

Article

sSfS: Segmented Shape from Silhouette Reconstruction of the Human Body

Wiktor Krajnik^{1,2}, Łukasz Markiewicz^{1,2} and Robert Sitnik^{1,2,*} 

¹ Mnemosis S. A., 8 Józefa Str., 31-056 Krakow, Poland; w.krajnik@mnemosis.pl (W.K.); lukasz.markiewicz.dokt@pw.edu.pl (Ł.M.)

² Institute of Micromechanics and Photonics, Warsaw University of Technology, 8 Sw. Andrzeja Boboli Str., 02-525 Warsaw, Poland

* Correspondence: robert.sitnik@pw.edu.pl; Tel.: +48-222348283

Abstract: Three-dimensional (3D) shape estimation of the human body has a growing number of applications in medicine, anthropometry, special effects, and many other fields. Therefore, the demand for the high-quality acquisition of a complete and accurate body model is increasing. In this paper, a short survey of current state-of-the-art solutions is provided. One of the most commonly used approaches is the Shape-from-Silhouette (SfS) method. It is capable of the reconstruction of dynamic and challenging-to-capture objects. This paper proposes a novel approach that extends the conventional voxel-based SfS method with silhouette segmentation—segmented Shape from Silhouette (sSfS). It allows the 3D reconstruction of body segments separately, which provides significantly better human body shape estimation results, especially in concave areas. For validation, a dataset representing the human body in 20 complex poses was created and assessed based on the quality metrics in reference to the ground-truth photogrammetric reconstruction. It appeared that the number of invalid reconstruction voxels for the sSfS method was 1.7 times lower than for the state-of-the-art SfS approach. The root-mean-square (RMS) error of the distance to the reference surface was also 1.22 times lower.



Citation: Krajnik, W.; Markiewicz, Ł.; Sitnik, R. sSfS: Segmented Shape from Silhouette Reconstruction of the Human Body. *Sensors* **2022**, *22*, 925. <https://doi.org/10.3390/s22030925>

Academic Editor: Alessandro Bevilacqua

Received: 13 December 2021

Accepted: 20 January 2022

Published: 25 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: Shape from Silhouette; visual hull; human body segmentation; 3D reconstruction; pose estimation; volumetric methods; computer vision; multi-view images

1. Introduction

In past decades, the detailed high-resolution three-dimensional (3D) scanning of human body shapes has rapidly developed. Reconstruction accuracy, quality, and speed have improved significantly over recent years to the point where four-dimensional (4D, i.e., 3D over time) dynamic movement scanners (e.g., Diers International GmbH [1,2], 3dMD LLC [3,4], and Microsoft Corporation [5]) have become an industry standard. Both dynamic and static whole-body scanners have countless applications, such as in special effects in movies/video games [6], in human body pose and deformation analyses for medical diagnostics [7–9], and in anthropometry [10]. However, most of these systems have various limitations, i.e., insufficient precision, high cost, and noisy or incomplete data. The latter is usually caused by inevitable body self-occlusion and results in defective data that are full of holes. Therefore, it is one of the most important areas of interest in the 3D/4D research field.

This paper presents a novel approach to addressing one of those issues, namely surface reconstruction errors. A human body geometry is obtained using a variation of the Shape-from-Silhouette (SfS) voxel-based visual hull estimation method [11]. In the proposed segmented Shape-from-Silhouette (sSfS) approach, the reconstruction is performed for each body segment separately. The independent division into body segments for each image allows for individual 3D reconstruction of shapes with fewer concavities. Consequently,

after combining the reconstructed 3D partial models, the representation of the human body geometry is significantly better compared to the classic SfS method results. For silhouette estimation in images, solutions based on an existing convolutional neural network (CNN) can be used. Therefore, this technique requires only an efficient method for silhouette extraction from the image and a calibrated set of cameras. A state-of-the-art survey of other well-established methods is also presented.

2. Related Works

2.1. 3D Reconstruction Methods Used for Human Body Scanning

The acquisition of the 3D human body shape is a demanding process, requiring a high resolution and short acquisition time due to the body's complexity and sway. A 3D whole-body scanning approach can be developed using various measurement methods [12]. The most common ones are presented in Table 1.

Table 1. A short summary of the 3D reconstruction methods used for human body scanning.

Method Name	Measurement Technique	Type of Illumination
Laser triangulation (LT) [13]	Detection of laser stripes projected onto the object's surface	Laser
Time-of-Flight (ToF) [14]	Measurement of the depth data of the subject surface from return time of light impulse or phase shift	Infrared laser
Structured Light (SL) [15]	Analysis of a light pattern (e.g., fringes) projected onto the measured object's surface	Structured two-dimensional (2D) pattern projected by laser or digital projectors
Photogrammetry [16]	Detection and analysis of the key point correspondences between images of the object taken from different angles simultaneously	Natural sunlight or shadowless illumination

Of these four methods, LT is typically the slowest in terms of acquisition time due to the utilization of a laser line that is incapable of scanning the whole human body surface at once. A single static scan takes a significant amount of time (e.g., 9 s with Vitus Bodyscan [17]) and is not suitable for dynamic motion capture [18]. Despite being faster, the ToF method provides the lowest resolution of the 3D measurement. It can be increased by adding more camera views, but it is limited by the interference of light sources on the measured subject's surface [19]. By comparison, the SL method provides fast and comprehensive 3D shape acquisition [20]. It requires structured pattern projection. One of its drawbacks is light interference from multiple projectors distributed around the measured subject. This issue can be solved with an example solution proposed in [21], in which spectral filters were used to avoid the crosstalk between neighboring cameras and projector modules. Finally, the photogrammetry method benefits from a short image acquisition time, a lack of interference issues, and a relatively simple reconstruction process, with high overall performance [22]. Systems with fixed camera positions are emerging for the measurement of bodies in motion. One of the most common photogrammetric methods is Structure from Motion (SfM), typically followed by the use of the multi-view stereo (MVS) [23] algorithm, responsible for the densification of the SfM point cloud. Multiple open-source and commercial software packages exist for photogrammetric reconstruction, such as VisualSFM [24] and Agisoft Metashape [25]. However, the photogrammetric systems must provide enough views to avoid occlusions in the measured area, as proposed in [26]. In addition, they provide poor results for low-repeating textures, shaded areas, and reflecting spots.

2.2. SfS Visual Hull Estimation

Recently, many authors have proposed reconstruction methods based on a visual hull [11] extraction from a set of multi-view binary silhouette images. Visual hulls can be either computed as an exact polyhedral representation or approximated as voxels.

Polyhedral-based visual hull estimation approaches [27,28] estimate visual cones for each silhouette by casting rays from the camera center to the silhouette edge. Then, the subject's shape is formed based on the intersection of those visual cones. The result is the surface of the subject, by design, without the concave parts of the subject. This approximation can be further enhanced by eliminating rough edges of the visual hull [29,30].

By comparison, voxel-based methods [31,32] divide the 3D space into a grid of voxels. Then, the visual hull is formed based on the voxels projected onto the silhouette images based on the voting threshold (in which each successful voxel projection receives a vote). Moreover, the SfS reconstruction method [33] proposes a solution for non-binary silhouettes containing probability maps, where the formation of the visual hull is introduced as a pseudo-Boolean optimization problem [34].

Due to the ease of computing parallelization, SfS methods are suitable for working in real time [35]. Nevertheless, such implementations are limited by the number of views and the reconstruction resolution. For example, Perez et al. in [36] achieved 30 frames per second rate in a $256 \times 256 \times 128$ voxel resolution. Thus, the resolution achieved through real-time SfS implementation is far from meeting the demands of human body reconstruction.

This approach is not only limited, similarly to other reconstruction methods, by body self-occlusions, but also by the extraction quality of the silhouettes, which makes it even harder to properly reconstruct dynamic objects such as limbs. Corraza et al. in [37] matched rigid segments of an a priori human 3D model to visual hull reconstruction in order to estimate human poses. In contrast, Kanaujia et al. in [38] utilized a previously prepared mesh. Reconstructed points from the visual hull were segmented using mesh vertices. Then, the segmented body parts were used for human joint estimations. Other works have proposed improvements in the visual hull by refining its shape using depth information from RGB (red, green, blue) stereo image matching [39,40].

2.3. Parametric Body Models

In general, these methods use the input 2D image data to estimate the human body model parameters, which represent human poses and body shapes. Examples of such body models are the Skinned Multi-Person Linear (SMPL) model [41] and Shape Completion and Animation of People (SCAPE) [42]. Recent works have proven that it is possible to estimate the body shape using only a single-view video, 2D positions of human joints, or multi-view silhouette images with the SCAPE [43–45] and SMPL [46,47] body models. The reconstruction method proposed in [48] is an example of an encoder-decoder CNN architecture that predicts the SMPL model to fit a set of silhouettes. Alternatively, Dibra et al. in [49] presents a CNN architecture that is first trained to find a human shape model from a set of 3D meshes' body shape invariants. Then, the human pose is estimated based on a single or two-view silhouette.

2.4. Deep Neural Network Architectures

The latest research shows that it is possible to perform 3D human geometry reconstruction using deep learning without a parametric body model [50]. For example, Gilbert et al. [51] used an auto-encoder architecture to estimate human body shapes by improving its volumetric visual hull reconstruction. This solution is similar to super-resolution image enhancement because it augments a coarse visual hull to a high-fidelity body model. Natsume et al. [52] generated a 3D pose from a single view to produce a set of multi-view silhouette images. Then, the silhouettes were used to estimate the final 3D human body reconstruction with deep visual hull prediction, similarly to [50]. In addition, Xu et al. in [53] estimated an actor's rig with a multi-view photogrammetric scan and a skeleton. The proposed algorithm fits the rig to the given pose with the use of silhouette images.

Finally, the solution presented in [54] introduces a coarse-to-fine 3D reconstruction method for the human body. A coarse reconstruction from images was demonstrated with the 2D feature-based Pixel-aligned Implicit Function based on Multi-scale Features (MF-PIF) method, inspired by [55], and this was then refined with a voxel super-resolution network.

2.5. Local Shape Approximation and Meshing

The problem of the generation of watertight 3D models is typically addressed by meshing algorithms. This includes, inter alia, screened Poisson reconstruction over a sparse point cloud [56], triangular mesh hole-filling using a moving least squares approach [57], and volumetric diffusion [58]. Modeling of the unknown underlying geometry of the cavities can also be performed with algebraic surfaces [59] or the fitting of Bezier patches [60] to the neighborhood surface points.

2.6. Summary

Human-body-model-based methods have huge potential and provide a great solution to attaining the complete shape of the desired pose. However, the resulting models are still not fully realistic, especially in the case of unusual body positions. In contrast, meshing and local shape methods are merely mathematical approximations, unaware of the human body's peculiarities. Furthermore, deep-learning-based methods require significant computing resources with a large amount of graphics processing unit (GPU) memory to reconstruct the models in high resolution.

3. Materials

For verification purposes, a dynamic fighting sequence was synchronously captured using 34 color (RGB) cameras in a 4096×2160 (4 K) pixel resolution. The cameras were calibrated and their positions in a global coordinate system were known. Figure 1 shows a 3D visualization of the measurement scene with a subject in the center and an even distribution of the cameras. On each of the 15 columns, 2 cameras were hung at heights ranging from 0.5 to 3 m. Additionally, 4 cameras were placed on the 3.5 m ceiling.

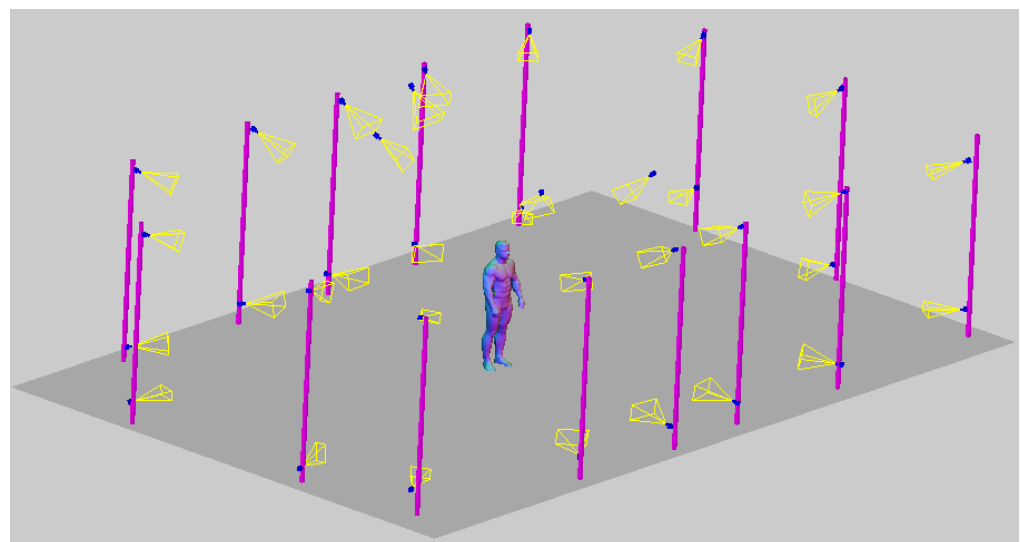






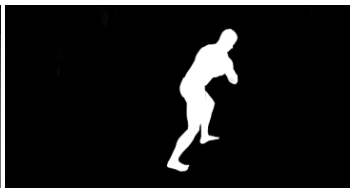



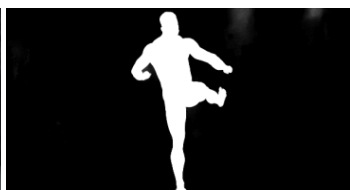
Figure 1. Visualized measurement scene with 34 camera distribution and sample-measured subject SfM point cloud.

In order to obtain the most accurate silhouette segmentation results on the 2D images, the Omnimatte pre-trained CNN model was utilized [61]. This takes as an input a rough mask of an object, and on the output returns precisely segmented images of the object and a background. To estimate the rough mask of the object, the Omnimatte utilizes a pre-trained DeepLabV3 model [62]. The chosen CNN solution provided satisfactory

results for silhouette estimation purposes. Nevertheless, it should be noted that the sSfS method is not restricted to it and, in the future, it can be replaced with a better silhouette estimation method.

As a reference, the photogrammetric reconstruction was performed using Agisoft Metashape software [25]. To ensure comparability, the same RGB images were used for both proposed and reference methods. The exemplary camera images and calculated masks from the validation dataset for 2 of 20 different subjects' poses are shown in Table 2 with respective ground-truth 3D reconstructions. As can be clearly seen, reconstruction defects are present in the form of holes. The complete test data collection can be seen in Appendix A, in Table A1.

Table 2. Sample test poses (2 out of the total of 20) used to validate the sSfS algorithm with corresponding RGB images, estimated silhouettes for 2 chosen cameras out of the set of 34, and photogrammetric reconstruction.

Sample RGB Images	Sample Silhouettes	Ground-Truth SfM Reconstruction
		
		
		
		

4. Methods

In this section, the details of our sSfS reconstruction method are described. First, the adaptation of the typical SfS visual hull estimation method is presented, along with the object volume estimation and voxel voting processes. Then, the sSfS method is detailed, including the novel step of human body segmentation on the silhouette images.

4.1. The Conventional SfS VISUAL Hull Reconstruction of the Entire Human Body

The implemented SfS reconstruction approach extracts the visual hull of the measured object in the voxel representation, similarly to the state-of-the-art SfS solutions [11,36]. The visual hull is based on creating the voxel grid space in the selected volume and estimating the voxels of the reconstruction with a projection of the voxel centers on the silhouettes. The visual hull estimation steps of the SfS algorithm are shown in Figure 2.

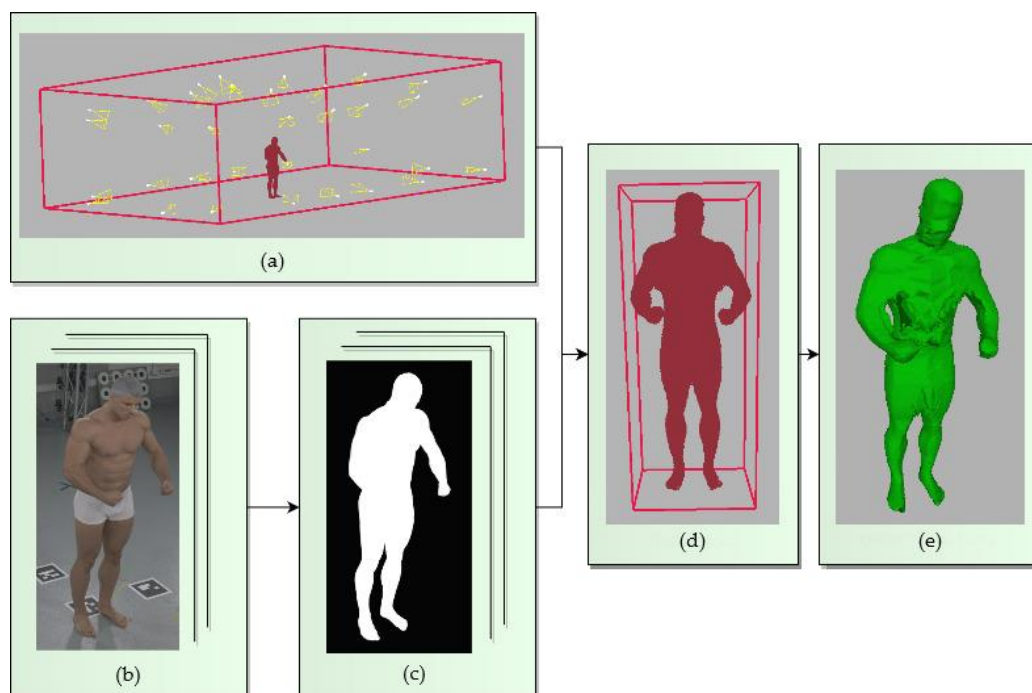


Figure 2. Steps involved in the conventional SfS reconstruction method of an object: (a) the volume of the whole reconstruction system, estimated with the cameras' positions; (b) input RGB images; (c) silhouette images of the subject; (d) approximate subject volume from the transition step (see Figure 3 for details); (e) final visual hull estimation with the target voxel size.

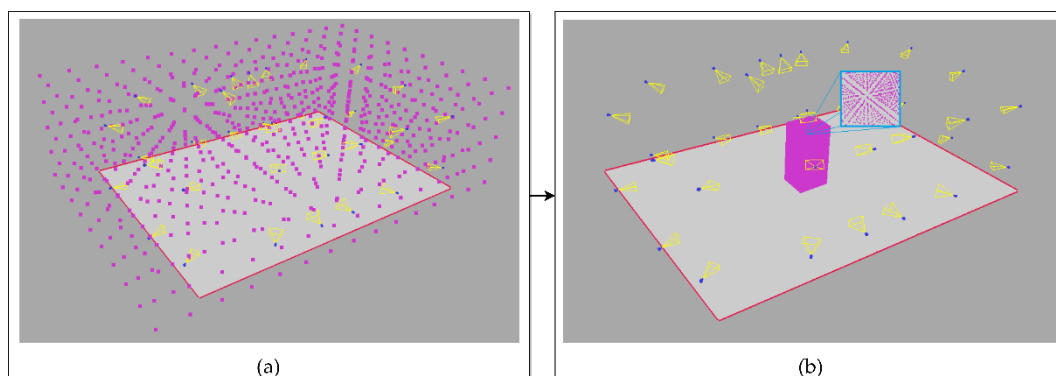


Figure 3. Estimation of the subject's volume: (a) the initial voxel grid with voxel size D_0 in the volume of the system cameras; (b) the voxel grid with voxel size D_1 of the subject's volume.

4.1.1. Calculation of the Subject's Volume

The output of the SfS reconstruction is a set of voxels with a known dimension D and center coordinates. Before the actual SfS reconstruction of the subject, the 3D volume of the cameras' bounding box is divided into a voxel grid with a given single voxel size D_0 . Then, a rough reconstruction of the subject is performed inside this volume to obtain its coarse dimensions. Finally, the reconstruction of the subject is carried out with a lower voxel size D_1 to obtain more accurate and dense results. The transition from the initial voxel size D_0 to the target D_1 is depicted in Figure 3. The voxel size transition allows one to solve the potential memory-overload issue, which may occur if all of the cameras' volume was filled with a dense grid of a voxel size D_1 .

4.1.2. Visual Hull Estimation with the Voxel Projections

For the silhouette estimation, we decided to utilize the state-of-the-art Omnimate CNN-based method [61], which separates objects of a specified class from the background, particularly humans. The silhouette image obtained using Omnimate net is not binary. Instead, it contains a quality parameter that defines the probability of a correct silhouette determination in each pixel encoded as an image intensity from 0 to 255. During the projection of voxels onto the silhouettes, the algorithm counts probabilities for voxels from all of the cameras. The sum is then compared with a threshold to determine the voxels of the visual hull. Figure 4 shows the process of the silhouette estimation and the projection of the voxels onto a silhouette. The magnified areas in the projection images in Figure 4 show a rough probability transition on the edge of the silhouette.

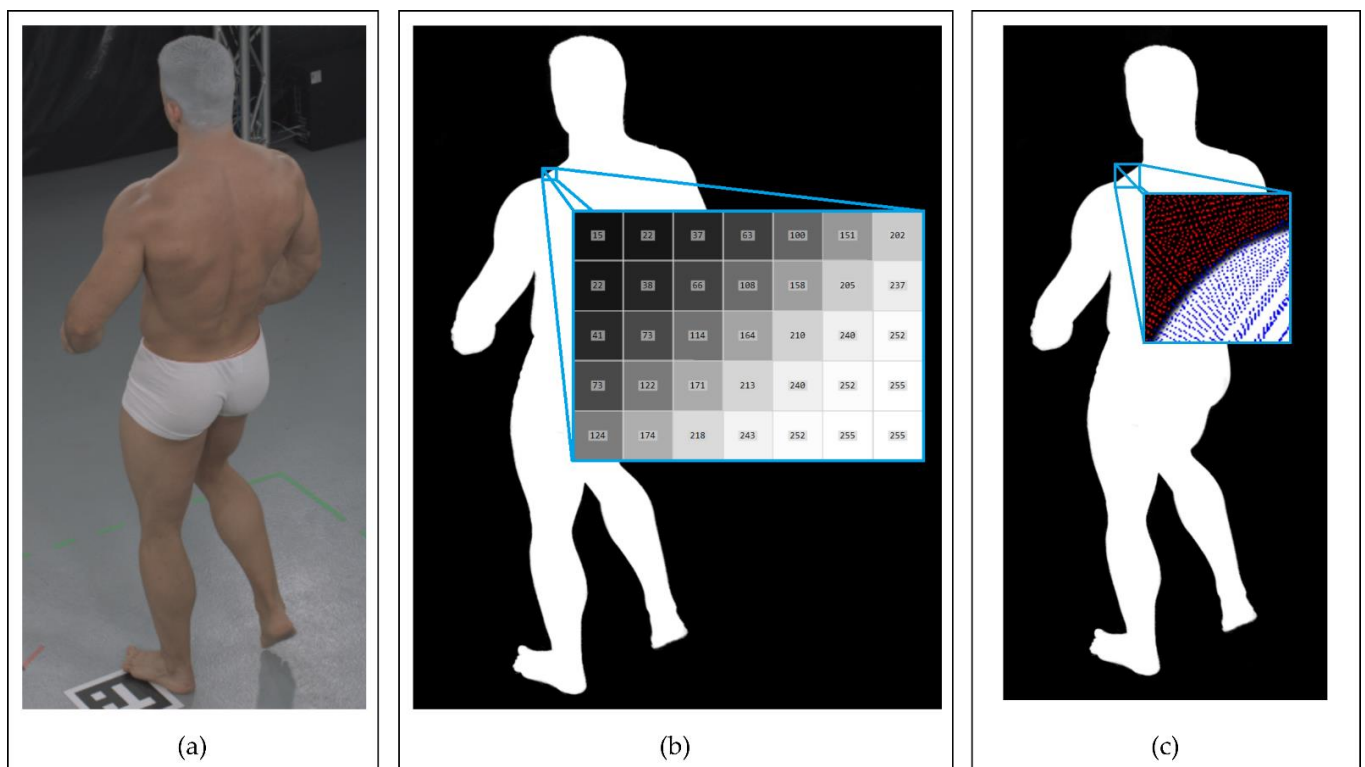


Figure 4. Example of Omnimate [61] silhouette estimation performance with voxel projections onto the image (data sample #9): (a) input RGB image; (b) silhouette image with magnified silhouette quality values on its edge; (c) voxel center projections onto the silhouette. The projections in the image are marked with pixel-sized dots. The blue pixels represent the silhouette, and the red represent the background.

4.2. sSfS Body Segment Reconstruction

The proposed coarse-to-fine sSfS method reconstructs the selected body parts separately and merges those partial results into one. Thus, each body part's visual hull estimation algorithm is the same as for the whole-body reconstruction method shown in Figure 2. However, to do this, the 2D human body segmentation of the silhouette images is required. Therefore, a custom method was implemented to map a 3D human body segmentation to 2D silhouettes. Nevertheless, this can also be achieved with several other methods, such as CNN body segmentation based on pose estimation from images [63,64].

To segment the body parts on silhouette images, the estimated positions of human joints obtained from the Human Pose CNN [65] on RGB images of the subject are used. Then, each of the approximate joint positions is found by means of ray casting from each corresponding 2D joint position to estimate the 3D joint position as the closest point to all 3D ray lines. As a result, the positions of the joints in a 3D space of cameras in the

system can be estimated, as shown in Figure 5c. Next, the 3D joints are used to segment the visual hull reconstruction by assigning each voxel to the closest bone (defined as a section between two joints). However, this step could also be performed with other human body 3D segmentation approaches [66,67]. The result of the proposed 3D segmentation is shown in Figure 5d. Finally, visual hull voxels from each segment are projected onto the silhouette images to obtain segmented silhouettes, as can be seen in Figure 5e. The detailed silhouette estimation process for the segments is shown in Figure 6, with an example of head segment detection.

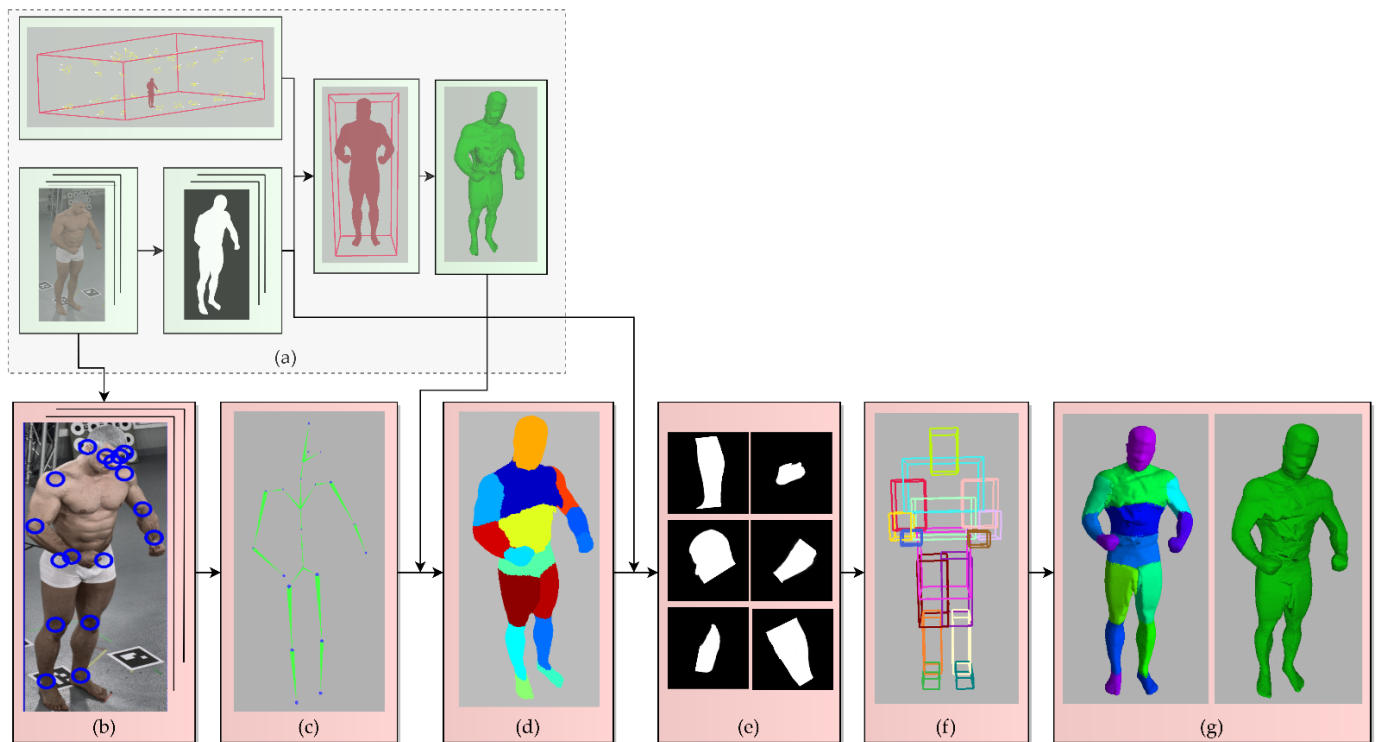


Figure 5. Proposed sSfS reconstruction approach flowchart. The steps added to the conventional SfS approach are highlighted in pink: (a) conventional visual hull estimation with silhouette images of an entire subject’s body (see Figure 3 for details); (b) results of the estimation of human joint positions on the 2D color images with a CNN-based Human Pose pre-trained model [65]; (c) retrieval of the 3D joint positions by casting the rays leading from each camera center’s 3D position to each joint and calculating the best intersections; (d) segmentation of the subject’s coarse 3D visual hull voxel reconstruction; (e) silhouette image segmentation by projecting the segment points onto the silhouettes (see Figure 6 for details); (f) estimation of the 3D volume of each body segment; (g) SfS reconstruction results for each body part separately and merged as a whole sSfS body model.

After the silhouette body part segmentation, each segment is reconstructed according to the conventional SfS algorithm, using the same voxel vote threshold value for the reconstruction of each body segment. Finally, the individual voxel models of body parts are merged and form the final sSfS reconstruction (Figure 5g).

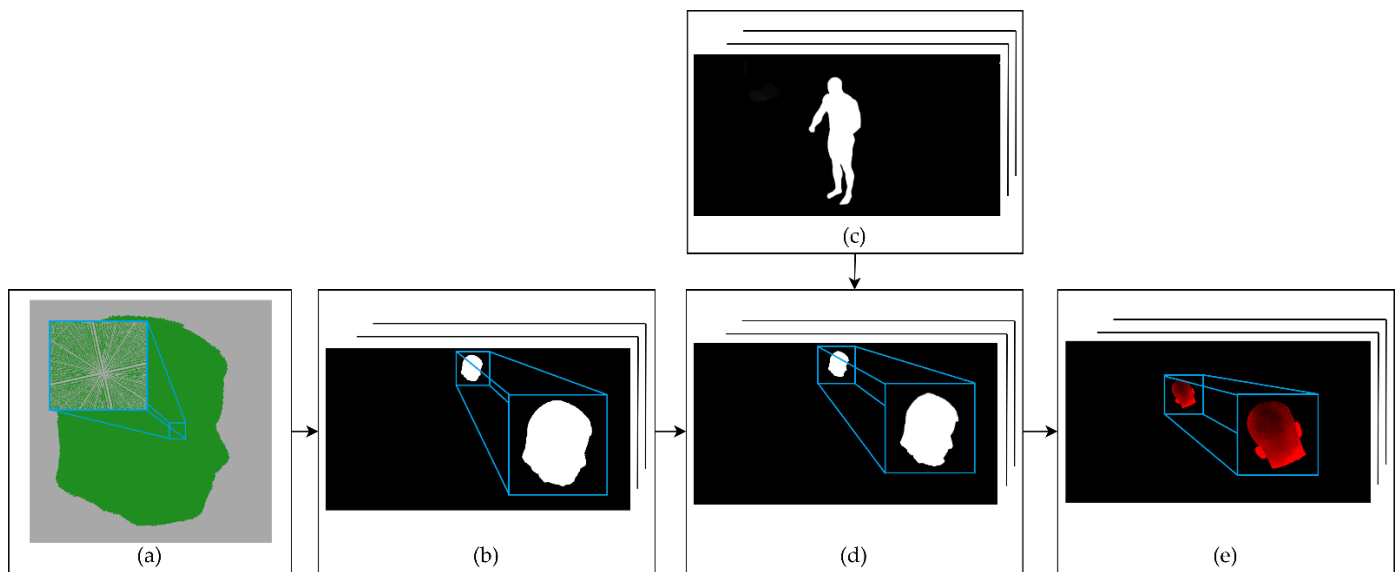


Figure 6. The estimation of the body segment silhouettes and their volumetric reconstructions. (a) Segment from the initial coarse volumetric reconstruction step with a 4 mm voxel size. (b) Projection of the voxels from the segment reconstruction onto the camera images for all the system's cameras. The pixels where the projection was performed are changed to a value of 255. (c) Input silhouettes of the subject's entire body from a set used to estimate an initial visual hull reconstruction. (d) Body segment detection for all silhouette images. The projection from step (b) is used as a mask for the segment on the silhouette image. (e) Detection of erroneous silhouettes by calculating the ratio of the segment projection pixels to the uncertain silhouette pixels, with values in the range $<1, 254>$.

As a result of the silhouette segmentation, it is possible to perform the reconstruction for each recognized body segment separately. This method allows us to find the errors in the silhouette estimation and skip the views for poorly detected body segments. To identify the erroneous segments, the ratio of the silhouette's uncertain pixels number (pixels with intensity values in the range $<1, 254>$) to the segment projection pixels number (pixels with intensity value 255) is calculated. The projection image refers to Figure 6b, and the silhouette segment image refers to Figure 6d. The uncertain silhouette pixels ratio calculation is described with the following equation:

$$F = \frac{M}{N} \quad (1)$$

where:

F uncertain pixels ratio,

M number of pixels in the silhouette segment image in range $<1, 254>$,

N number of pixels in the segment projection image with intensity 255.

By applying a simple F_t threshold value for each silhouette segment image, only the high-quality segments of the body silhouette are obtained for reconstruction, thus increasing the final visual hull's quality.

4.3. Hardware and Software Environment

The SfS reconstruction algorithms [11] were implemented in C++, and the interface for the CNN silhouette [61] and 2D joint position estimations [65] was implemented in Python. For the method's evaluation, we used a computer with a $2 \times$ Intel(R) Xeon(R) Silver 4215 processor @ 2.5GHz, $2 \times$ NVIDIA RTX 2080Ti and 512GB RAM.

5. Results

This section presents the results of the SfS and sSfS reconstructions for selected data samples compared with the SfM reference. Additionally, the difference between the SfS and sSfS outputs is presented using quality metrics. Both SfS and sSfS reconstructions were calculated with voxel sizes D_0 64 mm and D_1 4 mm for the dataset of 20 samples described in Section 3. The conventional SfS was reconstructed according to the algorithm steps in Figure 2. Then, this coarse SfS reconstruction was used in the sSfS phase, as shown in Figure 5.

5.1. Surface Shape Comparison

A visual comparison between the SfS and sSfS reconstruction surfaces is shown in Figure 7. The surface voxel of the visual hull was found using morphological erosion [68] (every eroded voxel was considered a surface voxel). The results of both methods were compared to the ground-truth SfM point cloud to assess their accuracy and their ability to fill the holes. In contrast to the SfM reference, SfS and sSfS provided results without significant distortions on the head and legs. For example, in data samples #2 and #4 in Figure 7, the SfS and sSfS correctly reconstructed the top head area, whereas the ground-truth SfM failed to do so. Similarly, the missing surface on the legs in the SfM reconstruction can be recovered in SfS and sSfS, as in data samples #3–#6 in Figure 7. In addition, the sSfS reconstruction surface was less noisy and rough than that of SfS, for example, on the chest in data sample #1 in Figure 7. The abovementioned reconstruction results for all 20 of the test poses are attached in Appendix A, Figure A1.

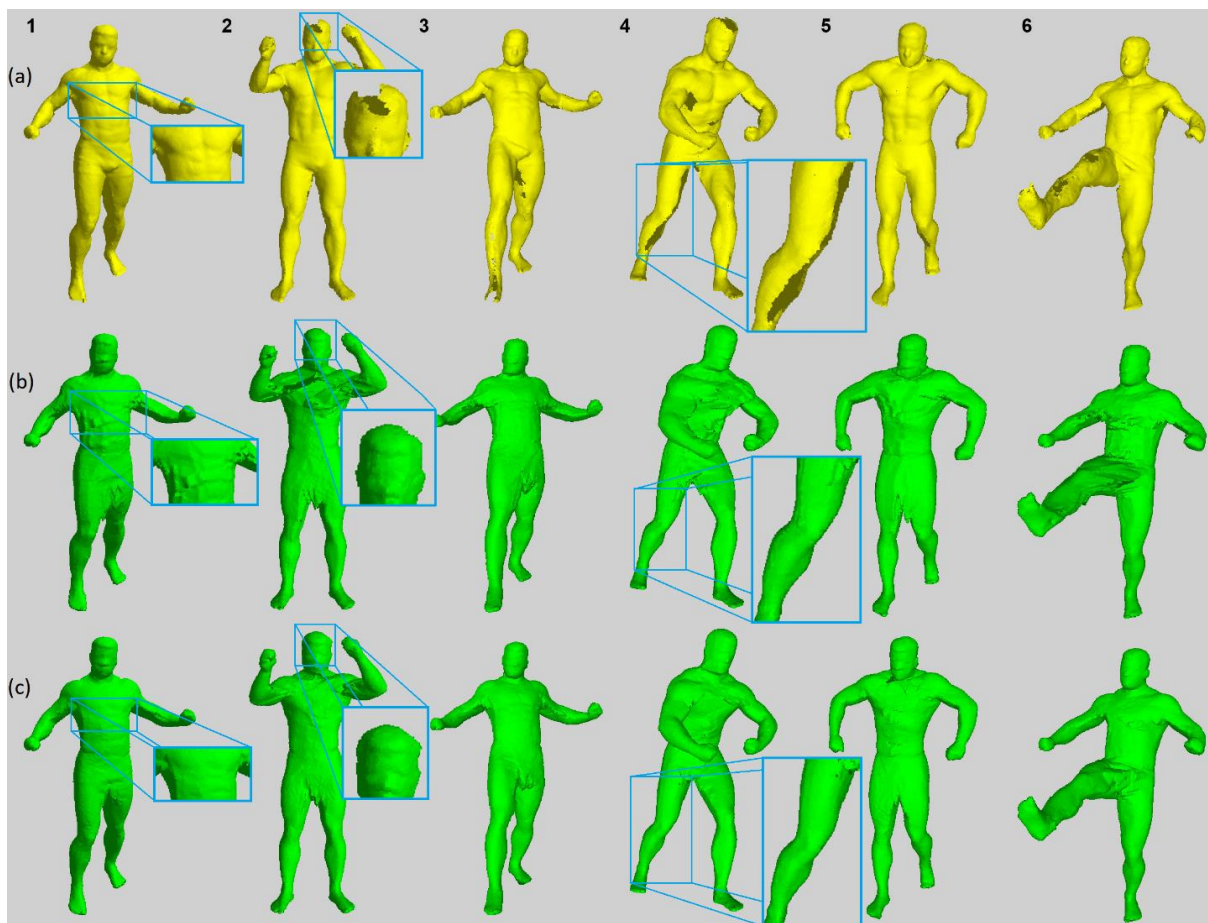


Figure 7. Surface reconstruction results comparison for selected dataset samples; every column ID number corresponds to a data sample from Table A1: (a) SfM; (b) SfS; (c) sSfS.

5.2. Metric Description and Results

Further quantitative analysis of the surfaces from the novel SfS and conventional sSfS was carried out in reference to the photogrammetric point clouds using two metrics:

- The number of erroneous visual hull voxels outside of the reference surface—For each voxel center, the closest point of the SfM reconstruction was found. Then, the dot product A between the vector from the voxel center to the SfM point and the SfM point's normal vector was estimated. The voxel was considered erroneous when the value of the dot product A was less than 0. Additionally, when the distance between the voxel center and SfM point was less than the voxel size (4 mm), the voxel was not counted as erroneous.
- The distance between the SfS/sSfS reconstruction voxel and the reference cloud's surface (point to surface, P2S)—The Euclidean 3D distance between the voxel center on the surface of the SfS/sSfS reconstruction and the closest SfM point.

5.2.1. The Number of Erroneous Voxels

The number of erroneous voxels provides information about the visual hull performance on the concave parts of the subject's pose. The subject's placement in the scene has a high impact on the number of voxels outside the reference surface, which determines the system cameras' capability to extract the visual hull. However, this metric can show the robustness of the visual hull estimation in relation to the body part occlusions. The erroneous voxel metric results for the SfS and sSfS reconstructions for the entire dataset are presented in Figure 8. It can be seen that, in some cases, our novel sSfS reconstruction contained about two times fewer erroneous voxels than the conventional SfS results. The approximate mean number of invalid voxels for the entire validation dataset in the SfS method was 285,000, with a standard deviation of 35,000 voxels. By comparison, the average number of invalid voxels in sSfS was 168,000, with a standard deviation of 25,800 voxels. The division of those mean values shows that the sSfS allowed us to obtain 1.7-fold better results for the erroneous voxels metric.

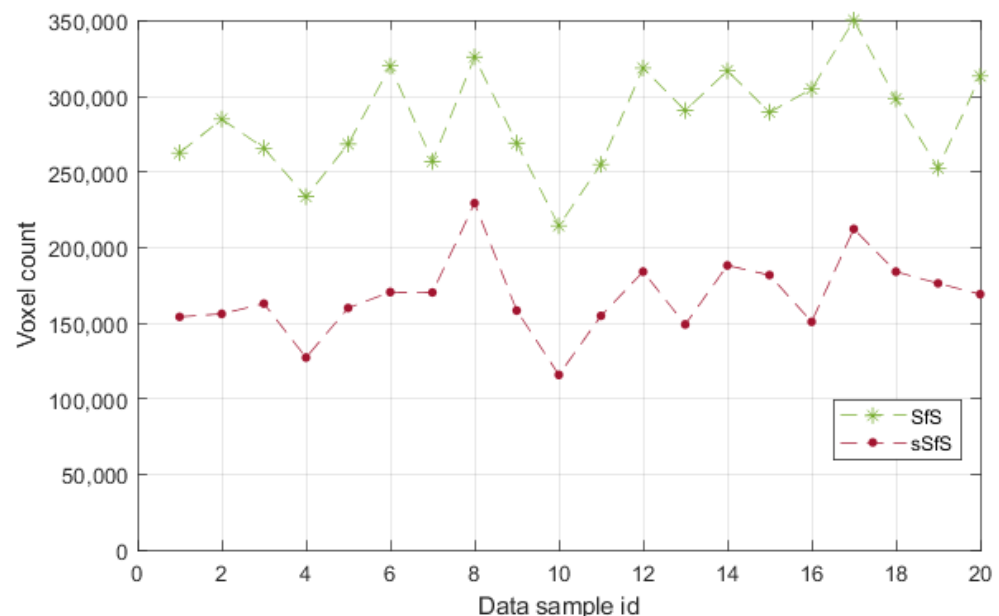


Figure 8. The erroneous voxel counts for SfS and sSfS reconstructions of the validation dataset.

5.2.2. Surface Voxels' Distance to the Reference

Calculating the distances of the surface voxels in relation to the reference surface allows the visualization of the visual hull error and tracking of the most problematic human body areas to reconstruct. Figure 9 shows the examples of the distance-to-surface

error for both the SfS and sSfS reconstructions (the rest of the data samples' results are shown in Appendix A, in Figure A2). In the visualization, it can be observed that the most problematic pose scenarios were observed mostly for occluded poses, especially when the limbs were near the torso area, in which both tested reconstruction approaches showed errors. Even though the results were imperfect for both approaches, the advantage of the sSfS approach is still clearly visible. Furthermore, the sSfS method achieved better results over the entire body surface. The advantage is primarily visible on the most challenging body parts, such as the armpit and crotch (magnified examples shown in Figure 9). High reconstruction quality on these body parts is crucial because the crotch and armpits are examples of the most problematic spots for SfM reconstructions.

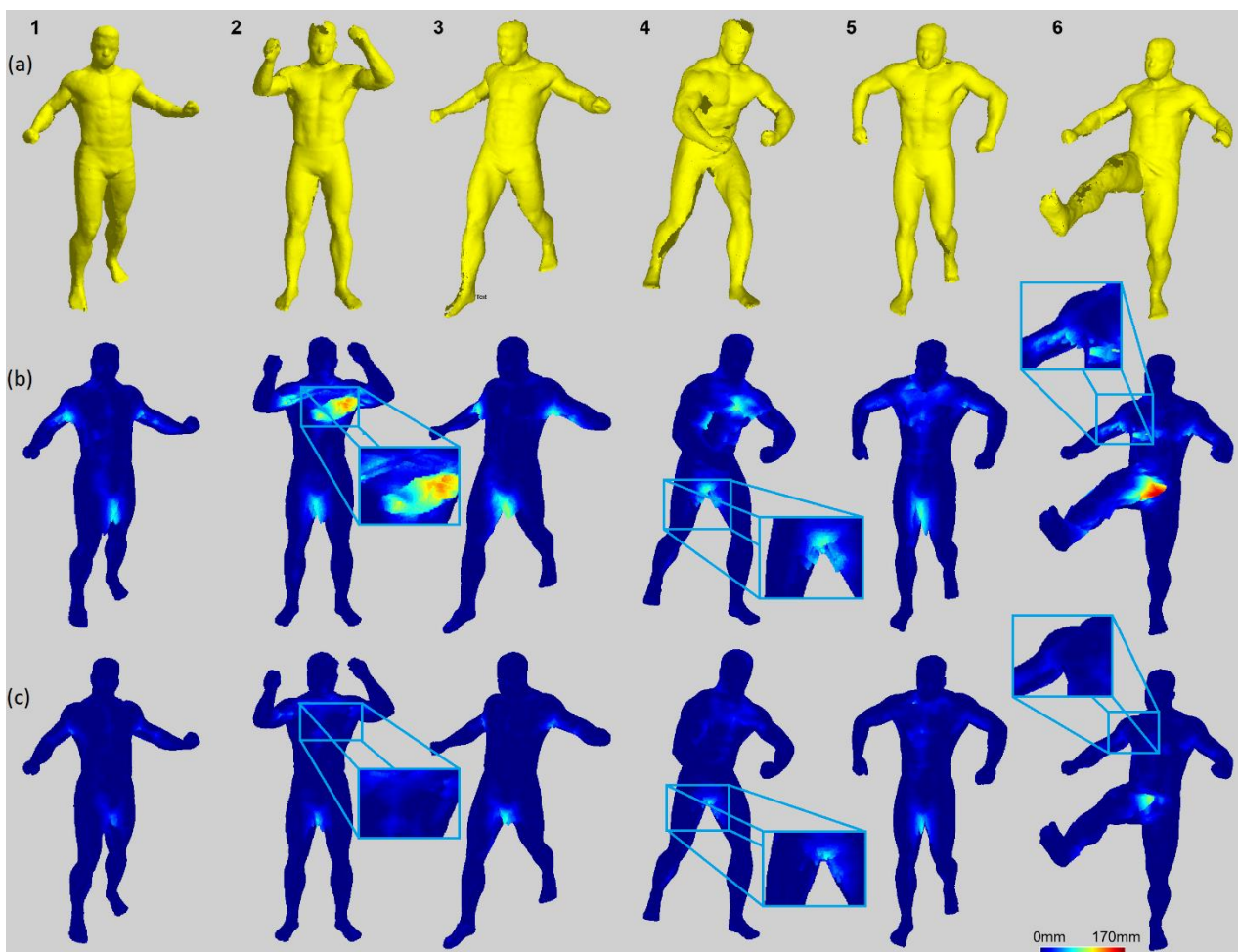


Figure 9. Visualization of P2S distances for SfS and sSfS results: (a) SfM; (b) SfS with P2S error map; (c) sSfS with P2S error map. Each data sample in the column corresponds to a data sample from Table A1; the numbers at the top of each column relate to the data sample ID.

In addition, histograms of the distances of the visual hull surface voxels to the reference surface were calculated. The example of the overlapping histograms of SfS and sSfS P2S in Figure 10a demonstrates that sSfS was characterized by more points with P2S distances up to 6 mm, in contrast to SfS, which contained more voxels with higher distance errors. Moreover, SfS had a high number of voxels with a P2S distance of more than 40 mm, which is also visible in Figure 9. Furthermore, the root-mean-square (RMS) of the P2S distances plot (Figure 10b) shows an advantage of the sSfS method on every data sample. The mean RMS of the P2S error for the SfS method was 3.62 mm, with a standard deviation of 0.34 mm for the entire validation dataset. For the sSfS method, the mean RMS was 2.92 mm, with a

standard deviation of 0.32 mm. The RMS of the P2S error for the validation dataset was 1.22-fold better in the sSfS method.

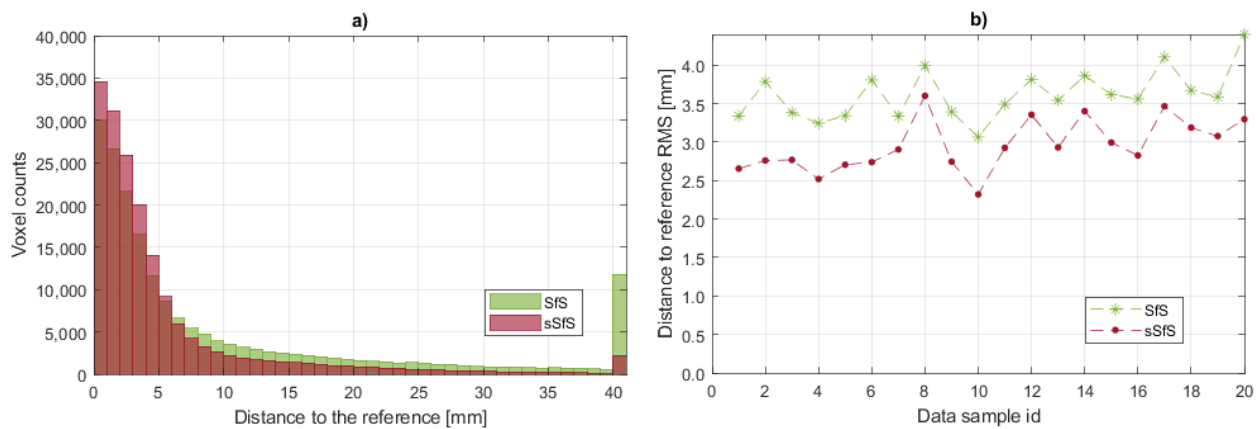


Figure 10. P2S error histograms: (a) overlapping SfS and sSfS histograms for dataset sample #6. The x -axis represents the histogram bin ranges with a bin of 1 mm. The last bin contains the sum of all samples with a distance >40 mm; (b) RMS of the P2S distances for SfS and sSfS reconstructions for the entire validation dataset.

5.3. Computation Time Comparison

The computation time of the implemented SfS reconstruction was around 30 s for a single pose, including silhouettes' estimation. In comparison, the sSfS reconstruction took around 250 s for a single pose, involving the joint estimation and reconstruction with the coarse SfS visual hull segmentation.

6. Discussion

The complete reconstruction of a subject's body with SfS is not a trivial task, as it is prone to silhouette estimation errors. Using a simple fixed voxel vote threshold is insufficient to solve the erroneous silhouette issue. If the vote threshold is high, accurately estimated body parts are reconstructed with high quality. However, if estimation is uncertain, some body parts may not be included in the reconstruction model, as they would not accumulate enough votes. Conversely, a lower voxel vote threshold value can yield superfluous artefacts on the reconstructed surface. Another approach is to skip views containing a significant number of uncertain pixels, for example, by analyzing uncertain pixels in the silhouette. In this manner, the reconstruction quality of the body parts deteriorates, but the body shape remains undistorted. The impact of the voting threshold and the view selection on SfS in terms of reconstruction quality and surface noise is shown in Figure 11. The SfS results are also compared to those of the sSfS approach, which demonstrated the best quality regardless of the voting threshold.

Examples of silhouette estimation issues are presented in Figure 12, which shows two silhouettes containing significant distortions. The problem with uncertain silhouette estimation appears mostly in places with a low color difference between the subject's body and the background or around the overlapping regions of body parts. However, Figure 12 shows that correctly estimated silhouette segments could be used at least partially in the sSfS reconstruction process per segment and thus increase the overall quality. For example, Figure 12a shows that the entire right arm's silhouette is well estimated. In Figure 12b, only the silhouette part on the feet is distorted, and the rest of the silhouette can be used for sSfS reconstruction.

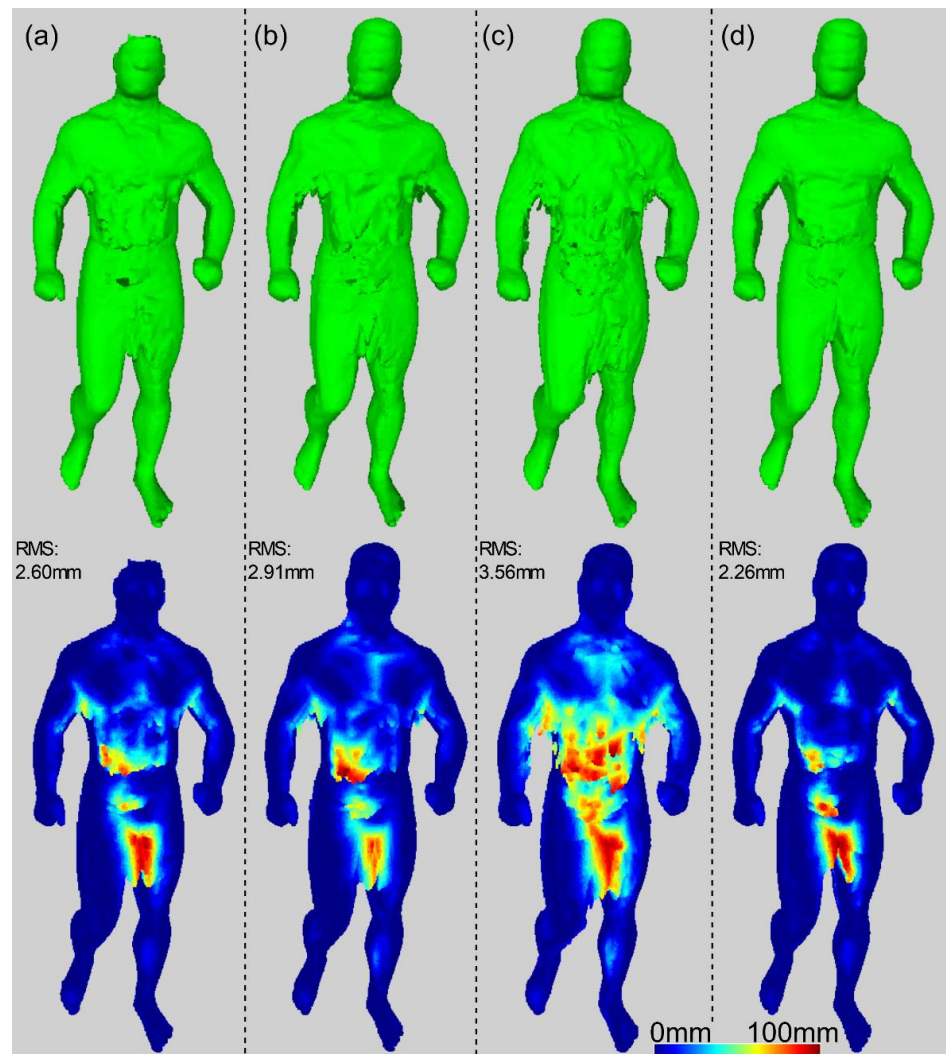


Figure 11. Comparison of the results of three different SfS reconstruction approaches and sSfS for data sample #9: (a) SfS for a high voting threshold for all views; (b) SfS for a high voting threshold for selected views; (c) SfS for a low voting threshold for all views; (d) sSfS.

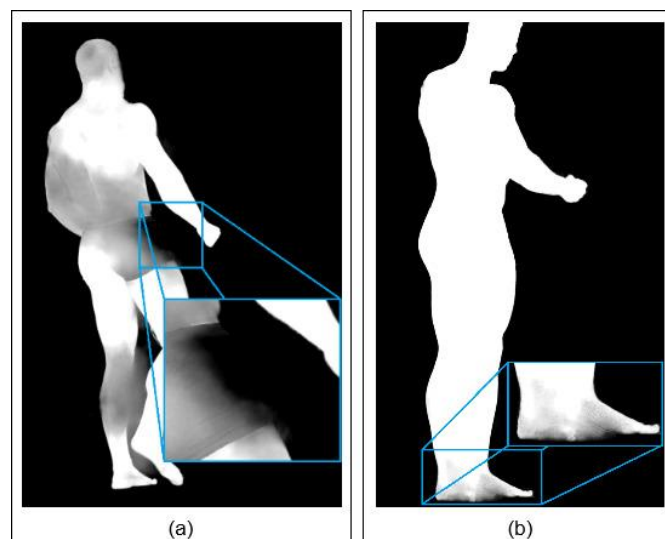


Figure 12. Examples of erroneous silhouette areas: (a) torso, left leg, and head; (b) feet.

Despite having significantly better reconstruction quality than the conventional SfS reconstruction, the sSfS method also has some weaknesses. A primary disadvantage of the sSfS approach is its dependence on the coarse 3D body reconstruction and the subsequent segmentation. The 3D body segmentation can be erroneous in the case of noisy data or complex body poses. The sSfS method provides fewer noise voxels, but is also less universal due to the need for silhouette segmentation. By comparison, the current version of the sSfS reconstruction contains manually adjusted thresholds for voxel voting and view skipping, specifically for the example dataset. The use of manual thresholds may be an issue for larger datasets with irregular lighting or different numbers of camera views. Furthermore, the sSfS computation time can be improved by adjusting algorithms to parallel computing.

7. Conclusions

Detailed 3D reconstruction of the human body is a challenging task, and many possible areas exist for improvement in the existing solutions. This paper proposes a robust sSfS method that significantly enhances the conventional SfS reconstruction quality. It utilizes 2D body extraction algorithms to achieve high-quality silhouettes, in addition to human body segmentation. Performing 2D body segmentation on silhouettes enabled the introduction of the segmented Shape from Silhouette (sSfS) approach, which demonstrated robustness for silhouette estimation distortions and showed better reconstruction quality than the conventional SfS on every tested data sample. Furthermore, it was shown that sSfS can be used for the filling of cavities in the human body point cloud obtained through another reconstruction method.

Despite the promising experimental results presented in this paper, further work is required to improve the sSfS results. For example, the sSfS is dependent on the 3D segmented coarse reconstruction, which determines the quality of the segmentation of silhouettes. Therefore, the coarse SfS reconstruction method that is an input for the sSfS segmentation step can be replaced with some other reconstruction method, such as a multi-view parametric human model-based method [46] or one of the CNN-based 3D reconstruction approaches [54].

Moreover, the rapid evolution of 2D segmentation algorithms may eliminate the need for initial 3D segmentation from our pipeline. Furthermore, the development of 2D segmentation solutions may allow the sSfS to improve the reconstruction of other object classes. Another means to improve the current sSfS approach is to develop a better silhouette estimation algorithm by replacing the current method or utilizing efficient silhouette postprocessing.

Author Contributions: Conceptualization, W.K. and R.S.; methodology, R.S.; software, W.K. and Ł.M.; validation, W.K., Ł.M.; formal analysis, R.S.; investigation, W.K.; resources, W.K.; data curation, W.K.; writing—original draft preparation, W.K.; writing—review and editing, W.K. and Ł.M.; visualization, W.K.; supervision, R.S.; project administration, R.S.; funding acquisition, R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by grant number “POIR.01.01.01-00-0510/19-00” by the National Center for Research and Development (Poland).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Sample RGB images with estimated silhouettes for 34 cameras and SfM reconstructions used to validate the SfS/sSfS reconstruction. The dataset contains image sets and 3D scans of 20 different poses of the subject.


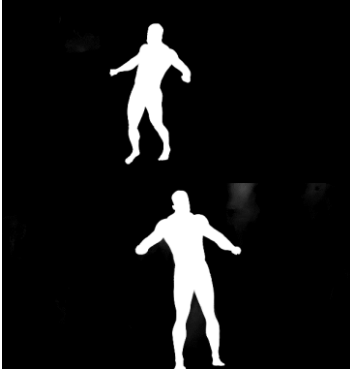
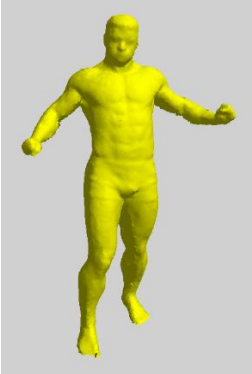
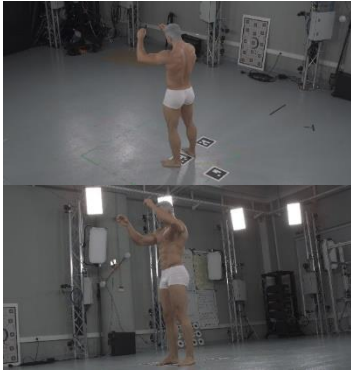
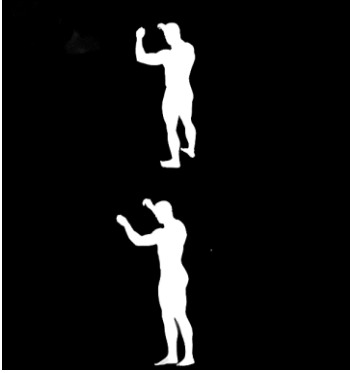



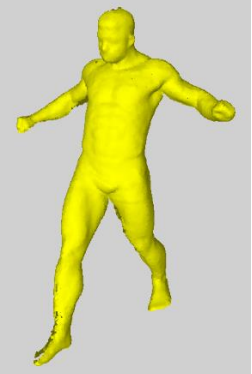

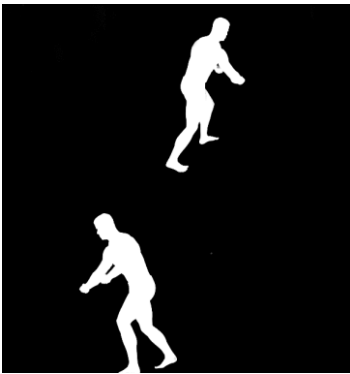
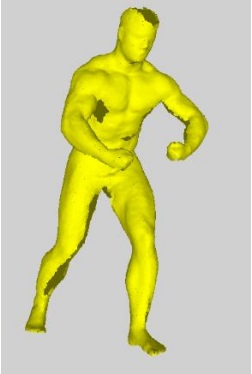
Id.	Sample RGB Images	Sample Silhouettes	SfM Ground-Truth Reconstruction
1			
2			
3			
4			

Table A1. Cont.

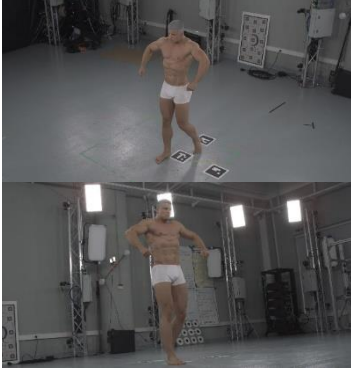
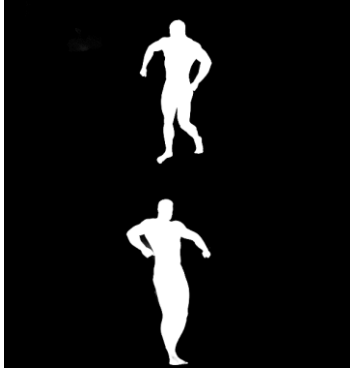
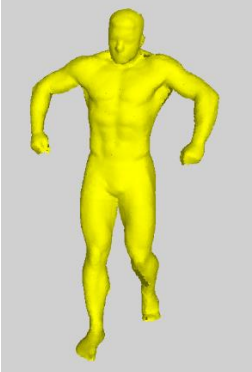
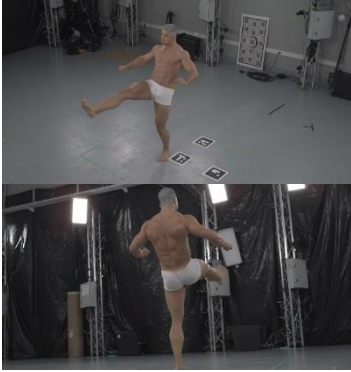

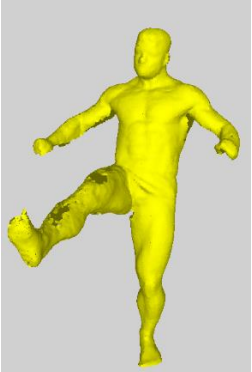
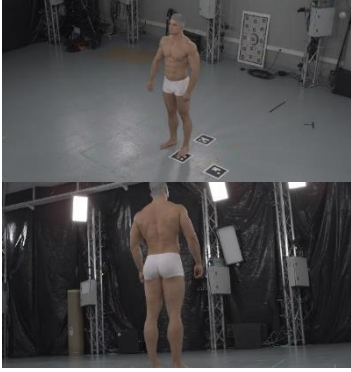
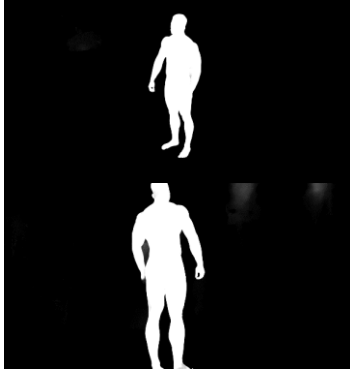
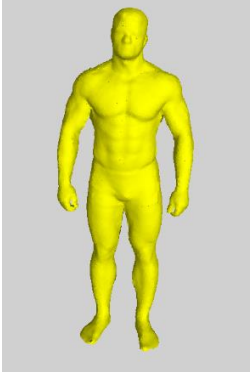
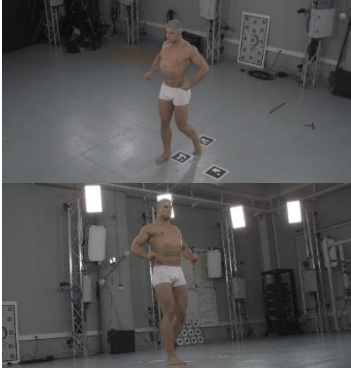

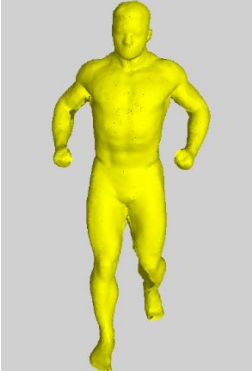
Id.	Sample RGB Images	Sample Silhouettes	SfM Ground-Truth Reconstruction
5			
6			
7			
8			

Table A1. Cont.


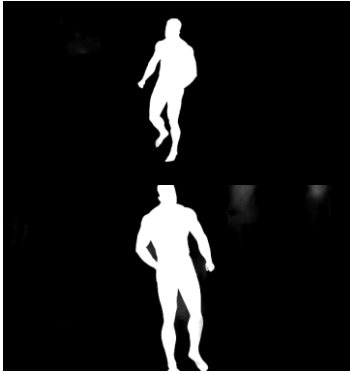
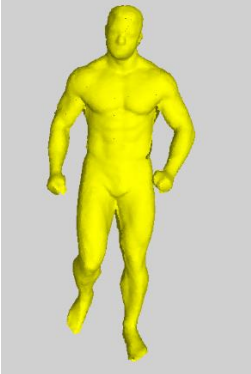

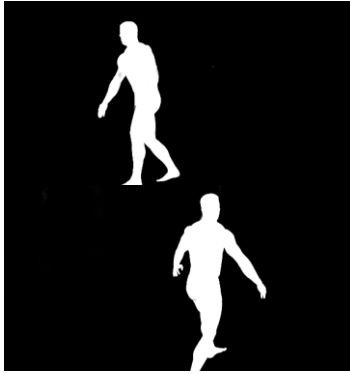
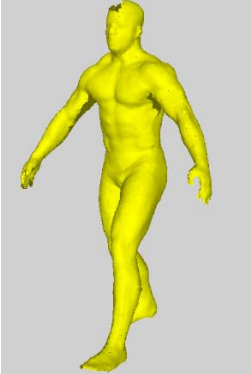
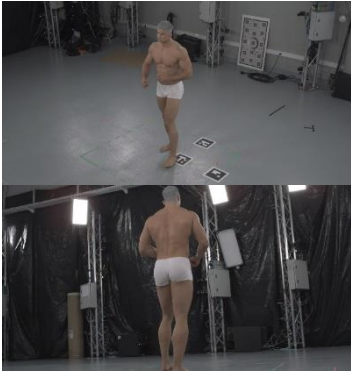



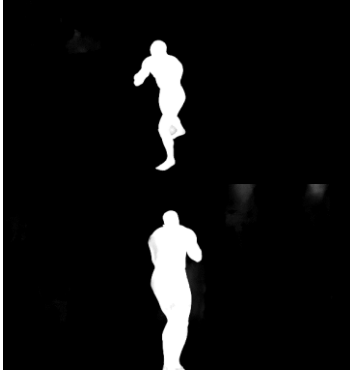

Id.	Sample RGB Images	Sample Silhouettes	SfM Ground-Truth Reconstruction
9			
10			
11			
12			

Table A1. Cont.

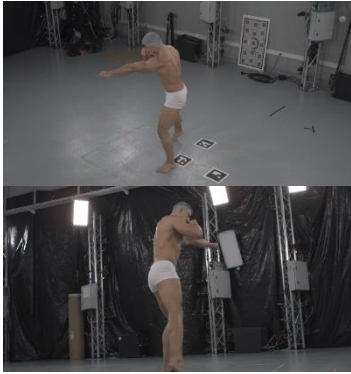
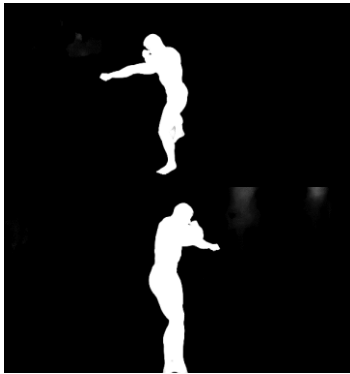

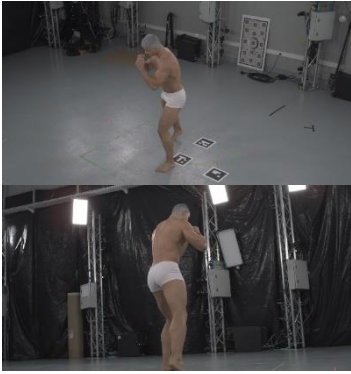
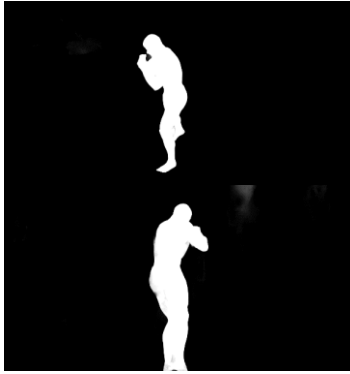
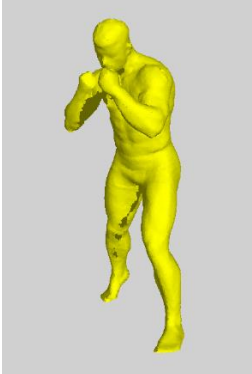

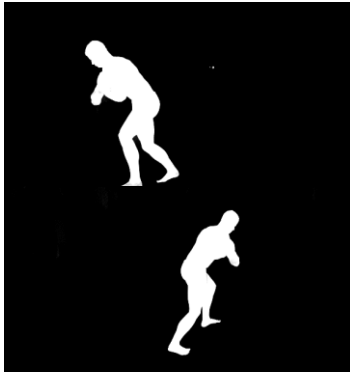

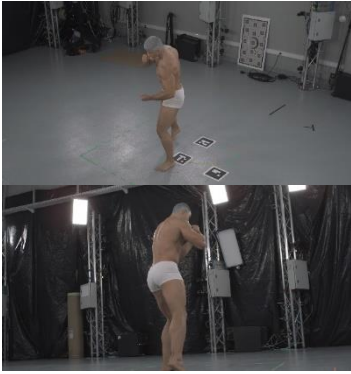
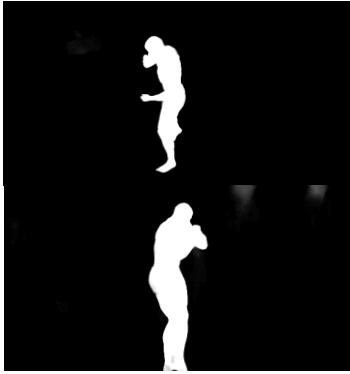

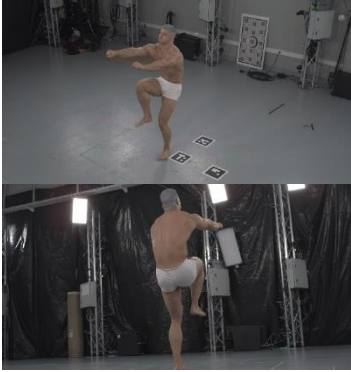
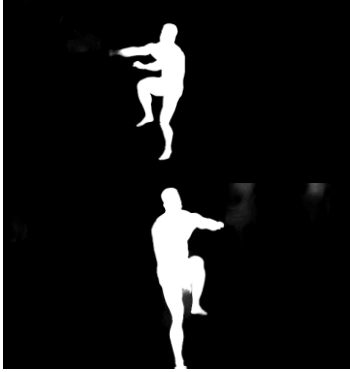

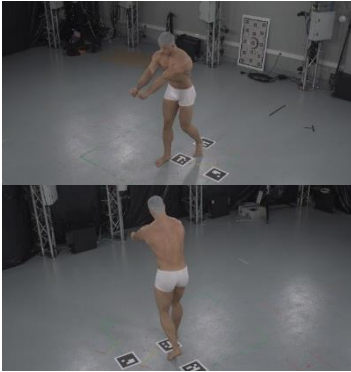
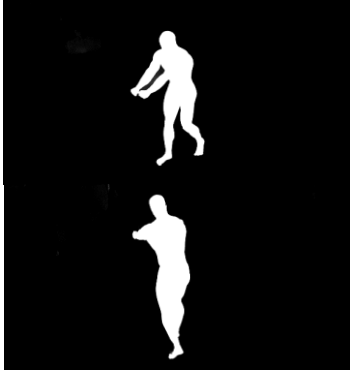
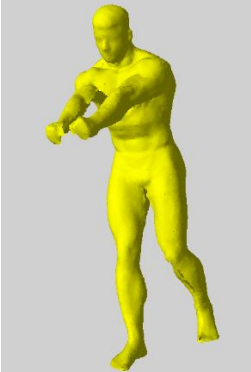

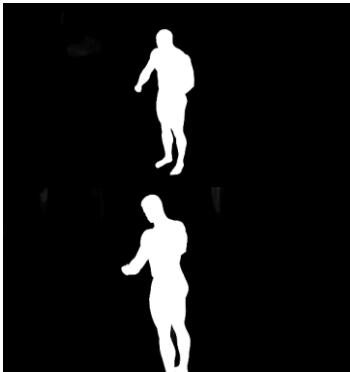
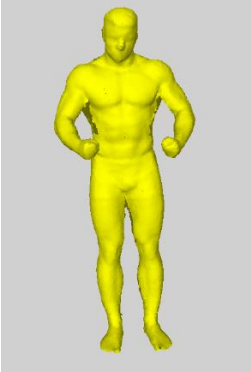

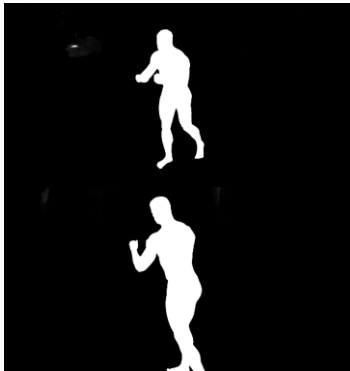

Id.	Sample RGB Images	Sample Silhouettes	SfM Ground-Truth Reconstruction
13			
14			
15			
16			

Table A1. Cont.

Id.	Sample RGB Images	Sample Silhouettes	SfM Ground-Truth Reconstruction
17			
18			
19			
20			

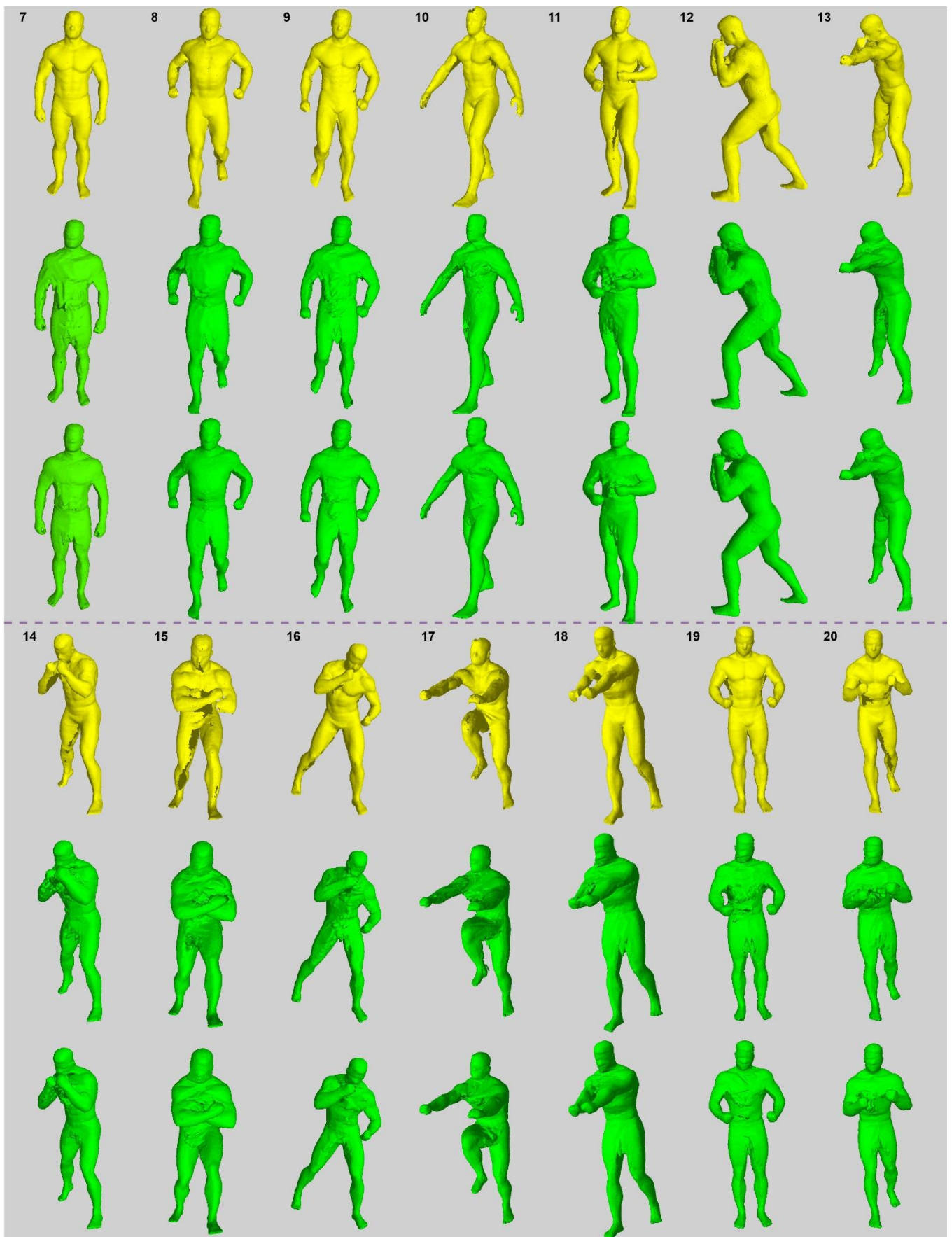


Figure A1. Visualization of the SfS/sfS reconstructions for data samples #7–#20.

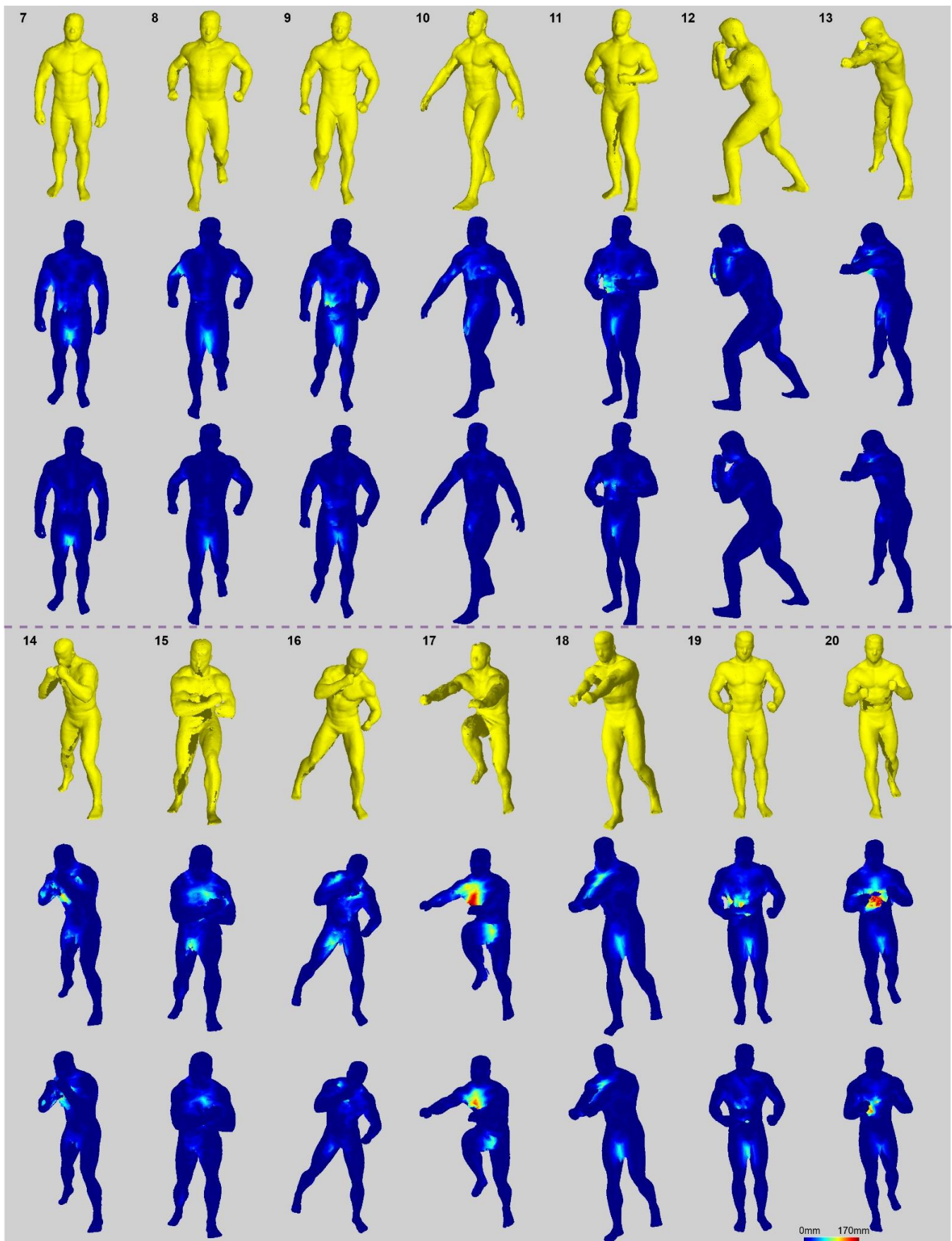


Figure A2. Visualization of the P2S distances for SfS/sSfS reconstructions for data samples #7–#20.

References

1. Gipsman, A.; Rauschert, L.; Daneshvar, M.; Knott, P. Evaluating the Reproducibility of Motion Analysis Scanning of the Spine during Walking. *Adv. Med.* **2014**, *2014*, 721829. [[CrossRef](#)] [[PubMed](#)]
2. Betsch, M.; Wild, M.; Johnstone, B.; Jungbluth, P.; Hakimi, M.; Kühlmann, B.; Rapp, W. Evaluation of a Novel Spine and Surface Topography System for Dynamic Spinal Curvature Analysis during Gait. *PLoS ONE* **2013**, *8*, e70581. [[CrossRef](#)] [[PubMed](#)]
3. Pons-Moll, G.; Romero, J.; Mahmood, N.; Black, M.J. Dyna: A model of dynamic human shape in motion. *ACM Trans. Graph.* **2015**, *34*, 1–14. [[CrossRef](#)]
4. Zhang, C.; Pujades, S.; Black, M.; Pons-Moll, G. Detailed, Accurate, Human Shape Estimation from Clothed 3D Scan Sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
5. Available online: <https://www.microsoft.com/en-us/mixed-reality/capture-studios> (accessed on 1 December 2021).
6. Available online: <https://scanable.com/mobile-photogrammetry/> (accessed on 1 December 2021).
7. Michoński, J.; Glinkowski, W.; Witkowski, M.; Sitnik, R. Automatic recognition of surface landmarks of anatomical structures of back and posture. *J. Biomed. Opt.* **2012**, *17*, 056015. [[CrossRef](#)]
8. Liberadzki, P.; Markiewicz, L.; Witkowski, M.; Sitnik, R. Novel 4D Whole Body Scanning Solution and its Medical Application. In Proceedings of the 9th International Conference and Exhibition on 3D Body Scanning and Processing Technologies, Lugano, Switzerland, 16–17 October 2018.
9. Treleaven, P.; Wells, J. 3D Body Scanning and Healthcare Applications. *Computer* **2007**, *40*, 28–34. [[CrossRef](#)]
10. Markiewicz, L.; Witkowski, M.; Sitnik, R.; Mielicka, E. 3D anthropometric algorithms for the estimation of measurements required for specialized garment design. *Expert Syst. Appl.* **2017**, *85*, 366–385. [[CrossRef](#)]
11. Cheung, K.-M.; Baker, S.; Kanade, T. Shape-From-Silhouette across Time Part I: Theory and Algorithms. *Int. J. Comput. Vis.* **2005**, *62*, 221–247. [[CrossRef](#)]
12. Daanen, H.A.M.; Haar, F.B.T. 3D whole body scanners revisited. *Displays* **2013**, *34*, 270–275. [[CrossRef](#)]
13. Ebrahim, M. 3D Laser Scanners: History, Applications and Future. 2014. Available online: https://www.researchgate.net/profile/Mostafa-Ebrahim-3/publication/267037683_3D_LASER_SCANNERS_HISTORY_APPLICATIONS_AND_FUTURE/links/5442bdf10cf2e6f0c0f93727/3D-LASER-SCANNERS-HISTORY-APPLICATIONS-AND-FUTURE.pdf (accessed on 1 December 2021).
14. Foix, S.; Alenya, G.; Torras, C. Lock-in Time-of-Flight (ToF) Cameras: A Survey. *IEEE Sens. J.* **2011**, *11*, 1917–1926. [[CrossRef](#)]
15. Salvi, J.; Fernandez, S.; Pribanic, T.; Llado, X. A state of the art in structured light patterns for surface profilometry. *Pattern Recognit.* **2010**, *43*, 2666–2680. [[CrossRef](#)]
16. James, M.R.; Robson, S. Straightforward reconstruction of 3D surfaces and topography with a camera: Accuracy and geoscience application. *J. Geophys. Res. Earth Surf.* **2012**, *117*. [[CrossRef](#)]
17. Available online: <https://www.vitronic.com/en-us/3d-bodyscan/scanner-for-performance-diagnostics> (accessed on 1 December 2021).
18. D’Apuzzo, N. 3D body scanning technology for fashion and apparel industry. In Proceedings of the IS&T/SPIE Electronic Imaging 2007, San Jose, CA, USA, 28 January–1 February 2007.
19. Pribanic, T.; Petkovic, T.; Bojanic, D.; Bartol, K. Smart Time-Multiplexing of Quads Solves the Multicamera Interference Problem. In Proceedings of the 2020 International Conference on 3D Vision (3DV), Fukuoka, Japan, 25–28 November 2020.
20. Jeught, S.V.; Dirckx, J.J.J. Real-time structured light profilometry: A review. *Opt. Lasers Eng.* **2016**, *87*, 18–31. [[CrossRef](#)]
21. Liberadzki, P.; Adamczyk, M.; Witkowski, M.; Sitnik, R. Structured-Light-Based System for Shape Measurement of the Human Body in Motion. *Sensors* **2018**, *18*, 2827. [[CrossRef](#)] [[PubMed](#)]
22. Bartol, K.; Bojanic, D.; Petkovic, T.; Pribanic, T. A Review of Body Measurement Using 3D Scanning. *IEEE Access* **2021**, *9*, 67281–67301. [[CrossRef](#)]
23. Nocerino, E.; Stathopoulou, E.K.; Rigon, S.; Remondino, F. Surface Reconstruction Assessment in Photogrammetric Applications. *Sensors* **2020**, *20*, 5863. [[CrossRef](#)]
24. Available online: <http://ccwu.me/vsfm/> (accessed on 1 December 2021).
25. Available online: <https://www.agisoft.com> (accessed on 1 December 2021).
26. Joo, H.; Simon, T.; Li, X.; Liu, H.; Tan, L.; Gui, L.; Banerjee, S.; Godisart, T.; Nabbe, B.; Matthews, I.; et al. Panoptic Studio: A Massively Multiview System for Social Interaction Capture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *41*, 190–204. [[CrossRef](#)]
27. Matusik, W.; Buehler, C.; Raskar, R.; Gortler, S.J.; McMillan, L. Image-based visual hulls. In Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, New Orleans, LA, USA, 23–28 July 2000.
28. Franco, J.-S.; Boyer, E. Efficient Polyhedral Modeling from Silhouettes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 414–427. [[CrossRef](#)]
29. Franco, J.-S.; Lapierre, M.; Boyer, E.; Boyer, J.-S.F.M.L.E. Visual Shapes of Silhouette Sets. In Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT’06), Chapel Hill, NC, USA, 14–16 June 2006; pp. 1–8.
30. Furukawa, Y.; Ponce, J. LNCS 3951—Carved Visual Hulls for Image-Based Modeling. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006.
31. Mulayim, A.Y.; Yilmaz, U.; Atalay, V. Silhouette-based 3-D model reconstruction from multiple images. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2003**, *33*, 582–591. [[CrossRef](#)]
32. Yoon, G.-J.; Cho, H.; Won, Y.-Y.; Yoon, S.M. Three-Dimensional Density Estimation of Flame Captured From Multiple Cameras. *IEEE Access* **2019**, *7*, 8876–8884. [[CrossRef](#)]

33. Tabb, A. Shape from Silhouette Probability Maps: Reconstruction of Thin Objects in the Presence of Silhouette Extraction and Calibration Error. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.
34. Boros, E.; Hammer, P.L. Pseudo-Boolean optimization. *Discret. Appl. Math.* **2002**, *123*, 155–225. [[CrossRef](#)]
35. Loop, C.; Zhang, C.; Zhang, Z. Real-time high-resolution sparse voxelization with application to image-based modeling. In Proceedings of the 5th High-Performance Graphics Conference, Anaheim, CA, USA, 19–21 July 2013.
36. Perez, J.M.; Aledo, P.G.; Sanchez, P.P. Real-time voxel-based visual hull reconstruction. *Microprocess. Microsyst.* **2012**, *36*, 439–447. [[CrossRef](#)]
37. Corazza, S.; Mündermann, L.; Chaudhari, A.M.; Demattio, T.; Cobelli, C.; Andriacchi, T.P. A Markerless Motion Capture System to Study Musculoskeletal Biomechanics: Visual Hull and Simulated Annealing Approach. *Ann. Biomed. Eng.* **2006**, *34*, 1019–1029. [[CrossRef](#)] [[PubMed](#)]
38. Kanaujia, A.; Kittens, N.; Ramanathan, N. Part Segmentation of Visual Hull for 3D Human Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013.
39. Roeck, S.D.; Cornelis, N.; Gool, L.V. Augmenting fast stereo with silhouette constraints for dynamic 3D capture. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006.
40. Lin, H.-Y.; Wu, J.-R. 3D reconstruction by combining shape from silhouette with stereo. In Proceedings of the 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008.
41. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.* **2015**, *34*, 1–16. [[CrossRef](#)]
42. Anguelov, D.; Srinivasan, P.; Koller, D.; Thrun, S.; Rodgers, J.; Davis, J. SCAPE: Shape completion and animation of people. *ACM Trans. Graph.* **2005**, *24*, 408–416. [[CrossRef](#)]
43. Balan, A.O.; Sigal, L.; Black, M.J.; Davis, J.E.; Haussecker, H.W. Detailed Human Shape and Pose from Images. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007.
44. Guan, P.; Weiss, A.; Balan, A.O.; Black, M.J. Estimating human shape and pose from a single image. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009.
45. Dibra, E.; Jain, H.; Oztireli, C.; Ziegler, R.; Gross, M. HS-Nets: Estimating Human Body Shape from Silhouettes with Convolutional Neural Networks. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016.
46. Li, Z.; Heyden, A.; Oskarsson, M. Parametric Model-Based 3D Human Shape and Pose Estimation from Multiple Views. In Proceedings of the Scandinavian Conference on Image Analysis, Norrköping, Sweden, 11–13 June 2019; pp. 336–347.
47. 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 Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 561–578.
48. Tan, V.; Budvytis, I.; Cipolla, R. Indirect deep structured learning for 3D human body shape and pose prediction. In Proceedings of the British Machine Vision Conference 2017, London, UK, 4–7 September 2017.
49. Dibra, E.; Jain, H.; Oztireli, C.; Ziegler, R.; Gross, M. Human Shape from Silhouettes Using Generative HKS Descriptors and Cross-Modal Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
50. Huang, Z.; Li, T.; Chen, W.; Zhao, Y.; Xing, J.; LeGendre, C.; Luo, L.; Ma, C.; Li, H. Deep Volumetric Video From Very Sparse Multi-view Performance Capture. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 351–369.
51. Gilbert, A.; Volino, M.; Collomosse, J.; Hilton, A. Volumetric performance capture from minimal camera viewpoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
52. Natsume, R.; Saito, S.; Huang, Z.; Chen, W.; Ma, C.; Li, H.; Morishima, S. SiCloPe: Silhouette-Based Clothed People. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
53. Xu, W.; Chatterjee, A.; Zollhöfer, M.; Rhodin, H.; Mehta, D.; Seidel, H.-P.; Theobalt, C. MonoPerfCap. *ACM Trans. Graph.* **2018**, *37*, 1–15. [[CrossRef](#)]
54. Li, Z.; Oskarsson, M.; Heyden, A. Detailed 3D Human Body Reconstruction from Multi-view Images Combining Voxel Super-Resolution and Learned Implicit Representation. *Appl. Intell.* **2020**. Available online: <https://link.springer.com/content/pdf/10.1007/s10489-021-02783-8.pdf> (accessed on 1 December 2021).
55. Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Li, H.; Kanazawa, A. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.
56. Kazhdan, M.; Hoppe, H. Screened poisson surface reconstruction. *ACM Trans. Graph.* **2013**, *32*, 1–13. [[CrossRef](#)]
57. Tekumalla, L.; Cohen, E. A Hole-Filling Algorithm for Triangular Meshes. 2004. Available online: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.142.3960&rep=rep1&type=pdf> (accessed on 1 December 2021).
58. Davis, J.; Marschner, S.R.; Garr, M.; Levoy, M. Filling holes in complex surfaces using volumetric diffusion. In Proceedings of the First International Symposium on 3D Data Processing Visualization and Transmission, Padua, Italy, 19–21 June 2002.

59. Chalmovianský, P.; Jüttler, B. Filling Holes in Point Clouds. In *Mathematics of Surfaces*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 196–212.
60. Nowak, M.; Michoński, J.; Sitnik, R. Filling cavities in point clouds representing human body surface using Bezier patches. *Multimed. Tools Appl.* **2021**, *80*, 15093–15134. [[CrossRef](#)]
61. Lu, E.; Cole, F.; Dekel, T.; Zisserman, A.; Freeman, W.T.; Rubinstein, M. Omnimate: Associating Objects and Their Effects in Video. *arXiv* **2021**, arXiv:2105.06993.
62. Available online: https://pytorch.org/hub/pytorch_vision_deeplabv3_resnet101/ (accessed on 1 December 2021).
63. Lin, K.; Wang, L.; Luo, K.; Chen, Y.; Liu, Z.; Sun, M.-T. Cross-Domain Complementary Learning Using Pose for Multi-Person Part Segmentation. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 1066–1078. [[CrossRef](#)]
64. Li, P.; Xu, Y.; Wei, Y.; Yang, Y. Self-Correction for Human Parsing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *1*. [[CrossRef](#)]
65. Xiao, B.; Wu, H.; Wei, Y. Simple Baselines for Human Pose Estimation and Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
66. Jercec, A.; Bojanic, D.; Bartol, K.; Pribanic, T.; Petkovic, T.; Petrak, S. On using PointNet Architecture for Human Body Segmentation. In Proceedings of the 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA), Dubrovnik, Croatia, 23–25 September 2019.
67. Ueshima, T.; Hotta, K.; Tokai, S.; Zhang, C. Training PointNet for human point cloud segmentation with 3D meshes. In Proceedings of the Fifteenth International Conference on Quality Control by Artificial Vision, Tokushima, Japan, 12–14 May 2021.
68. Jonker, P.P. Morphological Operations on 3D and 4D Images: From Shape Primitive Detection to Skeletonization. In Proceedings of the International Conference on Discrete Geometry for Computer Imagery, Uppsala, Sweden, 13–15 December 2000; pp. 371–391.