process. The normal maps need to be predicted from the HR input image with an additional network. The use of HR images and normal maps increases the computation cost and memory use. Other approaches reconstruct high-detail shape using multiple images [13, 45, 20, 49], which is a different objective from single image 3D reconstruction. Our proposed framework aims to reconstruct super-resolution geometries from a single LR image (256×256) without leveraging auxiliary data.

3 Methodology

We introduce a novel framework that aims to learn a high-detail implicit shape representation to recover Super-Resolution Shape (SuRS) from a single LR RGB image. A LR image does not contain significant information about fine details and previous approaches for shape reconstruction from a single image cannot reconstruct detailed shape if the fine detail are not embedded in the input (Section 4.3). Inspired by the 2D image super-resolution objective of mapping LR images to their HR counterparts, SuRS learns to map a LR surface S_{LR} to its HR counterpart S_{HR} . By learning the missing information of S_{LR} with respect to S_{HR} , our approach can reproduce the fine details of the HR shape even if they are not embedded in the input. We define Super-Resolution shape S_{SR} as a shape with fine details reconstructed from a LR input image.

3.1 Implicit Function Representations of Super-Resolution Shape

An implicit function defines a surface as a level set of a function f , e.g. $f(X) = 0$ [36] where X is a set of 3D points in \Re^3 . To represent a 3D surface S , this function $f(X)$ is modelled with a Multi-Layer Perceptron (MLP) that classifies

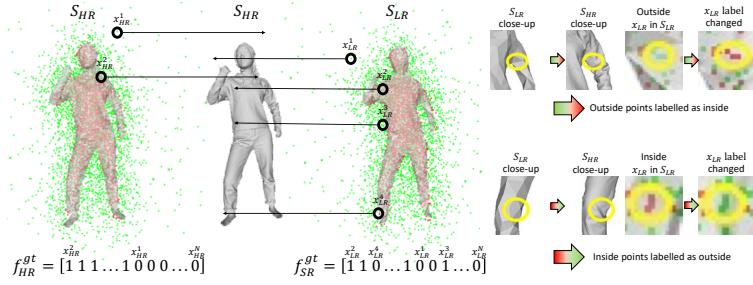


Fig. 3. On the left, comparison between ordinary implicit representation of 3D points x_{HR} sampled from the HR surface S_{HR} space and implicit function representation of super-resolution shape with 3D points x_{LR} sampled from the LR surface S_{LR} space and classified with respect to S_{HR} . The red points are ‘inside’ the surfaces while the green ones are ‘outside’. On the right, close-up of the labelling process: some points ‘outside’ S_{LR} are represented as ‘inside’ points (first row) while some other points ‘inside’ S_{LR} are labelled as ‘outside’ (second row).

the 3D points as either ‘inside’ or ‘outside’ the 3D surface S . A HR surface S_{HR} can be represented as a 0.5 level-set of a continuous 3D occupancy field:

$$f_{HR}^{gt}(x_{HR}) = \begin{cases} 1, & \text{if } x_{HR} \text{ is inside } S_{HR} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where x_{HR} is a point in the 3D space around the HR surface S_{HR} . This means that f_{HR}^{gt} is a vector of length N , where N is the total number of 3D points x_{HR} that are sampled from S_{HR} space and classified as ‘inside’ or ‘outside’ points:

- $f_{HR}^{gt}[0 : \frac{N}{2}] = 1$ correspond to the $\frac{N}{2}$ 3D points inside S_{HR} ;

- $f_{HR}^{gt}[\frac{N}{2} + 1 : N] = 0$ correspond to the $\frac{N}{2}$ 3D points outside S_{HR} ;

To map the LR surface to the HR surface, we adapt the implicit function representation to the 3D super-resolution shape. We define a new ground truth for the high-detail implicit function of the estimated super-resolution shape S_{SR} :

$$f_{SR}^{gt}(x_{LR}) = \begin{cases} 1, & \text{if } x_{LR} \text{ is inside } S_{HR} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where x_{LR} are the 3D points sampled from the space of the LR surface S_{LR} . In contrast to the common implicit function, f_{SR}^{gt} is created by classifying the 3D points x_{LR} with respect to the HR surface instead of the LR one: some x_{LR} points are labelled as ‘outside’ even if they are ‘inside’ S_{LR} , and viceversa (Fig. 3). The vector f_{SR}^{gt} contains $N_{inside} \neq \frac{N}{2}$ values equal to 1 and $N_{outside} \neq N_{inside} \neq \frac{N}{2}$ values equal to 0. This will lead the MLP that models f_{SR}^{gt} to classify points of the LR surface S_{LR} as points of an HR surface S_{HR} , learning a map from S_{LR} to S_{HR} . The points that change classification represent the difference between S_{HR} and S_{LR} , which is primarily in the fine details of the shapes. SuRS learns to infer this difference to estimate super-resolution shape detail from a LR image.

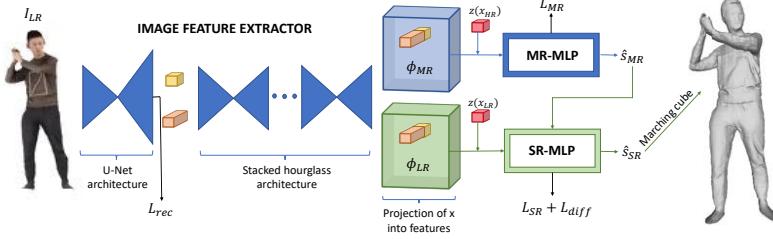


Fig. 4. Network architecture of SuRS. The image feature extractor creates an embedding from the LR input image. 3D points are projected on the embedding and classified first by MR-MLP to produce a mid-level resolution estimation and then by SR-MLP to obtain a super-resolution estimation, from which the 3D SR shape is reconstructed.

3.2 Super-Resolution Single-View Shape Estimation

We apply SuRS to the task of 3D human digitization from a single LR RGB image. To reconstruct a 3D human shape from a single high-resolution RGB image I_{HR} , Saito et al. [33] introduced pixel aligned implicit function (PIFu), whose objective is to predict the binary occupancy value for any given 3D position x in continuous camera space. Instead of classifying directly the 3D points X of the surface, PIFu estimates the level-set by exploiting the image feature $\phi(p)$ of the 2D projection $p = \pi(x_{HR})$ of the 3D point x_{HR} in the input image I_{HR} . The function f is represented by a MLP that is conditioned on the depth value $z(x_{HR})$ of x_{HR} in the camera coordinate space:

$$f_{HR}(\phi(p, I_{HR}), z(x_{HR})) = \hat{s}_{HR}, \hat{s}_{HR} \in \mathbb{R} \quad (3)$$

We modify the pixel-aligned implicit function representation by adapting it to the reconstruction of super-resolution shape from a low-resolution image I_{LR} :

$$f_{SR}(\phi(p_{LR}, I_{LR}), z(x_{LR}), \hat{s}_{MR}) = \hat{s}_{SR}, \hat{s}_{SR} \in \mathbb{R} \quad (4)$$

where x_{LR} are the 3D points sampled from the space of the LR surface S_{LR} , $p_{LR} = \pi(x_{LR})$, \hat{s}_{MR} is a mid-resolution estimation of the shape computed with $f_{HR}^{gt}(x_{HR})$ from a LR input image and \hat{s}_{SR} is a super-resolution estimation of the shape computed with $f_{SR}^{gt}(x_{LR})$. SuRS is modelled via a neural network architecture composed of 3 modules and trained end-to-end (Fig. 4).

- **Image features extractor:** an image feature embedding is extracted from the projected 2D location of the 3D points in the LR image I_{LR} . Specifically, the embedding is the concatenation of two feature vectors extracted from the input image of size $(N_I \times N_I)$: one with the resolution of $(\frac{N_I}{2} \times \frac{N_I}{2})$ to maintain holistic reasoning and the second of higher resolution $(2N_I \times 2N_I)$ to embed the fine detail of the input image. The combination of holistic reasoning and high-resolution image features is essential for high-fidelity 3D reconstruction [34].
- **Mid-resolution multi-layer perceptron (MR-MLP):** a first classification \hat{s}_{MR} of the 3D points x_{HR} sampled from S_{HR} space is computed with an MLP

that models the pixel aligned implicit function $f_{MR}(\phi(p, I_{LR}), z(x_{HR}))$. x_{HR} are projected through orthogonal projection to the LR input image I_{LR} and indexed to the feature embedding retrieved before. This is concatenated with the depth value $z(x_{HR})$ of x_{HR} in the camera coordinate space and given as input to MR-MLP. f_{MR} differs from f_{HR} because the input is a low-resolution image: the output representation of MR-MLP does not reconstruct high-detail shape since information is missing in the LR input image and it cannot be learned with the standard implicit function modelled by MR-MLP (Section 4.2).

- **Super-resolution multi-layer perceptron (SR-MLP):** the final estimation \hat{s}_{SR} is obtained with a further MLP, which models the implicit function f_{SR} to represent super-resolution shape. The input of SR-MLP differs from the MR-MLP one since the 3D points x_{LR} are sampled from the space of the LR surface S_{LR} , projected to the input image and indexed to the feature embedding. Their projection is then concatenated with the depth value $z(x_{LR})$ of x_{LR} . To help the learning of a map from S_{LR} to S_{HR} , the estimation \hat{s}_{MR} is concatenated with the input embedding of SR-MLP to facilitate the learning of the lost information of S_{LR} . SR-MLP trained with f_{SR}^{gt} learns to classify x_{LR} according to the difference between S_{LR} and S_{HR} . Compared to MR-MLP, SR-MLP additionally infers this difference to the estimation \hat{s}_{SR} , representing fine details on the SR shape even if they are not represented in the LR input image.

During inference, a LR image is given as input to the network. A set of M random 3D points x of the 3D space are projected on the extracted embedding and then classified first with MR-MLP to estimate \hat{s}_{MR} and then with SR-MLP to estimate \hat{s}_{SR} . The final shape is obtained by extracting iso-surface $f = 0.5$ of the probability field \hat{s}_{SR} at threshold 0.5 applying the Marching Cube algorithm [21].

3.3 Training Objectives

Our objective function combines 4 different losses used to train the 3 modules:

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{MR} + \mathcal{L}_{SR} + \mathcal{L}_{diff} \quad (5)$$

2D reconstruction loss: the higher resolution feature extracted from the LR input image is double the size of the image. To ensure that the newly created features will preserve the spatial structure of the LR input image, SuRS minimises the L1 loss commonly used in the 2D image super-resolution task [41]:

$$\mathcal{L}_{rec} = \|I_{SR} - I_{GT}\|_1 \quad (6)$$

I_{SR} is an image reconstructed from the HR features extracted by the feature extractor module and processed by a further convolutional layer. I_{GT} is the ground-truth of the reconstructed image I_{SR} .

MR loss: since MR-MLP represents a pixel aligned implicit function, we train the network by minimising the average of mean squared error between MR-MLP output and the canonical ground-truth of implicit functions as done in PIFu [33]:

$$\mathcal{L}_{MR} = \frac{1}{N} \sum_{i=1}^N |f_{MR}(\phi_{MR}^i, z(x_{HR}^i)) - f_{HR}^{gt}(x_{HR}^i)|^2 \quad (7)$$

where N is the number of points x_{HR} sampled from the space of S_{HR} , $\phi_{MR}^i = \phi(p, I_{LR})$ and $f_{MR}(\phi_{MR}^i, z(x_{HR}^i)) = \hat{s}_{MR}$.

SR loss: to compute the final estimation, the average of mean squared error between the output of SR-MLP and SuRS ground-truth is minimised:

$$\mathcal{L}_{SR} = \frac{1}{N} \sum_{i=1}^N |f_{SR}(\phi_{LR}^i, z(x_{LR}^i), \hat{s}_{MR}) - f_{SR}^{gt}(x_{LR}^i)|^2 \quad (8)$$

where $\phi_{LR}^i = \phi(p_{LR}, I_{LR})$, $f_{SR}(\phi_{LR}^i, z(x_{LR}^i), \hat{s}_{MR}) = \hat{s}_{SR}$ and the N points x_{LR} are sampled from the space of S_{LR} .

Difference loss: to facilitate the learning of the map from S_{LR} to S_{HR} , we design a novel loss that minimises the average of mean squared error of the difference between S_{SR} and S_{HR} , expressed as the difference between the estimated values \hat{s}_{MR} and \hat{s}_{SR} :

$$\mathcal{L}_{diff} = \frac{1}{N} \sum_{i=1}^N |(f_{HR}^{gt} - f_{SR}^{gt}) - (\hat{s}_{MR} - \hat{s}_{SR})|^2 \quad (9)$$

Since f_{SR}^{gt} contains the missing information of S_{LR} , this loss leads the network to learn in which points S_{LR} differs from S_{HR} . Since these points represent the fine details, the network learns how to reproduce them in the estimated shape. This loss is important to learn the super-resolution shape detail (see Section 4.2).

4 Experiments

We quantitatively and qualitatively evaluate the proposed approach on the task of reconstructing 3D human shape from a single LR image. To train SuRS, LR surfaces must be retrieved from a set of HR example models. The performance will improve if the displacement between S_{HR} and S_{LR} is higher as shown in Section 4.2. Therefore, we use the THuman2.0 [44] dataset, which consists of 524 high-resolution surfaces which have a high level of fine details. We split the dataset into training (402 models) and testing (122 models). To evaluate the generalisation of the method, we test SuRS on 136 models taken from 3D People [1]. We evaluate the reconstruction accuracy with 3 quantitative metrics: the normal reprojection error introduced in [33], the average point-to-surface Euclidean distance (P2S) and the Chamfer distance (CD), expressed in cm.

4.1 Implementation Details

SuRS is trained with LR images I_{LR} of size $(N_I \times N_I, N_I = 256)$ obtained by downscaling by a factor of 2 ground truth images I_{GT} with bicubic degradation. We first apply [5] to make the HR surfaces watertight. For training, a LR shape S_{LR} is obtained by remeshing S_{HR} from $\sim 400k$ to $1k$ faces via quadric edge collapse decimation [10]. $N_T = 24000$ 3D points X are sampled with a mixture of uniform sampling and importance sampling around surface S_{HR} following PIFu [33]. A subset of X of $N = 6000$ x_{HR} points are sampled for S_{HR} by

selecting $\frac{N}{2}$ points inside and $\frac{N}{2}$ outside S_{HR} . A different subset of X of $N = 6000$ x_{LR} points are sampled for S_{LR} by selecting $\frac{N}{2}$ points inside and $\frac{N}{2}$ outside S_{LR} . The novel architecture of the image feature extractor is a combination of U-Net [31] and stacked hourglass architectures [24]. The former retrieves the features whose resolution is double that of the input image while the second has been proved to be effective for surface reconstruction [24]. We design a novel U-Net architecture to retrieve both high and low-resolution features from the input image (see supplementary material). The HR features are then processed by a convolution layer while the LR features by 3 stacks. The outputs of the image feature extractor are a HR feature vector of size $(2N_I \times 2N_I \times 64)$ and a LR one of size $(\frac{N_I}{2} \times \frac{N_I}{2} \times 256)$. The final embedding is obtained with the concatenation of these two vectors. MR-MLP has the number of neurons of (321, 1024, 512, 256, 128, 1) while SR-MLP of (322, 1024, 512, 256, 128, 1) with skip connections at 3rd, 4th and 5th layers.

4.2 Ablation Studies

Training configurations. In the first ablation study we demonstrate that training the losses of the network in an end-to-end manner improves the performance of SuRS. We evaluate the network by training separately the three modules; by training first the feature extractor (\mathcal{L}_{rec}) and then the two MLPs together ($\mathcal{L}_{diff} + \mathcal{L}_{MR} + \mathcal{L}_{SR}$); by first training the feature extractor and the MR-MLP together ($\mathcal{L}_{rec} + \mathcal{L}_{MR}$) and then the SR-MLP ($\mathcal{L}_{diff} + \mathcal{L}_{SR}$). \mathcal{L}_{diff} is always trained with \mathcal{L}_{SR} since it depends on the output of SR-MLP. The values of the considered metrics are the highest when the end-to-end training is adopted (Table 1). This is confirmed by the qualitative evaluation (Fig.5). The worst results are obtained when the feature extractor (\mathcal{L}_{rec}) is trained separately from the MLPs (case 1, 2). In these cases, the feature extractor learns how to create the embedding without being influenced by the classification part of the network: the extracted features embed information that is meaningless for the task of 3D human reconstruction. When the feature extractor is trained along with the MLPs (case 3, 4), the extracted features are influenced by the other losses and they are more informative for the objective of 3D human digitization. This proves that just applying a 2D image super-resolution network to upscale the input image as a pre-processing step is less efficient than the proposed approach. Training SR-MLP separately introduces noise in the reconstructed shapes while the end-to-end configuration produces the best results.

Decimation factors. To create the LR geometry from the HR shape, we apply quadric edge collapse decimation that reduces the number of faces. We evaluate our approach by changing the decimation factor of the number of faces of S_{LR} . Low-resolution geometries with 100000, 50000, 10000 and 1000 faces are created (examples are illustrated in the supplementary material). Both the quantitative (Table 2) and qualitative (Fig. 6) results show that the performance increases when the number of faces of S_{LR} is the lowest, with significantly sharper details on the output shapes. Our approach is more efficient if the difference between the HR ground-truth shape and the LR ground-truth one is higher. SuRS

Table 1. Quantitative results obtained by training SuRS with different configurations.

Training Configuration	THuman2.0		3D people	
	CD	Normal P2S	CD	Normal P2S
1: $\mathcal{L}_{rec}; \mathcal{L}_{MR}; \mathcal{L}_{SR} + \mathcal{L}_{disp}$	1.375	0.1360	1.454	1.509 0.1261 1.636
2: $\mathcal{L}_{rec}; \mathcal{L}_{MR} + \mathcal{L}_{SR} + \mathcal{L}_{disp}$	1.261	0.1347	1.496	1.243 0.1240 1.470
3: $\mathcal{L}_{rec} + \mathcal{L}_{MR}; \mathcal{L}_{SR} + \mathcal{L}_{disp}$	1.035	0.1106	1.083	1.121 0.1160 1.281
4: End to end (ours)	0.931	0.1065	1.151	1.057 0.1127 1.247

Table 2. Quantitative evaluation of using different decimation factors to create the low-resolution ground-truth shape.

Nr. faces	THuman2.0		3D people	
	CD	Normal P2S	CD	Normal P2S
100000	0.942	0.1084	1.117	1.147 0.1123 1.331
50000	0.950	0.1079	1.117	1.118 0.1120 1.291
10000	0.941	0.1113	1.116	1.084 0.1133 1.276
1000 (ours)	0.931	0.1065	1.151	1.057 0.1127 1.247

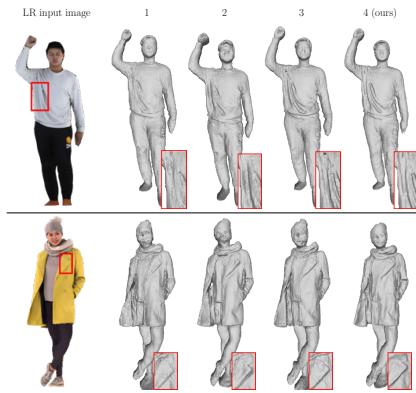


Fig. 5. Visual results obtained by changing the configuration of training. 1, 2, 3 and 4 are different configurations (refer to Table 1). The upper model is from THuman2.0, the below one from 3DPeople.

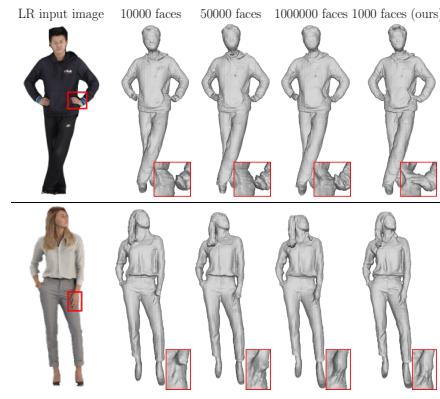


Fig. 6. Qualitative results obtained by using different decimation factors to create the low-resolution ground-truth shape. The upper model is from THuman2.0, the below one from 3DPeople.

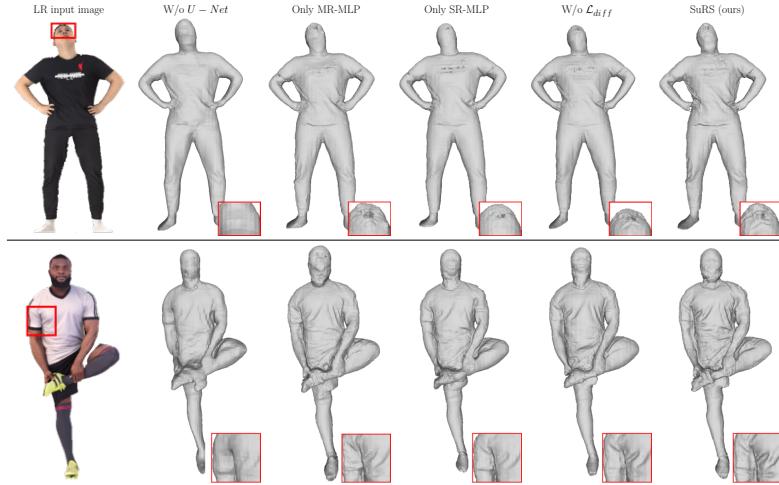
learns to estimate missing information to obtain the SR shape by mapping S_{LR} to S_{HR} . As S_{LR} is coarser, the information loss increases while if S_{LR} is similar to S_{HR} , there is no missing information, hence nothing to learn. The LR shapes with 1000 faces are the ones with the highest displacement from their HR counterparts, hence the improvement in the performance.

Efficiency of the proposed architecture and displacement loss. We demonstrate the significant improvement achieved by the combination of the 3 modules and by the introduction of \mathcal{L}_{diff} . We train and test 4 frameworks obtained by modifying the architecture of SuRS:

- **Without U-Net:** the U-Net architecture of the feature extractor is not implemented. The feature embedding, extracted only by the stacked hour-glass part of the future extractor, is composed by two feature vectors of size $(N_I \times N_I \times 64)$ and $(\frac{N_I}{2} \times \frac{N_I}{2} \times 256)$. \mathcal{L}_{rec} is not minimised during training.
- **Only MR-MLP:** this framework consists only of the image feature extractor and MR-MLP. SR-MLP is not implemented and the shape is reconstructed from \hat{s}_{MR} . \mathcal{L}_{SR} and \mathcal{L}_{diff} are not considered during training.

Table 3. Quantitative results obtained by changing the architecture of the approach.

Architecture	THuman2.0			3D people		
	CD	Normal P2S	CD	Normal P2S		
W/o U-Net	1.090	0.1228	1.204	1.260	0.1296	1.364
Only MR-MLP	1.044	0.1066	1.077	1.240	0.1129	1.276
Only SR-MLP	1.039	0.1146	1.051	1.223	0.1215	1.255
W/o \mathcal{L}_{diff}	1.020	0.1138	1.196	1.127	0.1200	1.255
SuRS (ours)	0.931	0.1065	1.151	1.057	0.1127	1.247

**Fig. 7.** Visual results obtained by changing the architecture of our approach. The upper model is from THuman2.0 while the below is from 3DPeople.

- **Only SR-MLP:** the feature embedding is directly processed by SR-MLP that outputs the estimation \hat{s}_{SR} without being conditioned on \hat{s}_{MR} . \mathcal{L}_{MR} and \mathcal{L}_{diff} are not minimised during training.
- **Without \mathcal{L}_{diff} :** this framework is the same as the proposed one but it is trained without considering \mathcal{L}_{diff} to check the importance of this loss.

Both the quantitative (Table 3) and the qualitative (Fig. 7) evaluations show the importance of implementing all the different modules of SuRS. If the HR features are not extracted (W/o U-Net), the fine details are lost in the reconstructed shape, which is just a coarse representation of the input image. Less sharp details are obtained when only MR-MLP is applied, with a lower resolution of the final reconstruction. This configuration does not use either the f_{SR}^{gt} ground-truth or the L_{diff} and this deteriorates the performance. In this case, the network does not learn the map from S_{LR} to S_{HR} and the lost information is not retrieved. Similarly, this map is not learned when only SR-MLP is applied and significant lower resolution shapes are obtained compared to when the map is learned (combination of SR-MLP and MR-MLP with HR features), confirming the efficiency of SuRS. The difference between training SuRS with or without

Table 4. Quantitative comparisons between state-of-the-art approaches with LR input image for training and testing. The highest scores are highlighted in red while the second highest scores are blue.

N_I	Methods	THuman2.0		3D people	
		CD	Normal P2S	CD	Normal P2S
2	DeepHuman [48]	1.956	0.1465	2.063	1.703 0.1253 1.733
5	PIFu [33]	1.518	0.1218	1.647	1.519 0.1198 1.581
6	PIFuHD _{no}	1.308	0.1237	1.346	1.440 0.1595 1.464
6	PIFuHD [34]	1.032	0.1116	1.046	1.062 0.1257 1.037
6	PaMIR [47]	1.713	0.1341	1.818	1.644 0.1410 1.740
6	Geo-PIFu [11]	1.786	0.1473	1.724	1.853 0.1658 1.652
	SuRS (ours)	0.931	0.1065	1.151	1.057 0.1127 1.247

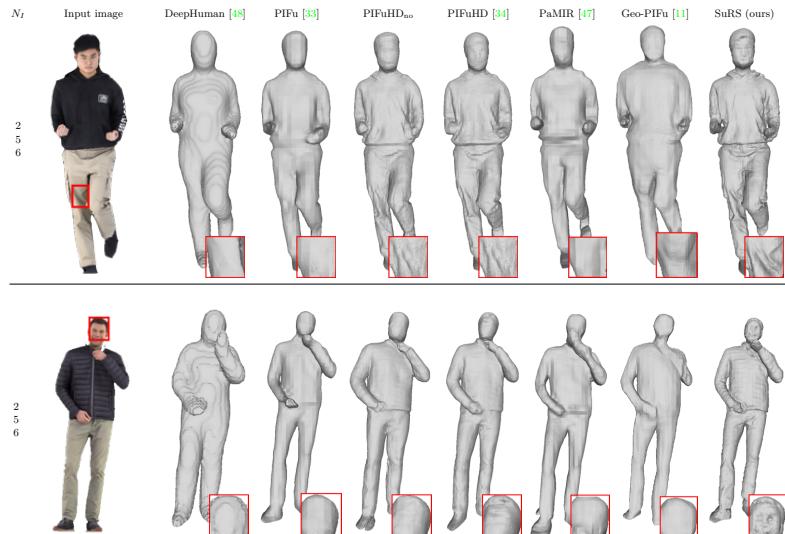


Fig. 8. Visual comparisons using LR input image for training and testing. The upper model is from THuman2.0, the below one is from 3DPeople.

L_{diff} is significant proving its efficiency: the wrinkles of the dress are sharper and human body parts are more realistic when L_{diff} is minimised.

4.3 Comparisons

We compare SuRS with related works on 3D human digitization from a single image. We evaluate DeepHuman [48], which do not use implicit representation but leverage parametric models to facilitate the reconstruction. We then compare SuRS with approaches that represent the 3D surfaces with implicit representation. Among these, PIFu [33] is the only one which uses only single RGB images. PIFuHD [34] exploits front and back normal maps while PaMIR [47] and Geo-PIFu [11] leverage parametric models. We train and test PIFuHD without using normal maps (PIFuHD_{no}). These approaches are trained and tested with the same datasets as SuRS except for Geo-PIFu, which cannot pre-processed THuman2.0 shapes. See supplementary material for further comparisons.

Qualitative evaluation: Fig. 8 illustrates two shapes reconstructed from a LR input image of sizes $N_I = 256$. Compared to other methods, SuRS reconstructs significantly higher resolution shapes, which contain the highest level of fine details in clothes and faces. Related works produce coarse shapes with smooth details even if other data such as normal maps or parametric models is leveraged.

Quantitative evaluation: Our approach outperforms related works when only LR input images are used in both training and testing. SuRS achieves the highest values of the considered metrics, proving its superiority over all the considered approaches (Table 4). SuRS outperforms also all the other approaches that use auxiliary data for the CD and Normal metrics. PiFUHD achieves lower values of P2S due to its use of back normal maps, which improve its performance on parts of the model that are not seen in just a single image.

Real data: We qualitatively evaluate our approach on images of people captured in real scenarios. More specifically, we reconstruct the 3D shape of a person from an HR image (> 1024) where the person resolution is at a relatively low-resolution sub-image of the original HR image ($< 350 \times 350$). We select random images from various datasets [18,15] and we compare SuRS with only the approaches that do not use auxiliary data in the reconstruction, namely PiFU and PIFuHD_{no}. For the reconstruction, a LR human body patch is cropped from the HR image. SuRS significantly outperforms the other approaches, reproducing higher level of fine details on the reconstructed shapes, which are coarser with smoother details if reconstructed by the other approaches (Fig. 9).

Limitations: our method cannot super-resolve parts of the human body that are not visible on the input image: since no auxiliary data are provided, SuRS reconstructs a coarse geometry of the hidden parts of the body. Like existing works, it may generate incorrect body structures when the input model presents features that significantly differ from the ones seen during training. It may also suffer problems related to depth ambiguity. Examples in supplementary material.

5 Conclusion

We propose Super-Resolution shape represented by an high-detail implicit representation and we tackle the problem of 3D human digitization from a single low-resolution image. To achieve this, we propose a novel architecture with a customised novel loss that can reconstruct super-resolution shape from a low-resolution image. As demonstrated by the evaluation, the reconstructed surfaces contain significantly higher level of details compared to the outputs of related methods when low-resolution images are used. The resolution of the shape obtained from a low-resolution image is significantly higher than the one obtained by state-of-the-art works that leverage also auxiliary data such as normal maps and parametric models. As future works, a super-resolution texture representation will be investigated and the approach will be extended to dynamic shape.

Acknowledgement: this research was supported by UKRI EPSRC Platform Grant EP/P022529/1

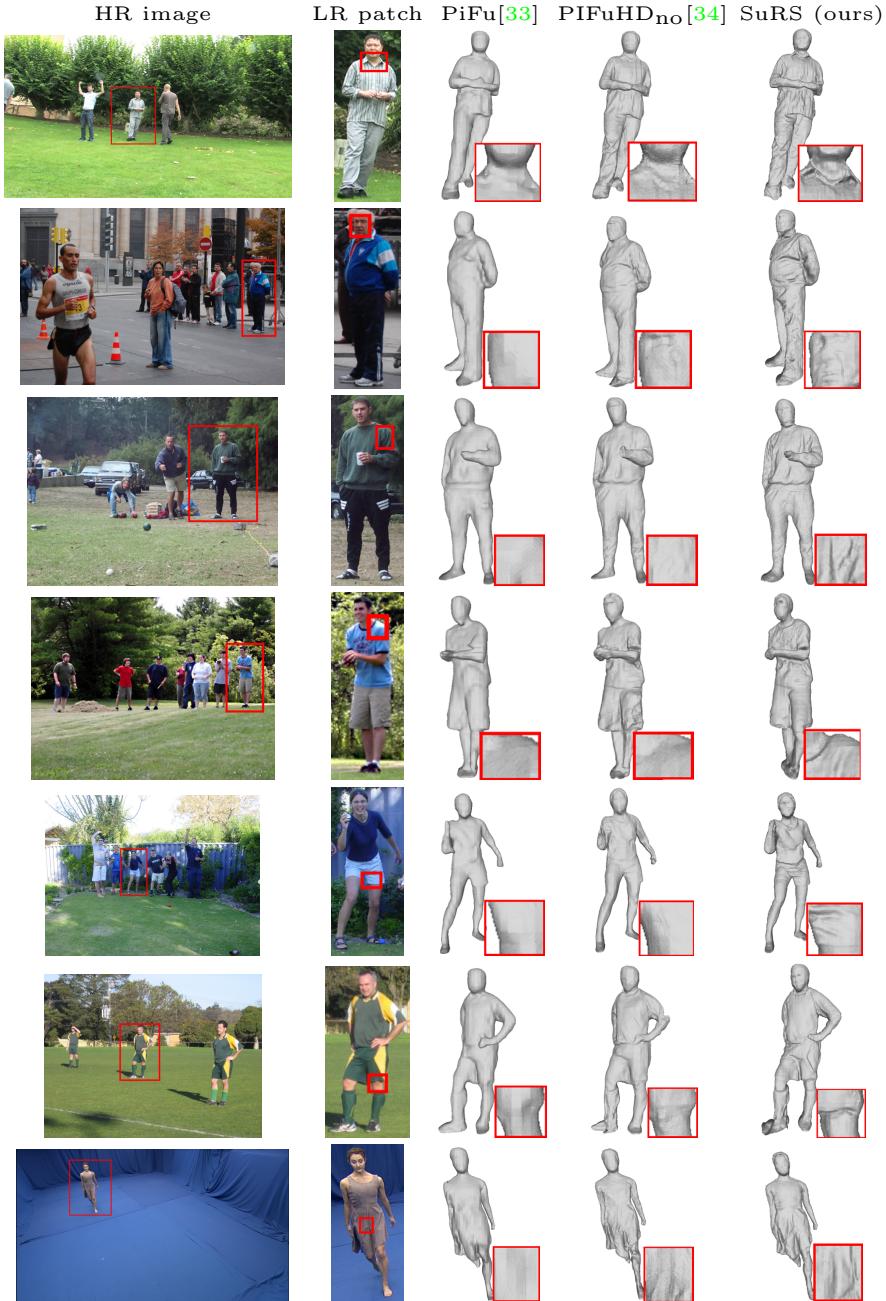


Fig. 9. Human shapes reconstructed from LR patches of HR images of people.

References

1. 3d people. <https://3dpeople.com/en/>, accessed: 2021-10-16
2. Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single rgb camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1175–1186 (2019)
3. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8387–8397 (2018)
4. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2293–2303 (2019)
5. Barill, G., Dickson, N., Schmidt, R., Levin, D.I., Jacobson, A.: Fast winding numbers for soups and clouds. ACM Transactions on Graphics (2018)
6. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: European conference on computer vision. pp. 561–578. Springer (2016)
7. Chen, L., Ye, J., Jiang, L., Ma, C., Cheng, Z., Zhang, X.: Synthesizing cloth wrinkles by cnn-based geometry image superresolution. Computer Animation and Virtual Worlds **29**(3-4), e1810 (2018)
8. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5939–5948 (2019)
9. Dinesh, C., Cheung, G., Bajic, I.V.: Super-resolution of 3d color point clouds via fast graph total variation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1983–1987. IEEE (2020)
10. Garland, M., Heckbert, P.S.: Simplifying surfaces with color and texture using quadric error metrics. In: Proceedings Visualization'98 (Cat. No. 98CB36276). pp. 263–269. IEEE (1998)
11. He, T., Collomosse, J., Jin, H., Soatto, S.: Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. arXiv preprint arXiv:2006.08072 (2020)
12. He, T., Xu, Y., Saito, S., Soatto, S., Tung, T.: Arch++: Animation-ready clothed human reconstruction revisited. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11046–11056 (2021)
13. Hong, Y., Zhang, J., Jiang, B., Guo, Y., Liu, L., Bao, H.: Stereopifu: Depth aware clothed human digitization via stereo vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 535–545 (2021)
14. Huang, Z., Xu, Y., Lassner, C., Li, H., Tung, T.: Arch: Animatable reconstruction of clothed humans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3093–3102 (2020)
15. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: Proceedings of the British Machine Vision Conference (2010), doi:10.5244/C.24.12
16. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7122–7131 (2018)
17. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5253–5263 (2020)

18. Li, L.J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition. In: 2007 IEEE 11th international conference on computer vision. pp. 1–8. IEEE (2007)
19. Li, Y., Tsiminaki, V., Timofte, R., Pollefeys, M., Gool, L.V.: 3d appearance super-resolution with deep learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9671–9680 (2019)
20. Li, Z., Oskarsson, M., Heyden, A.: Detailed 3d human body reconstruction from multi-view images combining voxel super-resolution and learned implicit representation. arXiv preprint arXiv:2012.06178 (2020)
21. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. ACM siggraph computer graphics **21**(4), 163–169 (1987)
22. Malleson, C., Collomosse, J., Hilton, A.: Real-time multi-person motion capture from multi-view video and imus. International Journal of Computer Vision **128**(6), 1594–1611 (2020)
23. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4460–4470 (2019)
24. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016)
25. Ni, M., Lei, J., Cong, R., Zheng, K., Peng, B., Fan, X.: Color-guided depth map super resolution using convolutional neural network. IEEE Access **5**, 26666–26672 (2017)
26. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 165–174 (2019)
27. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10975–10985 (2019)
28. Pesavento, M., Volino, M., Hilton, A.: Attention-based multi-reference learning for image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14697–14706 (2021)
29. Pesavento, M., Volino, M., Hilton, A.: Super-resolution appearance transfer for 4d human performances. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1791–1801 (2021)
30. Richard, A., Cherabier, I., Oswald, M.R., Tsiminaki, V., Pollefeys, M., Schindler, K.: Learned multi-view texture super-resolution. In: 2019 International Conference on 3D Vision (3DV). pp. 533–543. IEEE (2019)
31. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
32. Rossi, M., Frossard, P.: Geometry-consistent light field super-resolution via graph-based regularization. IEEE Transactions on Image Processing **27**(9), 4207–4218 (2018)
33. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2304–2314 (2019)

34. Saito, S., Simon, T., Saragih, J., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 84–93 (2020)
35. Sang, L., Haefner, B., Cremers, D.: Inferring super-resolution depth from a moving light-source enhanced rgbd sensor: A variational approach. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1–10 (2020)
36. Sclaroff, S., Pentland, A.: Generalized implicit functions for computer graphics. ACM Siggraph Computer Graphics **25**(4), 247–250 (1991)
37. Sinha, A., Unmesh, A., Huang, Q., Ramani, K.: Surfnet: Generating 3d shape surfaces using deep residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6040–6049 (2017)
38. Song, X., Dai, Y., Zhou, D., Liu, L., Li, W., Li, H., Yang, R.: Channel attention based iterative residual learning for depth map super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5631–5640 (2020)
39. Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., Schmid, C.: Bodynet: Volumetric inference of 3d human body shapes. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 20–36 (2018)
40. Voynov, O., Artemov, A., Egiazarian, V., Notchenko, A., Bobrovskikh, G., Burnaev, E., Zorin, D.: Perceptual deep depth super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5653–5663 (2019)
41. Wang, Z., Chen, J., Hoi, S.C.: Deep learning for image super-resolution: A survey. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3365–3387 (2020)
42. Wu, H., Zhang, J., Huang, K.: Point cloud super resolution with adversarial residual graph networks. arXiv preprint arXiv:1908.02111 (2019)
43. Xu, X., Chen, H., Moreno-Noguer, F., Jeni, L.A., De la Torre, F.: 3d human pose, shape and texture from low-resolution images and videos. IEEE transactions on pattern analysis and machine intelligence (2021)
44. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021) (June 2021)
45. Zhang, S., Liu, J., Liu, Y., Ling, N.: Dimnet: Dense implicit function network for 3d human body reconstruction. Computers & Graphics **98**, 1–10 (2021)
46. Zhang, S., Chang, S., Lin, Y.: End-to-end light field spatial super-resolution network using multiple epipolar geometry. IEEE Transactions on Image Processing **30**, 5956–5968 (2021)
47. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
48. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7739–7749 (2019)
49. Zins, P., Xu, Y., Boyer, E., Wuhrer, S., Tung, T.: Data-driven 3d reconstruction of dressed humans from sparse views. In: 2021 International Conference on 3D Vision (3DV). pp. 494–504. IEEE (2021)