Monocular Real-Time Volumetric Performance Capture

Ruilong Li[1,2] *, Yuliang Xiu[1,2] *, Shunsuke Saito[1,2], Zeng Huang[1,2], Kyle Olsewski[1,2], and Hao Li[1,2,3]

[1] University of Southern California
[2] USC Institute for Creative Technologies
[3] Pinscreen
{ruilongl, yxiu, zenghuan}@usc.edu, {shunsuke.saito16, olszewski.kyle}@gmail.com, hao@hao-li.com

Abstract. We present the first approach to volumetric performance capture and novel-view rendering at real-time speed from monocular video, eliminating the need for expensive multi-view systems or cumbersome pre-acquisition of a personalized template model. Our system reconstructs a fully textured 3D human from each frame by leveraging Pixel-Aligned Implicit Function (PIFu). While PIFu achieves high-resolution reconstruction in a memory-efficient manner, its computationally expensive inference prevents us from deploying such a system for real-time applications. To this end, we propose a novel hierarchical surface localization algorithm and a direct rendering method without explicitly extracting surface meshes. By culling unnecessary regions for evaluation in a coarse-to-fine manner, we successfully accelerate the reconstruction by two orders of magnitude from the baseline without compromising the quality. Furthermore, we introduce an Online Hard Example Mining (OHEM) technique that effectively suppresses failure modes due to the rare occurrence of challenging examples. We adaptively update the sampling probability of the training data based on the current reconstruction accuracy, which effectively alleviates reconstruction artifacts. Our experiments and evaluations demonstrate the robustness of our system to various challenging angles, illuminations, poses, and clothing styles. We also show that our approach compares favorably with the state-of-the-art monocular performance capture. Our proposed approach removes the need for multi-view studio settings and enables a consumer-accessible solution for volumetric capture.

1 Introduction

Videoconferencing using a single camera is still the most common approach face-to-face communication over long distances, despite recent advances in virtual and augmented reality and 3D displays that allow for far more immersive and compelling interaction. The reason for this is simple: convenience. Though the technology exists to obtain high-fidelity digital representations of one's specific appearance that can be rendered from arbitrary viewpoints, existing methods

* indicates equal contribution

to capture and stream this data [7, 10, 15, 46, 59] require cumbersome capture technology, such as a large number of calibrated cameras or depth sensors, and the expert knowledge to install and deploy these systems. Videoconferencing, on the other hand, simply requires a single video camera, such as those found on common consumer devices, *e.g.* laptops and smartphones. Thus, if we can capture a complete model of a person's unique appearance and motion from a single consumer-grade camera, we can bridge the gap preventing novice users from engaging in immersive communication in virtual environments.

However, successful reconstruction of not only the geometry but also the texture of a person from a single viewpoint poses significant challenges due to depth ambiguity, changing topology, and severe occlusions. To address these challenges, data-driven approaches using high-capacity deep neural networks have been employed, demonstrating significant advances in the fidelity and robustness of human modeling [40, 53, 60, 73]. In particular, Pixel-Aligned Implicit Function (PIFu) [53] achieves fully-textured reconstructions of clothed humans with a very high resolution that is infeasible with voxel-based approaches. On the other hand, the main limitation of PIFu is that the subsequent reconstruction process is not fast enough for real-time applications: given an input image, PIFu densely evaluates 3D occupancy fields, from which the underlining surface geometry is extracted using the Marching Cubes algorithm [35]. After the surface mesh reconstruction, the texture on the surface is inferred in a similar manner. Finally, the colored meshes are rendered from arbitrary viewpoints. The whole process takes tens of seconds per object when using a 256^3 resolution. Our goal is to achieve such fidelity and robustness with the highly efficient reconstruction and rendering speed for real-time applications.

To this end, we introduce a novel surface reconstruction algorithm, as well as a direct rendering method that does not require extracting surface meshes for rendering. The newly introduced surface localization algorithm progressively queries 3D locations in a coarse-to-fine manner to construct 3D occupancy fields with a smaller number of points to be evaluated. We empirically demonstrate that our algorithm retains the accuracy of the original reconstruction, while being two orders of magnitude faster than the brute-force baseline. Additionally, combined with the proposed surface reconstruction algorithm, our implicit texture representation enables direct novel-view synthesis without geometry tessellation or texture mapping, which halves the time required for rendering. As a result, we enable 15 fps processing time with a 256^3 spatial resolution for volumetric performance capture.

In addition, we present a key enhancement to the training method of [53] to further improve the quality and efficiency of reconstruction. To suppress failure cases that rarely occur during training due to the unbalanced data distribution with respect to viewing angles, poses and clothing styles, we introduce an adaptive data sampling algorithm inspired by the Online Hard Example Mining (OHEM) method [55]. We incrementally update the sampling probability based on the current prediction accuracy to train more frequently with hard examples without manually selecting these samples. We find this automatic sampling approach highly effective for reducing artifacts, resulting in state-of-the-art accuracy.

Our main contributions are:

- The first approach to full-body performance capture at real-time speed from monocular video not requiring a template. From a single image, our approach reconstructs a fully textured clothed human under a wide range of poses and clothing types without topology constraints.
- A progressive surface localization algorithm that makes surface reconstruction two orders of magnitude faster than the baseline without compromising the reconstruction accuracy, thus achieving a better trade-off between speed and accuracy than octree-based alternatives.
- A direct rendering technique for novel-view synthesis without explicitly extracting surface meshes, which further accelerates the overall performance.
- An effective training technique that addresses the fundamental imbalance in synthetically generated training data. Our Online Hard Example Mining method significantly reduces reconstruction artifacts and improves the generalization capabilities of our approach.

2 Related Work

Volumetric Performance Capture Volumetric performance capture has been widely used to obtain human performances for free-viewpoint video [24] or high-fidelity geometry reconstruction [62]. To obtain the underlining geometry with an arbitrary topology, performance capture systems typically use general cues such as silhouettes [7, 37, 57, 63], mutli-view correspondences [13, 24], and reflectance information [62]. While these approaches successfully reconstruct geometry with an arbitrary topology, they require a large number of cameras with accurate calibration and controlled illumination. Another approach is to leverage commodity depth sensors to directly acquire 3D geometry. Volumetric fusion approaches have been used to jointly optimize for the relative 3D location and 3D geometry, incrementally updated from the captured sequence using a single depth sensor in real-time [20,42]. Later, this incremental geometry update was extended to non-rigidly deforming objects [19, 41] and joint optimization with reflectance [16]. While these approaches do not require a template or category-specific prior, they only support relatively slow motions. Multi-view systems combined with depth sensors significantly improve the fidelity of the reconstructions [7, 10, 46] and both hardware and software improvements further facilitate the trend of high-fidelity volumetric performance capture [15, 29]. However, the hardware requirements make it challenging to deploy these systems for non-professional users.

Template-based Performance Capture To relax the constraints of traditional volumetric performance capture, one common approach is to use a template model as an additional prior. Early works use a precomputed template model to reduce the number of viewpoints [8, 64] and improve the reconstruction quality [61]. Template models are also used to enable performance capture from RGBD input [68, 71]. However, these systems still rely on well-conditioned input

from multiple viewpoints. Instead of a personalized template model, articulated morphable models such as SCAPE [2] or SMPL [34] are also widely used to recover human pose and shapes from video input [14], a single image [4, 30], or RGBD input [69, 72]. More recently, components corresponding to hands [52] and faces [5, 32] were incorporated into a body model to perform more holistic performance capture from multi-view input [23], which was later extended to monocular input as well [48, 65]. Although the use of a parameteric model greatly eases the ill-posed nature of monocular performance capture, the lack of personalized details such as clothing and hairstyles severely impairs the authenticity of the captured performance. Recently Xu et al. [66] demonstrated that articulated personalized avatars can be tracked from monocular RGB videos by incorporating inferred sparse 2D and 3D keypoints [38]. The most relevant work to our approach is [17], which is the real-time extension of [66] with the reconstruction fidelity also improved with an adaptive non-rigidity update. Unlike the aforementioned template-based approaches, our method is capable of representing personalized details present in the input image without any preprocessing, as our approach is based on a template-less volumetric representation, enabling topological updates and instantaneously changing the subject.

Deep Learning for Human Modeling To infer fine-grained 3D shape and appearance from unconstrained images, where designing hand-crafted features is non-trivial, we need a high-capacity machine learning algorithm. The advent of deep learning showed promise by eliminating the need for hand-crafted features and demonstrated groundbreaking performance for human modeling tasks in the wild [1, 25, 38]. Fully convolutional neural networks have been used to infer 3D skeletal joints from a single image [38, 49, 51], which are used as building blocks for monocular performance capture systems [17, 66]. For full-body reconstruction from a single image, various data representations have been explored, including meshes [25, 28], dense correspondences [1], voxels [21, 60, 73], silhouettes [40], and implicit surfaces [18, 53, 54]. Notably, deep learning approaches using implicit shape representations have demonstrated significantly more detailed reconstructions by eliminating the discretization of space [6, 39, 47]. Saito et al. [53] further improve the fidelity of reconstruction by combining fully convolutional image features with implicit functions, and demonstrate that these implicit field representations can be extended to continuous texture fields for effective 3D texture inpainting without relying on precomputed 2D parameterizations. However, the major drawback of these implicit representations is that the inference is time-consuming due to the dense evaluation of the network in 3D space, which prevents its use for real-time applications. Though we base our 3D representation on [53] for high-fidelity and

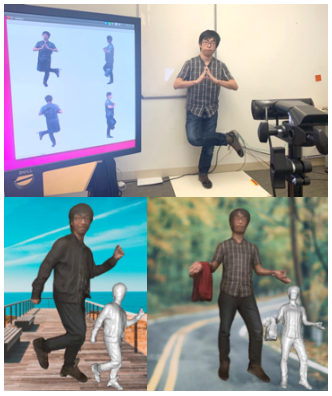

Fig. 1. A performance captured and re-rendered system in real-time from a monocular input video.

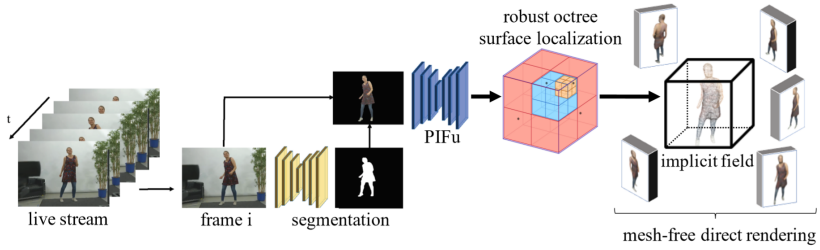

Fig. 2. System overview.

memory-efficient 3D reconstruction, our novel surface inference and rendering algorithms significantly accelerate the reconstruction and visualization of the implicit surface.

3 Method

In this section, we describe the overall pipeline of our algorithm for real-time volumetric capture (Fig. 2). Given a live stream of RGB images, our goal is to obtain the complete 3D geometry of the performing subject in real-time with the full textured surface, including unseen regions. To achieve an accessible solution with minimal requirements, we process each frame independently, as tracking-based solutions are prone to accumulating errors and sensitive to initialization, causing drift and instability [42, 75]. Although recent approaches have demonstrated that the use of anchor frames [3, 10] can alleviate drift, ad-hoc engineering is still required to handle common but extremely challenging scenarios such as changing the subject.

For each frame, we first apply real-time segmentation of the subject from the background. The segmented image is then fed into our enhanced Pixel-Aligned Implicit Function (PIFu) [53] to predict continuous occupancy fields where the underlining surface is defined as a 0.5-level set. Once the surface is determined, texture inference on the surface geometry is also performed using PIFu, allowing for rendering from any viewpoint for various applications. As this deep learning framework with effective 3D shape representation is the core building block of the proposed system, we review it in Sec. 3.1, describe our enhancements to it, and point out the limitations on its surface inference and rendering speed. At the heart of our system, we develop a novel acceleration framework that enables real-time inference and rendering from novel viewpoints using PIFu (Sec. 3.2). Furthermore, we further improve the robustness of the system by sampling hard examples on the fly to efficiently suppress failure modes in a manner inspired by Online Hard Example Mining [55] (Sec. 3.3).

3.1 Pixel-Aligned Implicit Function (PIFu)

In volumetric capture, 3D geometry is represented as the level set surface of continuous scalar fields. That is, given an input frame **I**, we need to determine whether a point in 3D space is inside or outside the human body. While this can

be directly regressed using voxels, where the target space is explicitly discretized [21, 60], the Pixel-Aligned Implicit Function (PIFu) models a function $O(\mathbf{P})$ that queries any 3D point and predicts the binary occupancy field in normalized device coordinates $\mathbf{P} = (P_x, P_y, P_z) \in \mathbb{R}^3$. Notably, with this approach no discretization is needed to infer 3D shapes, allowing reconstruction at arbitrary resolutions.

PIFu first extracts an image feature obtained from a fully convolutional image encoder $g_O(\mathbf{I})$ by a differentiable sampling function $\Phi(\mathbf{P}_{xy}, g_O(\mathbf{I}))$ (following [53], we use a bilinear sampling function [22] for Φ). Given the sampled image feature, a function parameterized by another neural network f_O estimates the occupancy of a queried point $\mathbf{P}$ as follows:

$$O(\mathbf{P}) = f_O(\Phi(\mathbf{P}_{xy}, g_O(\mathbf{I})), P_z) = \begin{cases} 1 & \text{if } \mathbf{P} \text{ is inside surface} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

PIFu [53] uses a fully convolutional architecture for g_O to obtain image features that are spatially aligned with the queried 3D point, and a Multilayer Perceptron (MLP) for the function f_O, which are trained jointly in an end-to-end manner. Aside from the memory efficiency for high-resolution reconstruction, this representation especially benefits volumetric performance capture, as the spatially aligned image features ensure the 3D reconstruction retains details that are present in input images, *e.g.* wrinkles, hairstyles, and various clothing styles. Instead of $L2$ loss as in [53], we use a Binary Cross Entropy (BCE) loss for learning the occupancy fields. As it penalizes false negatives and false positives more harshly than the $L2$ loss, we obtain faster convergence when using BCE.

Additionally, the same framework can be applied to texture inference by predicting vector fields instead of occupancy fields as follows:

$$\mathbf{T}(\mathbf{P}, \mathbf{I}) = f_T(\Phi(\mathbf{P}_{xy}, g_T(\mathbf{I})), \Phi(\mathbf{P}_{xy}, g_O(\mathbf{I})), P_z) = \mathbf{C} \in \mathbb{R}^3, \tag{2}$$

where given a surface point $\mathbf{P}$, the implicit function $\mathbf{T}$ predicts RGB color $\mathbf{C}$. The advantage of this representation is that texture inference can be performed on any surface geometry including occluded regions without requiring a shared 2D parameterization [31, 67]. We use the $L1$ loss from the sampled point colors.

Furthermore, we made several modifications to the original implementation of [53] to further improve the accuracy and efficiency. For shape inference, instead of the stacked hourglass [43], we use HRNetV2-W18-Small-v2 [58] as a backbone, which demonstrates superior accuracy with less computation and parameters. We also use conditional batch normalization [9, 11, 39] to condition the MLPs on the sampled image features instead of the concatenation of these features to the queried depth value, which further improves the accuracy without increasing computational overhead. Additionally, inspired by an ordinal depth regression approach [12], we found that representing depth P_z as a soft one-hot vector more effectively propagates depth information, resulting in faster convergence. For texture inference, we detect the visible surface from the reconstruction and directly use the color from the corresponding pixel, as these regions do not require any inference, further improving the realism of free viewpoint rendering. We provide additional ablation studies to validate our design choices in the supplemental material.

Inference for Human Reconstruction. In [53], the entire digitization pipeline starts with the dense evaluation of the occupancy fields in 3D, from which the surface mesh is extracted using Marching Cubes [35]. Then, to obtain the fully textured mesh, the texture inference module is applied to the vertices on the surface mesh. While the implicit shape representation allows us to reconstruct 3D shapes with an arbitrary resolution, the evaluation in the entire 3D space is prohibitively slow, requiring tens of seconds to process a single frame. Thus, acceleration by at least two orders of magnitude is crucial for real-time performance.

3.2 Real-Time Inference and Rendering

To reduce the computation required for real-time performance capture, we introduce two novel acceleration techniques. First, we present an efficient surface localization algorithm that retains the accuracy of the brute-force reconstruction with the same complexity as naive octree-based reconstruction algorithms. Furthermore, since our final outputs are renderings from novel viewpoints, we bypass the explicit mesh reconstruction stage by directly generating a novel-view rendering from PIFu. By combining these two algorithms, we can successfully render the performance from arbitrary viewpoints in real-time. We describe each algorithm in detail below.

Octree-based Robust Surface Localization. The major bottleneck of the pipeline is the evaluation of implicit functions represented by an MLP at an excessive number of 3D locations. Thus, substantially reducing the number of points to be evaluated would greatly increase the performance. The octree is a common data representation for efficient shape reconstruction [74] which hierarchically reduces the number of nodes in which to store data. To apply an octree for an implicit surface parameterized by a neural network, recently [39] propose an algorithm that subdivides grids only if it is adjacent to the boundary nodes (*i.e.*, the interface between inside node and outside node) after binarizing the predicted occupancy value. We found that this approach often produces inaccurate reconstructions compared to the surface reconstructed by the brute force baseline (see Fig. 3). Since a predicted occupancy value is a continuous value in the range $[0, 1]$, indicating the confidence in and proximity to the surface, another approach

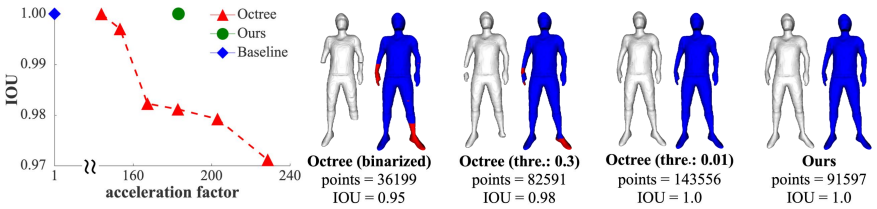

Fig. 3. Comparison of surface reconstruction methods. The plot shows the trade-off between the retention of the accuracy of the original reconstruction (*i.e.*, IOU) and speed. The acceleration factor is computed by dividing the number of evaluation points by that with the brute-force baseline. Note that the thresholds used for the octree reconstructions are 0.05, 0.08, 0.12, 0.2, 0.3, and 0.4 from left to right in the plot.

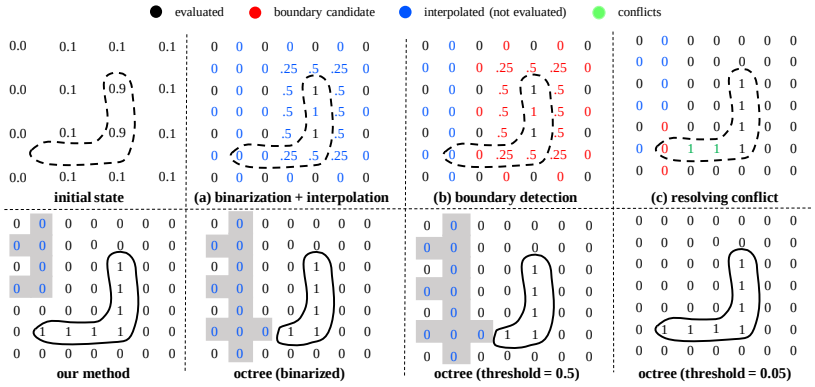

Fig. 4. Our surface localization algorithm overview. The dash and solid line denote the true surface and the reconstructed surface respectively. The nodes that are not used for the time-consuming network evaluation are shaded grey.

is to subdivide grids if the maximum absolute deviation of the neighbor coarse grids is larger than a threshold. While this approach allows for control over the trade-off between reconstruction accuracy and acceleration, we also found that this algorithm either excessively evaluates unnecessary points to perform accurate reconstruction or suffers from impaired reconstruction quality in exchange for higher acceleration. To this end, we introduce a surface localization algorithm that hierarchically and precisely determines the boundary nodes.

We illustrate our surface localization algorithm in Fig. 4. Our goal is to locate grid points where the true surface exists within one of the adjacent nodes at the desired resolution, as only the nodes around the surface matter for surface reconstruction. We thus use a coarse-to-fine strategy in which boundary candidate grids are progressively updated by culling unnecessary evaluation points.

Given the occupancy prediction at the coarser level, we first binarize the occupancy values with threshold of 0.5, and apply interpolation (*i.e.*, bilinear for 2D cases, and trilinear for 3D) to tentatively assign occupancy values to the grid points at the current level (Fig. 4(a)). Then, we extract the boundary candidates by extracting the grid points whose values are neither 0 nor 1. To cover sufficiently large regions, we apply a dilation operation to incorporate the 1-ring neighbor of these boundary candidates (Fig. 4(b)). These selected nodes are evaluated with the network and the occupancy values at these nodes are updated. Note that if we terminate at this point and move on to the next level, the true boundary candidates may be culled similar to the aforementioned acceleration approaches. Thus, as an additional step, we detect conflict nodes by comparing the binarized values of the interpolation and the network prediction for the boundary candidates. The key observation is that there must be a missing surface region when the value of prediction and the interpolation is inconsistent. The nodes adjacent to the conflict nodes are evaluated with the network iteratively until all the conflicts are resolved (Fig. 4(c)).

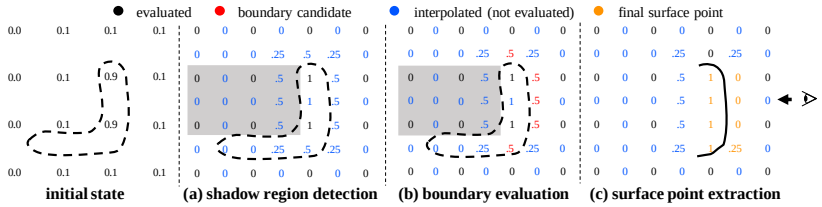

Fig. 5. Our mesh-free rendering overview. The dash and solid line denote the true surface and the reconstructed surface respectively.

Fig. 4 shows the octree-based reconstruction with binarization [39] and the subdivision with a higher threshold suffers from inaccurate surface localization. While the subdivision approach with a lower threshold can prevent inaccurate reconstruction, an excessive number of nodes are evaluated. On the other hand, our approach not only extracts the accurate surface but also effectively reduces the number of nodes to be evaluated (see the number of blue-colored nodes).

Mesh-free Rendering. While the proposed localization algorithm successfully accelerates surface localization, our end goal is rendering from novel viewpoints, and from any viewpoint a large portion of the reconstructed surface is not visible. Furthermore, PIFu allows us to directly infer texture at any point in 3D space, which can substitute the traditional rendering pipeline, where an explicit mesh is required to rasterize the scene. In other words, we can directly generate a novel-view image if the surface location is given from the target viewpoint. Motivated by this observation, we propose a view-based culling algorithm together with a direct rendering method for implicit data representations [39, 45, 53]. Note that while recently differentiable sphere tracing [33] and ray marching [44] approaches have been proposed to directly render implicit fields, these methods are not suitable for real-time rendering, as they sacrifice computational speed for differentiability to perform image-based supervision tasks.

Fig. 5 shows the overview of the view-based surface extraction algorithm. For efficient view-based surface extraction, note that the occupancy grids are aligned with the normalized device coordinates defined by the target view instead of the model or world coordinates. That is, the x and y axes in the grid are corresponding to the pixel coordinates and the z axis is aligned with camera rays. Thus, our first objective is to search along the z axis to identify the first two consecutive nodes within which the surface geometry exists.

First, we apply the aforementioned surface localization algorithm up to the $(L-1)$-th level, where $2^L \times 2^L \times 2^L$ is the target spatial resolution. Then, we upsample the binarized prediction at the $(L-1)$-th level using interpolation and apply the argmax operation along the z axis. The argmax operation provides the maximum value and the corresponding z index along the specified axis, where higher z values are closer to the observer. We denote the maximum value and the corresponding index at a pixel $\mathbf{q}$ by $O_{max}(\mathbf{q})$ and $i_{max}(\mathbf{q})$ respectively. Note that if multiple nodes contain the same maximum value, the function returns the smallest index. If $O_{max}(\mathbf{q}) = 1$, the nodes whose indices are greater than $i_{max}(\mathbf{q})$ are always occluded. Therefore, we treat these nodes as *shadow nodes* which are

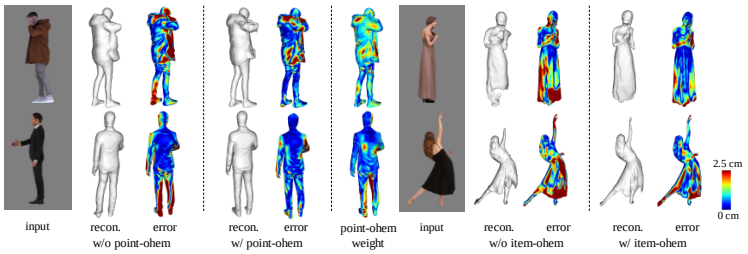

Fig. 6. Qualitative Evaluation of the OHEM sampling. The proposed sampling effectively selects challenging regions, resulting in significantly more robust reconstruction.

discarded for the network evaluation (Fig. 5(a)). Once shadow nodes are marked, we evaluate the remaining nodes with the interpolated value of 0.5 and update the occupancy values (Fig. 5(b)). Finally, we apply binarization to the current occupancy values and perform the argmax operation again along the z axis to obtain the updated nearest-point indices. For the pixels with $O_{max}(\mathbf{q}) = 1$, we take the nodes with the index of $i_{max}(\mathbf{q}) - 1$ and $i_{max}(\mathbf{q})$ as surface points and compute the 3D coordinates of surface $\mathbf{P}(\mathbf{q})$ by interpolating these two nodes by the predicted occupancy value (Fig. 5(c)). Then a novel-view image $\mathbf{R}$ is rendered as follows:

$$\mathbf{R}(\mathbf{q}) = \begin{cases} \mathbf{T}(\mathbf{P}(\mathbf{q}), \mathbf{I}) & \text{if } O_{max}(\mathbf{q}) = 1 \\ \mathbf{B} & \text{otherwise,} \end{cases} \tag{3}$$

where $\mathbf{B} \in \mathbb{R}^3$ is a background color. For virtual teleportation applications, we composite the rendering and the target scene using a transparent background.

3.3 Online Hard Example Mining for Data Sampling

As in [53], the importance-based point sampling for shape learning is more effective than uniform sampling within a bounding box to obtain highly detailed surfaces. However, we observe that this sampling strategy alone still fails to accurately reconstruct challenging poses and viewing angles, which account for only a small portion of the entire training data (see Fig. 6). Although one solution is to synthetically augment the dataset with more challenging training data, manually designing such a data augmentation strategy is non-trivial because various attributes (*e.g.*, poses, view angles, illuminations, and clothing types) may contribute to failure modes, and they are highly entangled.

Nevertheless, the success of importance sampling in [53] illustrates that changing the data sampling distribution directly influences the quality of the reconstruction. This observation leads us to a fundamental solution to address the aforementioned training data bias without domain-specifig knowledge. The key idea is to have the network automatically discover hard examples without manual intervention and adaptively change the sampling probability. We will first formulate the problem and solution in a general form and then develop an algorithm for our specific problem. While there are some works address the data bias problem using online hard negative mining (OHEM) strategy in various

tasks such as learning image descriptors [56], image classifiers [36], and object detection [55], each employs a mining strategy specific to their task. So it is non-trivial to extend there algorithms to another problem. On the contrary, our formulation is general and can be applied to any problem domain as it requires no domain-specific knowledge.

Given a dataset $\mathcal{M}$, a common approach for supervised learning is to define an objective function L_m per data sample m and reduce an error within a mini-batch using optimizers (*e.g.*, SGD, Adam [26]). Assuming uniform distribution for data sampling, we are minimizing the following function $\mathcal{L}$ w.r.t. variables (*i.e.*, network weights) over the course of iterative optimization:

$$\mathcal{L} = \frac{1}{\|\mathcal{M}\|} \sum_{m \in \mathcal{M}} \mathcal{L}_m. \tag{4}$$

Now suppose the dataset is implicitly clustered into S classes denoted as $\{\mathcal{M}_i\}$ based on various attributes (*e.g.*, poses, illumination). Eq. 4 can be written as:

$$\mathcal{L} = \frac{1}{\|\mathcal{M}\|} \sum_i \Big(\sum_{m \in \mathcal{M}_i} \mathcal{L}_m \Big) = \sum_i P_i \cdot \Big(\frac{1}{\|\mathcal{M}_i\|} \sum_{m \in \mathcal{M}_i} \mathcal{L}_m \Big), \tag{5}$$

where $P_i = \frac{\|\mathcal{M}_i\|}{\|\mathcal{M}\|}$ is the sampling probability of the cluster $\mathcal{M}_i$ among all the data samples. As shown in Eq. 5, the objective functions in each cluster are weighted by the probability P_i. This indicates that hard examples with lower probability are outweighed by the majority of the training data, resulting in poor reconstruction. On the other hand, if we modify the sampling probability of data samples in each cluster to be proportional to the inverse of the class probability P_i^{-1}, we can effectively penalize hard examples by removing this bias.

In our problem setting, the goal is to define the sampling probability per target image P_{im} and per 3D point P_{pt}, or alternatively to define the inverse of these directly. Note that the inverse of probability needs to be positive and not to go to infinity. By assuming the accuracy of prediction is correlated with class probability, we approximate the probability of occurrence of each image by an accuracy measurement as $P_{\text{im}} \sim \text{IoU}$, where IoU is computed by the sampled n_O points for each image. Similarly, we use a Binary Cross Entropy loss to approximate the original probability of sampling points. Based on these approximations, we model the inverse of the probabilities as follows:

$$P_{\text{im}}^{-1} = \exp(-\text{IoU}/\alpha_{\text{i}} + \beta_{\text{i}}), \qquad P_{\text{pt}}^{-1} = \frac{1}{\exp(-\mathcal{L}_{BCE}/\alpha_{\text{p}}) + \beta_{\text{p}}}, \tag{6}$$

where α and β are hyperparameters. In our experiments, we use $\alpha_{\text{i}} = 0.15$, $\beta_{\text{i}} = 10.0$, $\alpha_{\text{p}} = 0.7$ and $\beta_{\text{p}} = 0.0$. During training, we compute P_{im}^{-1} and P_{pt}^{-1} for each mini-batch and store the values for each data point, which are later used as the online sampling probability of each image and point after normalization. We refer to OHEM for images and points *item-ohem* and *point-ohem* respectively. Please refer to Sec. 4.1 for the ablation study to validate the effectiveness of our sampling strategy.

Metric	Chamfer		P2S		Std		Chamfer*		P2S*		Runtime
	RP	BUFF	RP	BUFF	RP	BUFF	RP	BUFF	RP	BUFF	fps
VIBE [27]	-	5.485	-	5.794	-	4.279	-	10.653	-	11.572	20
DeepHuman [73]	-	4.208	-	4.340	-	4.022	-	10.460	-	11.389	0.0066
PIFu [53]	1.684	3.629	1.743	3.601	1.953	3.744	6.796	8.417	9.127	8.552	0.033
Ours	1.561	3.615	1.624	3.613	1.624	3.631	6.456	8.675	9.556	8.934	15
Ours+P	**1.397**	**3.515**	**1.514**	**3.566**	**1.552**	**3.518**	6.502	8.366	7.092	8.540	
Ours+P+I	1.431	3.592	1.557	3.603	1.579	3.560	**4.682**	**8.270**	**5.874**	**8.463**	

Table 1. Quantitative results. * mean the results of *top-k* = 10 worst cases. P denotes *point-ohem* and I denotes *item-ohem*.

4 Results

We train our networks using NVIDIA GV100s with 512×512 images. During inference, we use a Logitech C920 webcam on a desktop system equipped with 62 GB RAM, a 6-core Intel i7-5930K processor, and 2 GV100s. One GPU performs geometry and color inference, while the other performs surface reconstruction, which can be done in parallel in an asynchronized manner when processing multiple frames. The overall latency of our system is on average 0.25 second.

We evaluate our proposed algorithms on the RenderPeople [50] and BUFF datasets [70], and on self-captured performances. In particular, as public datasets of 3D clothed humans in motion are highly limited, we use the BUFF datasets [70] for quantitative comparison and evaluation and report the average error measured by the Chamfer distance and point-to-surface (P2S) distance from the prediction to the ground truth. We provide implementation details, including the training dataset and real-time segmentation module, in the supplemental material.

In Fig. 1, we demonstrate our real-time performance capture and rendering from a single RGB camera. Because both the reconstructed geometry and texture inference for unseen regions are plausible, we can obtain novel-view renderings in real-time from a wide range of poses and clothing styles. We provide additional results with various poses, illuminations, viewing angles, and clothing in the supplemental document and video.

4.1 Evaluation

Fig. 3 shows a comparison of surface reconstruction algorithms. The surface localization based on a binarized octree [39] does not guarantee the same reconstruction as the brute-force baseline, potentially losing some body parts. The octree-based reconstruction with a threshold shows the trade-off between performance and accuracy. Our method achieves the best acceleration without any hyperparameters, retaining the original reconstruction accuracy while accelerating surface reconstruction from 30 seconds to 0.14 seconds (7 fps). By combining it with our mesh-free rendering technique, we require only 0.06 seconds per frame (15 fps) for novel-view rendering at the volumetric resolution of 256^3, enabling the first real-time volumetric performance capture from a monocular video.

In Tab. 1 and Fig. 6, we evaluate the effectiveness of the proposed Online Hard Example Mining algorithm quantitatively and qualitatively. Using the same training setting, we train our model with and without the *point-ohem* and

item-ohem sampling. Fig. 6 shows the reconstruction results and error maps from the worst 5 results in the training set. The *point-ohem* successfully improves the fidelity of reconstruction by focusing on the regions with high error (see the *point-ohem* weight in Fig. 6). Similarly, the *item-ohem* automatically supervises more on hard images with less frequent clothing styles or poses, which we expect to capture as accurately as more common poses and clothing styles. As a result, the overall reconstruction quality is significantly improved, compared with the original implementation of [53], achieving state-of-the-art accuracy (Tab. 1).

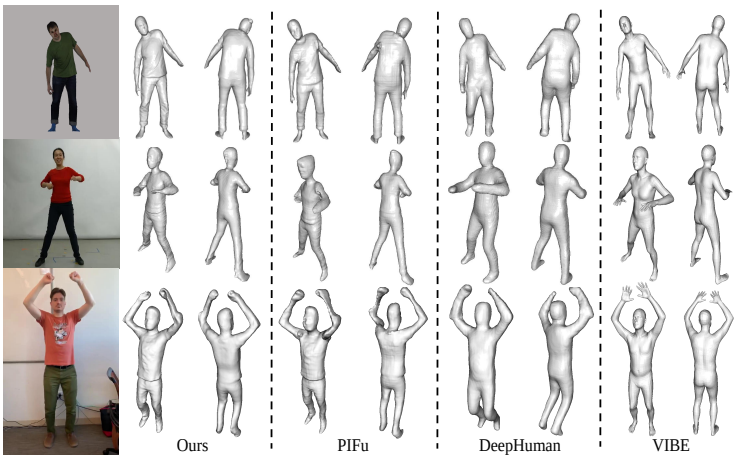

Fig. 7. Qualitative comparison with other reconstruction methods.

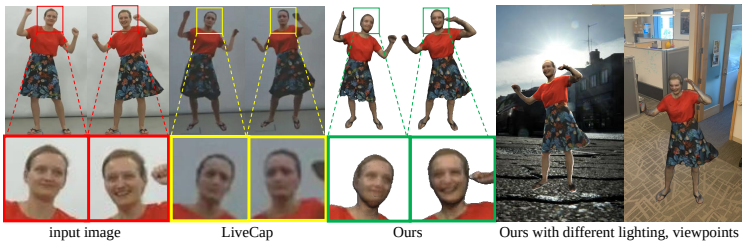

Fig. 8. Comparison template-based performance capture from monocular video.

4.2 Comparison

In Tab. 1 and Fig. 7, we compare our method with the state-of-the-art 3D human reconstruction algorithms from RGB input. Note that we train PIFu [53] using the

same training data with the other settings identical to [53] for a fair comparison, while we use the public pretrained models for VIBE [27] and DeepHuman [73] due to the custom datasets required by each method and their dependency on external modules such as the SMPL [34] model. Although a template-based regression approach [27] achieves robust 3D human estimations from images in the wild, the lack of fidelity and details severely impairs the authenticity of the performances. Similarly, a volumetric performance capture based on voxels [73] suffers from a lack of fidelity due to the limited resolution. While an implicit shape representation [53] achieves high-resolution reconstruction, the reconstructions become less plausible for infrequent poses and the inference speed (30 seconds) is too slow for real-time applications, both of which we address in this paper. We also qualitatively compare our reconstruction with the state-of-the-art real-time performance capture using a pre-captured template [17] (Fig. 8). While the reconstructed geometries are comparable, our method can render performances with dynamic textures that reflect lively expressions, unlike a tracking method using a fixed template. Our approach is also agnostic to topology changes, and can thus handle very challenging scenarios such as changing clothing (Fig. 1).

5 Conclusion

We have demonstrated that volumetric reconstruction and rendering of humans from a single input image is possible to achieve in near real-time speed without sacrificing the final image quality. Our novel progressive surface localization method allows us to vastly reduce the number of points queried during surface reconstruction, giving us a speedup of two orders of magnitude without reducing the final surface quality. Furthermore, we demonstrate that directly rendering novel viewpoints of the captured subject is possible without explicitly extracting a mesh or performing naive, computationally intensive volumetric rendering, allowing us to obtain real-time rendering performance with the reconstructed surface. Finally, our Online Hard Example Mining technique allows us to find and learn the appropriate response to challenging input examples, thereby making it feasible to train our networks with a tractable amount of data while attaining high-quality results with large appearance and motion variations. While we demonstrate our approach on human subjects and performances, our acceleration techniques are straightforward to implement and generalize to any object or topology. We thus believe this will be a critical building block to virtually teleport anything captured by a commodity camera anywhere.

6 Acknowledgement

This research was funded by in part by the ONR YIP grant N00014-17-S-FO14, the CONIX Research Center, a Semiconductor Research Corporation (SRC) program sponsored by DARPA, the Andrew and Erna Viterbi Early Career Chair, the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005, Adobe, and Sony.

References

1. Alp Güler, R., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7297–7306 (2018)
2. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: shape completion and animation of people. ACM Transactions on Graphics **24**(3), 408–416 (2005)
3. Beeler, T., Hahn, F., Bradley, D., Bickel, B., Beardsley, P., Gotsman, C., Sumner, R.W., Gross, M.: High-quality passive facial performance capture using anchor frames. ACM Transactions on Graphics (TOG) **30**(4), 75 (2011)
4. 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 (2016)
5. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics **20**(3), 413–425 (2013)
6. Chen, Z., Zhang, H.: Learning implicit fields for generative shape modeling. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5939–5948 (2019)
7. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. ACM Transactions on Graphics **34**(4), 69 (2015)
8. De Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. ACM Transactions on Graphics **27**(3), 98 (2008)
9. De Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., Courville, A.C.: Modulating early visual processing by language. In: Advances in Neural Information Processing Systems. pp. 6594–6604 (2017)
10. Dou, M., Khamis, S., Degtyarev, Y., Davidson, P., Fanello, S.R., Kowdle, A., Escolano, S.O., Rhemann, C., Kim, D., Taylor, J., et al.: Fusion4d: Real-time performance capture of challenging scenes. ACM Transactions on Graphics **35**(4), 114 (2016)
11. Dumoulin, V., Belghazi, I., Poole, B., Mastropietro, O., Lamb, A., Arjovsky, M., Courville, A.: Adversarially learned inference. arXiv preprint arXiv:1606.00704 (2016)
12. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2002–2011 (2018)
13. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(8), 1362–1376 (2010)
14. Guan, P., Weiss, A., Balan, A.O., Black, M.J.: Estimating human shape and pose from a single image. In: IEEE International Conference on Computer Vision. pp. 1381–1388 (2009)
15. Guo, K., Lincoln, P., Davidson, P., Busch, J., Yu, X., Whalen, M., Harvey, G., Orts-Escolano, S., Pandey, R., Dourgarian, J., et al.: The relightables: Volumetric performance capture of humans with realistic relighting. ACM Trans. Graph. **38**(6) (Nov 2019). https://doi.org/10.1145/3355089.3356571, https://doi.org/10.1145/3355089.3356571
16. Guo, K., Xu, F., Yu, T., Liu, X., Dai, Q., Liu, Y.: Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. ACM Transactions on Graphics (TOG) **36**(3), 32 (2017)

17. Habermann, M., Xu, W., Zollhoefer, M., Pons-Moll, G., Theobalt, C.: Livecap: Real-time human performance capture from monocular video. ACM Transactions on Graphics (TOG) **38**(2), 14 (2019)
18. 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)
19. Innmann, M., Zollhöfer, M., Nießner, M., Theobalt, C., Stamminger, M.: Volumedeform: Real-time volumetric non-rigid reconstruction. In: European Conference on Computer Vision. pp. 362–379. Springer (2016)
20. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., et al.: Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In: Proceedings of the 24th annual ACM symposium on User interface software and technology. pp. 559–568 (2011)
21. Jackson, A.S., Manafas, C., Tzimiropoulos, G.: 3D Human Body Reconstruction from a Single Image via Volumetric Regression. In: ECCV Workshop Proceedings. pp. 0–0. PeopleCap 2018 (2018)
22. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in neural information processing systems. pp. 2017–2025 (2015)
23. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8320–8329 (2018)
24. Kanade, T., Rander, P., Narayanan, P.: Virtualized reality: Constructing virtual worlds from real scenes. IEEE multimedia **4**(1), 34–47 (1997)
25. 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)
26. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
27. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
28. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE International Conference on Computer Vision (2019)
29. Kowdle, A., Rhemann, C., Fanello, S., Tagliasacchi, A., Taylor, J., Davidson, P., Dou, M., Guo, K., Keskin, C., Khamis, S., et al.: The need 4 speed in real-time dense visual tracking. In: SIGGRAPH Asia 2018 Technical Papers. p. 220. ACM (2018)
30. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6050–6059 (2017)
31. Lazova, V., Insafutdinov, E., Pons-Moll, G.: 360-degree textures of people in clothing from a single image. In: International Conference on 3D Vision (3DV) (sep 2019)
32. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Transactions on Graphics (TOG) **36**(6), 194 (2017)
33. Liu, S., Zhang, Y., Peng, S., Shi, B., Pollefeys, M., Cui, Z.: Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. arXiv preprint arXiv:1911.13225 (2019)
34. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transactions on Graphics **34**(6), 248 (2015)

35. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. ACM siggraph computer graphics **21**(4), 163–169 (1987)
36. Loshchilov, I., Hutter, F.: Online batch selection for faster training of neural networks. arXiv preprint arXiv:1511.06343 (2015)
37. Matusik, W., Buehler, C., Raskar, R., Gortler, S.J., McMillan, L.: Image-based visual hulls. In: ACM SIGGRAPH. pp. 369–374 (2000)
38. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. ACM Transactions on Graphics **36**(4), 44:1–44:14 (2017)
39. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. arXiv preprint arXiv:1812.03828 (2018)
40. Natsume, R., Saito, S., Huang, Z., Chen, W., Ma, C., Li, H., Morishima, S.: Siclope: Silhouette-based clothed people. In: CVPR. pp. 4480–4490 (2019)
41. Newcombe, R.A., Fox, D., Seitz, S.M.: DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 343–352 (2015)
42. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: Real-time dense surface mapping and tracking. In: Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on. pp. 127–136 (2011)
43. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. pp. 483–499 (2016)
44. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. arXiv preprint arXiv:1912.07372 (2019)
45. Oechsle, M., Mescheder, L., Niemeyer, M., Strauss, T., Geiger, A.: Texture fields: Learning texture representations in function space. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
46. Orts-Escolano, S., Rhemann, C., Fanello, S., Chang, W., Kowdle, A., Degtyarev, Y., Kim, D., Davidson, P.L., Khamis, S., Dou, M., et al.: Holoportation: Virtual 3d teleportation in real-time. In: Proceedings of the 29th Annual Symposium on User Interface Software and Technology. pp. 741–754 (2016)
47. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. arXiv preprint arXiv:1901.05103 (2019)
48. 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 Conference on Computer Vision and Pattern Recognition. pp. 10975–10985 (2019)
49. Popa, A.I., Zanfir, M., Sminchisescu, C.: Deep multitask architecture for integrated 2d and 3d human sensing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6289–6298 (2017)
50. Renderpeople: (2018), https://renderpeople.com/3d-people
51. Rogez, G., Weinzaepfel, P., Schmid, C.: LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. arXiv preprint arXiv:1803.00455 (2018)
52. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6) (Nov 2017)
53. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: ICCV (2019)

54. 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)
55. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 761–769 (2016)
56. Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Moreno-Noguer, F.: Fracking deep convolutional image descriptors. arXiv preprint arXiv:1412.6537 (2014)
57. Starck, J., Hilton, A.: Surface capture for performance-based animation. IEEE Computer Graphics and Applications **27**(3), 21–31 (2007)
58. Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J.: High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514 (2019)
59. Tang, D., Dou, M., Lincoln, P., Davidson, P., Guo, K., Taylor, J., Fanello, S., Keskin, C., Kowdle, A., Bouaziz, S., et al.: Real-time compression and streaming of 4d performances. In: SIGGRAPH Asia 2018 Technical Papers. p. 256. ACM (2018)
60. Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., Schmid, C.: BodyNet: Volumetric inference of 3D human body shapes. In: European Conference on Computer Vision. pp. 20–36 (2018)
61. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. ACM Transactions on Graphics **27**(3), 97 (2008)
62. Vlasic, D., Peers, P., Baran, I., Debevec, P., Popović, J., Rusinkiewicz, S., Matusik, W.: Dynamic shape capture using multi-view photometric stereo. ACM Transactions on Graphics **28**(5), 174 (2009)
63. Waschbüsch, M., Würmlin, S., Cotting, D., Sadlo, F., Gross, M.: Scalable 3D video of dynamic scenes. The Visual Computer **21**(8), 629–638 (2005)
64. Wu, C., Stoll, C., Valgaerts, L., Theobalt, C.: On-set performance capture of multiple actors with a stereo camera. ACM Transactions on Graphics **32**(6), 161 (2013)
65. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10965–10974 (2019)
66. Xu, W., Chatterjee, A., Zollhöfer, M., Rhodin, H., Mehta, D., Seidel, H.P., Theobalt, C.: Monoperfcap: Human performance capture from monocular video. ACM Transactions on Graphics **37**(2), 27:1–27:15 (2018)
67. Yamaguchi, S., Saito, S., Nagano, K., Zhao, Y., Chen, W., Olszewski, K., Morishima, S., Li, H.: High-fidelity facial reflectance and geometry inference from an unconstrained image. ACM Transactions on Graphics **37**(4), 162 (2018)
68. Ye, G., Liu, Y., Hasler, N., Ji, X., Dai, Q., Theobalt, C.: Performance capture of interacting characters with handheld kinects. European Conference on Computer Vision pp. 828–841 (2012)
69. Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., Liu, Y.: Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7287–7296 (2018)
70. Zhang, C., Pujades, S., Black, M.J., 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. pp. 4191–4200 (2017)
71. Zhang, P., Siu, K., Zhang, J., Liu, C.K., Chai, J.: Leveraging depth cameras and wearable pressure sensors for full-body kinematics and dynamics capture. ACM Transactions on Graphics (TOG) **33**(6), 221 (2014)

72. Zheng, Z., Yu, T., Li, H., Guo, K., Dai, Q., Fang, L., Liu, Y.: Hybridfusion: real-time performance capture using a single depth sensor and sparse imus. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 384–400 (2018)
73. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
74. Zhou, K., Gong, M., Huang, X., Guo, B.: Data-parallel octrees for surface reconstruction. IEEE transactions on visualization and computer graphics **17**(5), 669–681 (2010)
75. Zollhöfer, M., Nießner, M., Izadi, S., Rehmann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A., Loop, C., Theobalt, C., et al.: Real-time non-rigid reconstruction using an rgb-d camera. ACM Transactions on Graphics **33**(4), 156 (2014)