

Model-based Outdoor Performance Capture

Nadia Robertini^{1,2} Dan Casas¹ Helge Rhodin¹ Hans-Peter Seidel¹ Christian Theobalt¹

¹ Max-Planck-Institute for Informatics ² Intel Visual Computing Institute

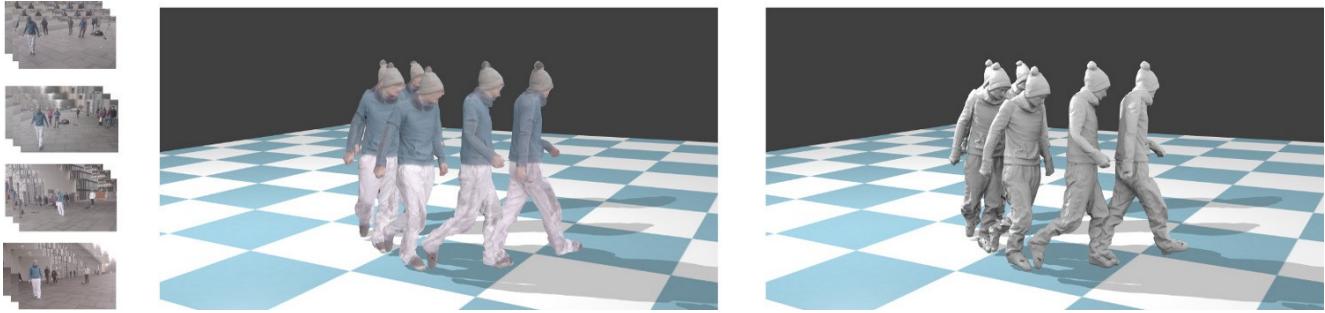


Figure 1: From a set of multi-camera input images (left), we reconstruct the human performance as a temporally consistent 3D mesh that accurately matches the captured motion, here visualized textured (center) and untextured (right) to better appreciate the results. Our novel formulation for model-based human performance capture enables reconstruction of outdoor performances without explicit silhouette segmentation.

Abstract

We propose a new model-based method to accurately reconstruct human performances captured outdoors in a multi-camera setup. Starting from a template of the actor model, we introduce a new unified implicit representation for both, articulated skeleton tracking and nonrigid surface shape refinement. Our method fits the template to unsegmented video frames in two stages – first, the coarse skeletal pose is estimated, and subsequently non-rigid surface shape and body pose are jointly refined. Particularly for surface shape refinement we propose a new combination of 3D Gaussians designed to align the projected model with likely silhouette contours without explicit segmentation or edge detection. We obtain reconstructions of much higher quality in outdoor settings than existing methods, and show that we are on par with state-of-the-art methods on indoor scenes for which they were designed.

1. Introduction

Marker-less human performance capture methods aim to reconstruct the motion as well as the temporally coherent non-rigid surface geometry of people in their general, potentially loosely deforming, apparel, from multi-view RGB

video [7, 15, 21, 42, 49]. Several state-of-the-art methods reconstruct highly detailed 3D mesh sequences by fitting a 3D template to the observed performance [15, 21, 48]; more general methods reconstruct per-frame geometry independently and without a prior model [20, 42, 49]. Most high-quality reconstruction methods fail on footage recorded in general outdoor scenes, as they expect constant lighting and crisp foreground/background subtraction which is best achieved in front of static indoor green screens.

Aiming to overcome this limitation, recent research in joint segmentation and reconstruction [45, 32, 33] successfully reconstructed deforming objects in less constrained setups. However, resulting 3D mesh detail is significantly lower than for previous in-studio silhouette-based methods. Orthogonal to performance capture, marker-less motion capture methods exist which work without silhouettes and in less controlled scenes, but they only estimate coarse human skeleton pose [43, 17, 1, 47].

We propose a new model-based performance capture method that takes a leap forward, and captures detailed human performance, including accurate motion and loose non-rigid surface shape, in less controlled and outdoor environments with moving background, *without* explicit silhouette extraction. To meet the challenges of less controlled environments, we use a new unified implicit formulation for both, articulated skeleton tracking and non-rigid sur-

face shape refinement. Our method fits an initial static surface mesh of an actor to unsegmented video frames in two stages – first skeletal pose is optimized, and subsequently non-rigid surface shape is refined. In both stages, we use a mathematical formulation as the minimization of an objective function that estimates the agreement of model and observation.

The coarse volumetric 3D body shape and body appearance used for skeletal pose estimation, as well as the fine-scale 3D surface geometry and appearance, which is coupled to the coarse body representation, are modeled using the same building block – an implicit function [25] defined over sets of Gaussians. Particularly for surface shape refinement we introduce a new combination 3D Gaussians (which we refer to as *Border Gaussians*), designed to align the projected model with likely silhouette contours without explicit segmentation or edge detection. Also each input image is transformed to a similar implicit representation. This scene representation enables effective and efficient pose optimization, and analytically differentiable smooth objective functions for both coarse fitting and refinement on unsegmented images.

Ours is the first integrated template representation and fitting approach for both coarse and fine geometry capable of accurately reconstructing the articulated motion and deforming surface geometry of actors in challenging outdoor scenes. We obtain reconstructions of much higher quality in outdoor settings than existing methods, and show that we are on par with state-of-the-art methods on indoor scenes for which they were designed.

2. Related Work

Markerless motion capture. Marker-less 3D skeletal(-only) motion capture from multi-view video has been extensively studied in the past [24]. *Generative* methods use actor models composed of primitives, such as ellipses, cylinders and cardboards [51, 9, 27], general mesh representations [4, 21], or parametric human models [39, 31]. They then optimize the pose that best explains the observations. Our approach is also inspired by methods using implicit surfaces [35] and volumetric models [43, 37]. *Discriminative* approaches do pose estimation based on trained classifiers [40, 1, 5]; they often play out their power in combination with generative approaches [39, 41, 17]. While early methods were constrained to controlled indoor settings, more recent methods succeeded on outdoor scenes, even when recorded with a low number of cameras [17, 37].

We use a variant of Stoll *et al.* [43] for coarse pose estimation. Generally motion capture methods are complemented with refinement approaches as they retain no surface detail and are inaccurate in situations of strong surface deformation, such as loose clothing forming folds. Nevertheless, they reconstruct the coarse human pose reliably.

Performance capture. Existing methods for performance capture can be split into two trends. On one hand, *model-based* methods [15, 21, 48, 29, 52, 54] deform a static template of an actor, usually acquired with a full body scanner, to best fit it into synchronized multi-camera input. Parametric body models [26, 2, 34, 30, 31, 6, 36] can also be used as a geometric template, however they are not well suited to represent surface deformations caused by cloth. On the other hand, *model-free* methods [42, 49, 20, 13] remove the need of an initial template by reconstructing a per-frame independent geometry — using visual hull on silhouettes [20] and multi-view photometric stereo reconstruction techniques [42]. Note that, in contrast to model-based methods, the output geometry is temporally incoherent (*e.g.* different amount of vertices and edges per frame), and needs to be later temporally tracked and aligned [10, 11] to achieve a compact temporally coherent 4D representation.

A key limitation of many existing performance capture methods is their dependency on explicit background segmentation for accurate silhouette alignment, an error-prone step which hinders their usage in uncontrolled environments. Progress has been made by multi-view segmentation [50, 16], joint segmentation and reconstruction [44, 22, 8, 32, 33, 14], and also aided by propagation of a manual initialization [23, 52, 45]. In uncontrolled environments the obtained segmentation is still noisy, enabling only skeleton pose [23] and rather coarse 3D reconstructions [32, 33]. Rhodin *et al.* propose a volumetric contour model and directly fit a parametric shape model to image edges, circumventing silhouette extraction entirely. However, only coarse shape without cloth-level detail is reconstructed [36].

Surface refinement in performance capture is particularly relevant to our work. Some methods deform an initial mesh fit by pulling the surface vertices towards the silhouette contours [48, 15]. More sophisticated methods use inverse rendering techniques that refine coarse geometry using shading cues [53, 52]. Most related to our work, recently Robertini *et al.* [38] efficiently recovered medium-frequency surface details optimizing the model-to-image photo-consistency using implicit representations of mesh and images. Similarly, Ilic *et al.* [25] also leverage the attractive properties of implicit surfaces — *e.g.* differentiability — for surface reconstruction. We also use an implicit representation, but tailored for efficient model-to-image edge similarity.

Despite all recent progress in performance capture and surface refinement, existing model-based methods do not cope well in uncontrolled outdoor scenes. Our goal is to enable more accurate model-to-image fitting in challenging scenes.

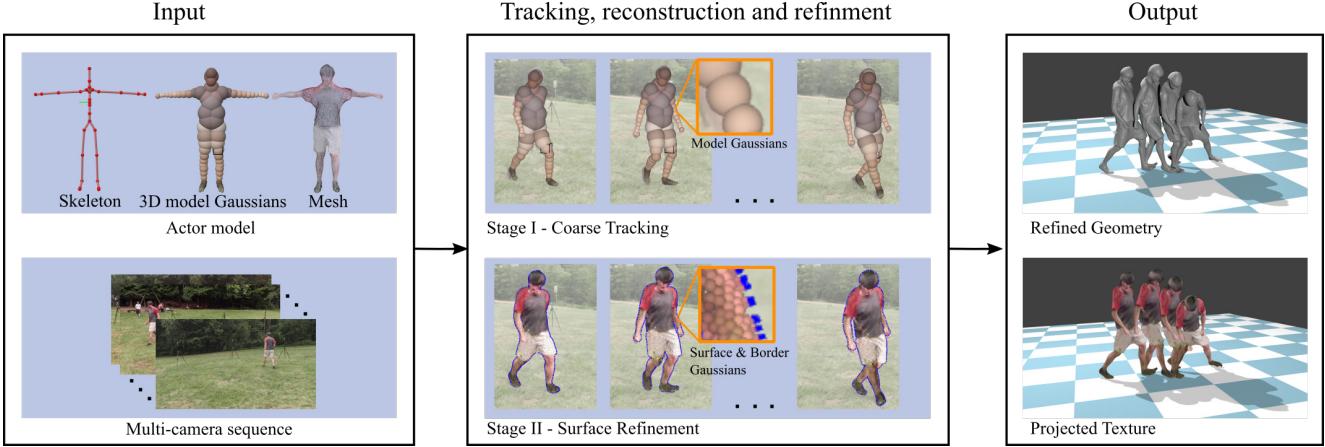


Figure 2: Overview of our method. Input to our optimization approach is an actor model and a multi-view sequence obtained from synchronized and calibrated cameras (right). The actor model consists of a skeleton with a colored implicit Gaussian-based volumetric representation of the actor as well as a colored static 3D mesh. We optimize model-image agreement in two stages (depicted in the middle), where we subsequently estimate the skeleton pose and then refine the surface using a new tracking approach that approximates the input shape with a set of Gaussians, as explained in Section 4.1. Output of our method (left) is a sequence of refined geometry and texture, which best resembles the input performance in terms of pose and surface details.

3. Overview

Our goal is to deform our template model such that it accurately reproduces the performance filmed with calibrated and synchronized multi-view video (Figure 2). To build our actor model we expect as input a colored static 3D mesh of the person obtained through a 3D laser scan, as in [15], through parametric model fitting [3, 31, 36], or through image-based reconstruction from manually created single-time step silhouettes [42, 20]. Then, a kinematic skeleton is fitted to this surface semi-automatically using linear-blend skinning.

We optimize the model-to-image agreement in two stages. Stage-I (Section 4.2) tracks the coarse skeleton motion of the performance, based on the approach from Stoll *et al.* [43], which does not require segmentation and attains high performance through a Gaussian representation of image and actor model (see Section 4.1). Output is an intermediate skeleton motion, which is used to drive the template mesh by skinning. It fails in reproducing non-rigid deformations caused by cloth and soft tissue deformation, and suffers from skinning artifacts. Stage-II estimates high-detail surface shape by maximizing the agreement between a fine-scale implicit representation of the surface mesh and the image. The detailed representation is obtained by placing small 3D *Surface Gaussians* on each vertex of the mesh and by additionally introducing special *Border Gaussians* used for contour alignment without background segmentation (see Section 4.3). While refining the surface, also the skeleton obtained in Stage-I is jointly refined.

The output of our method is a sequence of skeletal poses alongside with refined surface meshes that reproduce the input videos.

4. Performance Capture

4.1. Model Representation

We use a two-layer representation to express our actor model. *Layer-I* consists of a skeleton $\mathcal{S} = \{\theta, \hat{g}\}$, where $\theta = \{\theta_j\}_{j=1}^J$ is a degree of freedom vector that parameterizes the skeleton pose by its joint angles, and $\hat{g} = \{\hat{g}_m\}_{m=1}^M$ a collection of colored 3D Gaussians (*Model Gaussians*), rigidly attached to the skeleton, that approximate the 3D volume of the actor coarsely (see Figure 2 top-left and [43]). The skeleton is once fitted to an actor mesh \mathcal{M} during pre-processing, and skinning weights are computed in a static pose. Each *Model Gaussian* is assigned the average color of the surrounding mesh vertices.

Layer-II of the model is derived from the rigged, colored static surface mesh $\mathcal{M} = \{\mathbf{v}, \mathbf{W}\}$, where $\mathbf{v} = \{v_n\}_{n=1}^N$ are the vertices, and $\mathbf{W} \in \mathbb{R}^{J \times N}$ a matrix of rigging weights that define how each vertex v_n moves with respect to the degree of freedom θ_j . We do not explicitly use the mesh for capturing, but transform its vertices into an implicit representation $\mathcal{G} = \{\tilde{\mathbf{g}}^c, \check{\mathbf{g}}^c\}$ consisting of two subsets of Gaussians. The first set are *Surface Gaussians* $\tilde{\mathbf{g}}^c = \{\tilde{g}_s\}_{s=1}^{S_c}$ for the visible vertices from camera c that do not fall onto an occluding mesh contour at a given frame. A small 3D Gaussian is placed at each vertex position v_n and is assigned the color of the static mesh vertex [38]. The second set are *Border Gaussians* $\check{\mathbf{g}}^c = \{\check{g}_b\}_{b=1}^{B_c}$ for the vertices that fall onto an occluding mesh contour at a given frame. These are placed on the vertices that fall onto the mesh contour and are assigned the color of the static mesh vertex [38].

der Gaussians $\check{\mathbf{g}}^c = \{\check{g}_b\}_{b=1}^{B_c}$ for the vertices that lie on an outer occluding mesh contour in the original camera view c . Note that on *Layer-II*, we preserve the mesh connectivity information to define the surface topology of Surface and Border Gaussians, and assume they are coupled to the skeleton by the original mesh skinning weights. See Figure 2 top-left for a visualization of the actor model.

Additionally, our input image set $\mathcal{F} = \{\mathbf{f}^c\}_{c=1}^C$, where C is the number of cameras, is approximated with a set of 2D Gaussians (*Image Gaussians*) $\mathcal{I} = \{\mathbf{i}^c\}_{c=1}^C$, where \mathbf{i}^c is a vector of 2D Gaussians that approximate the image frame \mathbf{f}^c . The naïve approach is to create a single Image Gaussian for each pixel, and assign to it the pixel color. However, this results in an excessive amount of elements. Instead, as in [43], we use a quad-tree decomposition to efficiently cluster each frame with similar color areas into a single Gaussian.

4.2. Stage-I – Coarse Tracking

In this stage an initial skeletal pose Θ is estimated using the Layer-I model and the algorithm from [43], where we use the commercial implementation available through The Captury Studio [46]. This approach optimizes the skeleton such that the attached Model Gaussian $\hat{\mathbf{g}}$, projected on the image, agrees in color with the nearby Image Gaussians \mathbf{i}_c , see Figure 2 top-center. The core of the method is a pairwise Gaussian overlap measure, which estimates the proximity of Model and Image Gaussians,

$$E_{m,i} = \left[\int_{\Omega} \hat{g}_m(x) \hat{i}_c(x) dx \right]^2 = 2 \frac{\sigma_m \sigma_c}{\sigma_m^2 + \sigma_c^2} e^{-\frac{\|\mu_m - \mu_c\|^2}{\sigma_m^2 + \sigma_c^2}}, \quad (1)$$

where μ_i, σ_i are the Image Gaussian mean and standard deviation, respectively, μ_m is the Model Gaussian mean projected on the image plane by scaled orthographic projection, and σ_m the projected Model Gaussian size.

While any other outdoor motion capture algorithm could be used instead, the analytic and smooth form of the Gaussian overlap integral is of interest for our surface refinement method, which is explained in the next section.

4.3. Stage-II – Surface Refinement

Stage-I only captures the coarse articulated pose, but no fine-scale non-rigid surface deformations are well recovered by just rigging the template \mathcal{M} using the joint angles Θ . We therefore compute a refined estimate using the Layer-II model by finding (\mathbf{v}, Θ) maximizing

$$E(\mathbf{v}, \Theta) = E_{\text{surf}}(\mathbf{v}) + E_{\text{cont}}(\mathbf{v}) - w_{\text{skin}} E_{\text{skin}}(\mathbf{v}, \Theta) - w_{\text{smooth}} E_{\text{smooth}}(\mathbf{v}), \quad (2)$$

initialized with the pose Θ and the associated skinned mesh \mathbf{v} . In the following we explain the individual energy terms.

E_{surf} measures the photo consistency of the *visible* mesh surface vertices with the input images. It is implemented as a generalization of the volumetric Gaussian tracking described in Stage-I to represent the mesh surface implicitly.

$$E_{\text{surf}}(\mathbf{v}) = \sum_s \sum_i^{|g|} C(\delta_{s,i}) E_{s,i}, \quad (3)$$

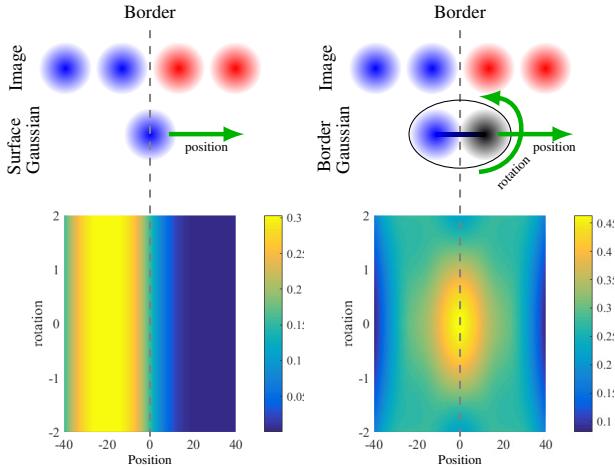
where $|g|$ is the number of Surface Gaussian, and $E_{s,i}$ is the Gaussian overlap between Surface and Image Gaussians, defined in Equation 1. The color similarity $C(\delta_{s,i})$, which maps the HSV color difference $\delta_{s,i}$ to the range $[0, 1]$ with a smooth step function, is a robust measure of Surface and Image Gaussian. This term is similar to [38] and accounts for fine detail refinement of the surface interior, when texture cues are available, but does not account for accurate contour alignment.

E_{cont} measures the model-to-image contour alignment. Our goal is to align each border vertex in color, space and direction with nearby image gradients, *i.e.* move border vertices such that their projection: (1) spatially coincides with a strong edge, (2) shows a strong gradient from vertex color to background color, and (3) the edge orientation aligns with the mesh contour direction. In our setting, we have an accurate shape and appearance model of the actor, but face unknown background, *e.g.* moving scenes, which hinders direct foreground-background gradient computation.

We propose a formulation that neither requires precomputations nor knowledge of the background color and is nevertheless efficient to optimize. For each mesh vertex v_n that is within a Δ -distance to the contour of the mesh in the camera plane, we create an implicit representation, which we referred to as *Border Gaussian* \check{g}_b , by placing two 3D Gaussians: an *Inside Gaussian*, displaced σ_b inside along the surface normal with $C(\delta_{b,i})$ as before; and an *Outside Gaussian*, displaced by σ_s to the outside and with color similarity $(1 - C(\delta_{b,i}))$. We therefore optimize

$$E_{\text{cont}}(\mathbf{v}) = \sum_b \sum_i^{|g|} C(\delta_{b,i}) E_{b_{\text{in}},i} + (1 - C(\delta_{b,i})) E_{b_{\text{out}},i} \quad (4)$$

where $|g|$ is the number of Border Gaussian, and $E_{b_{\text{in}},i}$ and $E_{b_{\text{out}},i}$ represent the Gaussian overlap (Equation 1) between the Inside and Outside 3D Gaussians and the Image Gaussian, respectively. This optimization causes attraction of the Inside Gaussian to the model color and the Outside Gaussian to the background color, because $(1 - C(\delta_{s,i}))$ is large when the color is dissimilar to the foreground, see Figure 3. This pair of Gaussians approximates a gradient from foreground to background color and has maximal response when the desired alignment of color, space and direction is maximal.



(a) Response of Surface Gaussian sampled along different positions and orientations, as in [38]. Notice it has only response to similarly colored Gaussians, and it is rotation insensitive. This representation is good to refine inner parts of the surface, where the color of the Surface Gaussian matches the Image (*e.g.* blue Gaussian over blue Gaussian, in top two rows), but fails for accurate model-to-image contour alignment.
(b) Response of a Border Gaussian (here represented by a blue and black spheres for the Inside and Outside Gaussian, respectively) sampled along different positions and orientations. Note that thanks to the pair structure, it has response to the gradient of foreground to background and are rotation sensitive. Maximum response is at the edge of the Image Gaussians, therefore pulling the Border Gaussian at the image edge as desired.

Figure 3: Energy landscape for a toy example using: (a) Surface Gaussian; and (b) Border Gaussian. From top to bottom, in each subfigure: 2D Gaussian test image, Gaussian sampled along the image, and energy landscape.

E_{skin} couples skeleton and surface motion. Some parts of the body are mostly rigid, such as shoes, and require less surface deformation refinement, while other parts, such as clothing, are more deformable. To model this we link the set of Surface Gaussian to the rigid transformations T_j of each skeleton joint j by minimizing the Surface Gaussian distance to the skinned position $\tilde{\mu}_s$,

$$\tilde{\mu}_s = \sum_{j=1}^J W(j, n) T_j d_{j,n}, \quad (5)$$

where $W(j, n)$ is the skinning weight of vertex n by joint j , and $d_{j,n}$ is the rigid offset between the vertex and the joint. The distance between refined and skinned position is estimated using the 3D overlap of the optimized Surface Gaussian \hat{g}_s with its rigid correspondent \check{g}_s , using Equation 1. A rigidity weight is defined for each vertex by pro-

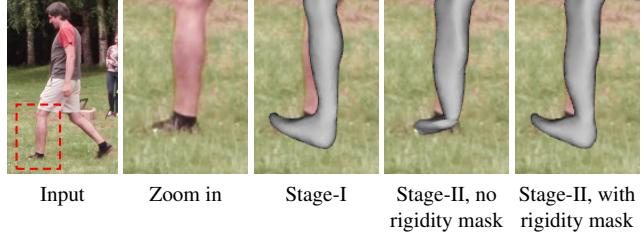


Figure 4: Influence of the rigidity mask on the E_{skin} component. In this challenging situation where the foot is occluded by the high grass, our refinement step in Stage-II, strongly influenced by the data term E_{surf} , tends to implausibly squash the geometry to maximize model-to-observation similarity. A rigidity mask can be used to define the local rigidity of each vertex. On the right, notice how the lower leg is refined such that it matches the image contour, while the foot maintains its original volume.

viding a mask, which enables to regularize different parts of the body differently (*e.g.* allow loose clothing to move *more* freely than hands or feet). This also allows us to deal with difficult situations where body parts are partly occluded, as demonstrated in Figure 4, where the shoe is hidden in high grass and the surface refinement step erroneously squeezes the geometry. When using a rigidity mask, we can enforce vertices on the feet to maintain the original $\tilde{\mu}_s$. Additionally, as a consequence of jointly optimizing \mathbf{v} and Θ , the E_{skin} term also refines the skeleton pose. See the supplementary video for a visualization.

E_{smooth} regularizes unnatural surface deformations with a smoothness prior term. We use a Laplacian smoothness term which penalizes deviations of the optimized mesh Laplacian from the mesh Laplacian of the template mesh, skinned by the initial Θ . Note that [38] constrained vertex motion explicitly to the surface normal direction, which suppresses tangential corrections entirely and can preclude convergence from large displacements, while the Laplacian regularization only ensures coherent motion of nearby vertices.

The unified energy representation enables joint and efficient optimization of surface interior, contour alignment, and skeleton pose, using traditional gradient ascent. In Section 5, we show that the proposed model applies to more general scenes than existing model based approaches and that exploiting of a detailed actor model gives superior results to existing model-free methods.

5. Results and Evaluation

We qualitatively and quantitatively evaluate our method, and compare our results with state-of-the-art human performance capture methods, both indoor and outdoors. See the supplementary video for further evaluation and results. The

supplementary document contains a description of each sequence and actor model.

All results presented in this section were computed on a desktop computer with a NVIDIA GeForce Titan X. Using unoptimized code, our Stage-II run time is about 5–30 minutes on the CPU and 1–4 minutes on the GPU, per frame.

5.1. Quantitative Evaluation

To quantitatively assess the performance of our method, we use a silhouette overlap metric between the actor model projected to the camera plane and the ground truth silhouette obtained with manual segmentation. We label as *positive* and *negative* the foreground and background pixels, respectively. When the label of the projected model and the ground truth image pixel agree, it is a *true* pixel, and *false* otherwise. The combination of true and false pixels is expressed with the F_1 score, commonly used in statistical analysis for binary classification, which can be interpreted as a weighted average of *precision* and *recall*. Note that figures in this section visualize the resulting overlap labels using **Green** for false negative, **Red** for false positive, **Purple** for true positive and **Black** for true negative.

Figure 5 presents a visualization of the qualitative evaluation of the `skirt` sequence — an indoor sequence, with ground truth silhouettes available — showing the overlap of the mesh and the ground truth contour (top row), and the resulting silhouette overlap labels (middle row), in 4 different settings. As expected, the mesh generated in Stage-I (first column) is incapable of capturing the non-rigid skirt contour and suffers from skinning artifacts in the shoulder area. Stage-II (second column) significantly improve these shortcomings, resulting in a much accurate alignment. To further evaluate our approach, we enforce Stage-II to work in ideal conditions where the background is known (third column). Instead of working with the input color images, we use the silhouette images — *i.e.* background is black — and therefore we assign the Outside Gaussians of the Border Gaussian also a black color, instead of the inverse of the inner Gaussian color as we do in uncontrolled conditions. Results under such ideal conditions are shown in the third column, and further validate our new implicit representation: *perfect* color assignment of the inner and Outside Gaussians refines the mesh such that it perfectly matches the ground truth. Our results are in fact comparable to Gall *et al.* [21] (fourth column), a state-of-the-art method that requires explicit silhouette segmentation.

Figure 6 visualizes the F_1 scores across 90 frames and 8 cameras of the `skirt` sequence, for different configurations. We evaluate each of the components of our energy, demonstrating that mesh refinement using Surface Gaussians and Border Gaussians significantly improves over Stage-I, as well as using Surface or Border Gaussian alone. Average F_1 score values when enforcing known

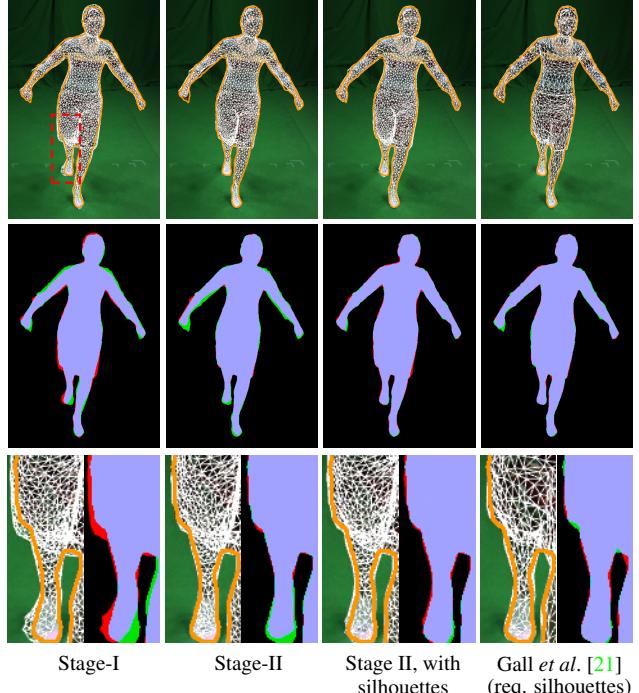


Figure 5: Evaluation of the proposed approach in the `skirt` sequence. In orange, the ground truth contour. Stage-II significantly improves the misalignment errors present in Stage-I, caused by non-rigid deformations and skinning artifacts. Additionally, we also compare the performance of our method in ideal conditions with known silhouettes, which generate results comparable to state-of-the-art silhouette-based methods [21]. See Section 5.1 for color scheme description and further details.

color background (0.9676 ± 0.0056) are comparable to the silhouette-based method from Gall *et al.* [21] (0.9683 ± 0.0045).

Extensive quantitative evaluation on outdoor footage is difficult due to the lack of ground truth data, which can only be generated with laborious manual segmentation. However, we manually segmented 10 frames of the publicly available cathedral dataset [28], as well as of our new sequences `unicampus` and `pablo`. Figure 7 shows the silhouette overlap evaluation in these sequences, demonstrating consistent improvement after mesh refinement. Despite the challenging scenes, with uncontrolled background, our method successfully reconstructs and refines the surface of the actor model, without requiring explicit manual silhouette segmentation. Table 1 presents the F_1 mean and standard deviation of evaluated frames in these sequences, consistently showing that our performance capture method achieves high scores even in such challenging datasets.

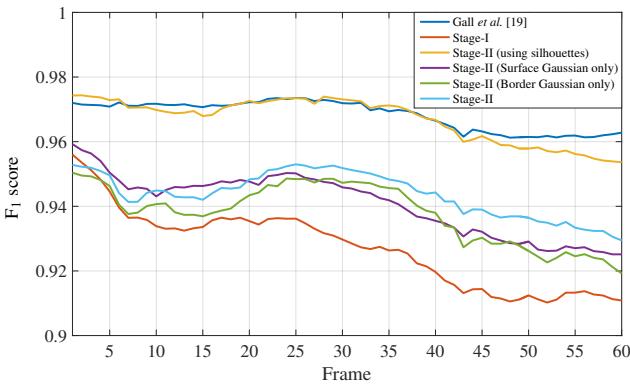


Figure 6: Quantitative evaluation of the silhouette overlap for the *skirt* sequence. Mesh refinement using Surface Gaussians and Border Gaussians (Stage-II, in light blue) significantly improves over Stage-I, as well as over using Surface or Border Gaussian alone. Additionally, we also show that in ideal conditions (*i.e.* known silhouettes), our method performs comparably to the indoor silhouette-based approach of Gall *et al.* [21].



Figure 7: Silhouette overlap evaluation in outdoor sequences. From right to left: original frame; Stage-I mesh and ground truth contour in orange; Stage-II refined mesh and ground truth contour in orange; Stage-I overlap silhouette evaluation; and Stage-II refined overlap silhouette evaluation.

	F ₁ score	
	Stage-I	Stage-II
cathedral	0.9114 ± 0.0077	0.9362 ± 0.0033
pablo	0.8812 ± 0.0156	0.9212 ± 0.0096
unicampus	0.8962 ± 0.0149	0.9223 ± 0.0083
skirt	0.9271 ± 0.0122	0.9676 ± 0.0056

Table 1: Quantitative evaluation of the sequences tested in this paper. The F₁ score of the Stage-II is consistently higher than in Stage-I.

5.2. Qualitative Results

In Figures 1 and 8, as well as in the supplementary video, we qualitatively show reconstruction results on 4 different sequences: *skirt*, *cathedral*, *unicampus* and *pablo*. Our results demonstrate that reconstructed meshes are temporally coherent, do not suffer from temporal noise, and maintain the level of detail of the input template without suffering from unnatural geometric deformation artifacts. We also show textured models by reprojecting the original image frames onto the refined models, which implicitly demonstrates the accuracy of our surface reconstruction. We believe that more advanced view-dependent texturing techniques [12] would alleviate some of the remaining ghosting artifacts in the appearance — however, we believe that ours is one of the first methods that demonstrates refined geometric and textured reconstructions of humans performing outdoors.

6. Discussion and Conclusion

We have presented one of the first model-based methods for human outdoor performance capture. Our new unified implicit representation for both skeleton tracking and non-rigid surface refinement allows to jointly optimize pose and shape, even in scenes with unknown moving background. Our method fits the template to unsegmented video frames in two stages – first skeletal pose is optimized, and subsequently both the pose and the non-rigid surface shape are refined.

While we believe our method takes a leap forward in the area of human performance capture, there are a number of challenges that remain open for future research. Explicit illumination estimation could be incorporated into our model to better handle changes in lighting as well as time-varying shading effects happening on the surface within the same sequence. Our model-based approach requires a colored template mesh for initialization and cannot cope well with complex shape or topology changes. Automating initialization [36] and means to handle topology changes [55] are interesting directions for future work. Enabling the method to handle moving and unsynchronized cameras [18, 19] could

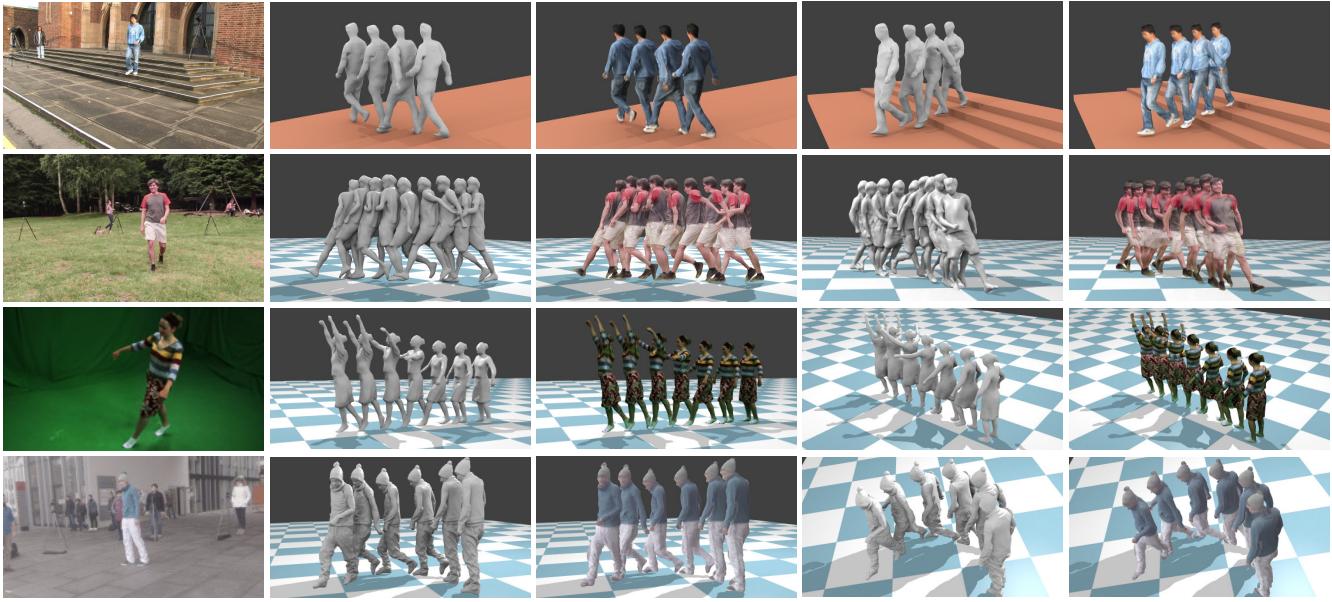


Figure 8: Qualitative results of our human performance capture approach. On the left, a representative input frame of each sequence. For each sequence, we show two pairs of untextured and textured reconstructions of various frames. From top to bottom: cathedral, pablo, skirt and unicampus.

also be further explored, but would require a per-frame estimation of the calibration parameters. High-frequency geometric detail could also be recovered using inverse rendering techniques – this would potentially add even finer detail to the refined meshes.

Acknowledgements

We thank all reviewers for their valuable feedback and The Foundry for license support. This research was funded by the ERC Starting Grant project CapReal (335545).

References

- [1] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-view pictorial structures. In *BMVC*, 2013. [1](#) [2](#)
- [2] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape Completion and Animation of People. *ACM Transactions on Graphics*, 24(3):408–416, 2005. [2](#)
- [3] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *CVPR*, 2007. [3](#)
- [4] L. Ballan and G. M. Cortelazzo. Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes. In *3DPVT*, 2008. [2](#)
- [5] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3D pictorial structures for multiple human pose estimation. In *CVPR*, pages 1669–1676, 2014. [2](#)
- [6] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *ICCV*, pages 2300–2308, 2015. [2](#)
- [7] D. Bradley, T. Popa, A. Sheffer, W. Heidrich, and T. Boubekeur. Markerless garment capture. *ACM Transactions on Graphics (TOG)*, 27(3), 2008. [1](#)
- [8] M. Bray, P. Kohli, and P. H. S. Torr. POSECUT: simultaneous segmentation and 3D pose estimation of humans using dynamic graph-cuts. In *ECCV*, pages 642–655, 2006. [2](#)
- [9] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *CVPR*, pages 8–15, 1998. [2](#)
- [10] C. Budd, P. Huang, M. Klaudiny, and A. Hilton. Global non-rigid alignment of surface sequences. *International Journal of Computer Vision*, 102(1-3):256–270, 2013. [2](#)
- [11] C. Cagniart, E. Boyer, and S. Ilic. Probabilistic deformable surface tracking from multiple videos. In *ECCV*, pages 326–339, 2010. [2](#)
- [12] D. Casas, M. Volino, J. Collomosse, and A. Hilton. 4D Video Textures for Interactive Character Appearance. *Computer Graphics Forum*, 33(2), 2014. [7](#)
- [13] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)*, 34(4), 2015. [2](#)
- [14] Y. Cui, W. Chang, T. Nöll, and D. Stricker. KinectAvatar: Fully automatic body capture using a single kinect. In *ACCV Workshops*, pages 133–147, 2012. [2](#)
- [15] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. *ACM Transactions on Graphics*, 27(3):98, 2008. [1](#) [2](#) [3](#)

- [16] A. Djelouah, J.-S. Franco, E. Boyer, F. Le Clerc, and P. Pérez. Sparse multi-view consistency for object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1890–1903, 2015. 2
- [17] A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *CVPR*, pages 3810–3818, 2015. 1, 2
- [18] A. Elhayek, C. Stoll, N. Hasler, K. I. Kim, H.-P. Seidel, and C. Theobalt. Spatio-temporal motion tracking with unsynchronized cameras. In *CVPR*, pages 1870–1877. IEEE, 2012. 7
- [19] A. Elhayek, C. Stoll, K. I. Kim, and C. Theobalt. Outdoor human motion capture by simultaneous optimization of pose and camera parameters. In *CGF*, volume 34, pages 86–98. Wiley Online Library, 2015. 7
- [20] J.-S. Franco and E. Boyer. Efficient polyhedral modeling from silhouettes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3):414–427, 2009. 1, 2, 3
- [21] J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In *CVPR*, pages 1746–1753, 2009. 1, 2, 6, 7
- [22] J.-Y. Guillemaut and A. Hilton. Joint multi-layer segmentation and reconstruction for free-viewpoint video applications. *International Journal of Computer Vision*, 93(1):73–100, 2011. 2
- [23] N. Hasler, B. Rosenhahn, T. Thormahlen, M. Wand, J. Gall, and H.-P. Seidel. Markerless motion capture with unsynchronized moving cameras. In *CVPR*, pages 224–231, 2009. 2
- [24] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund. Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. *IEEE Journal of Selected Topics in Signal Processing*, 6(5):538–552, 2012. 2
- [25] S. Ilic and P. Fua. Implicit meshes for surface reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):328–333, 2006. 2
- [26] A. Jain, T. Thormählen, H.-P. Seidel, and C. Theobalt. MovieReshape: Tracking and reshaping of humans in videos. *ACM Transactions on Graphics*, 29(5), 2010. 2
- [27] S. X. Ju, M. J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated image motion. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pages 38–44. IEEE, 1996. 2
- [28] H. Kim and A. Hilton. Influence of colour and feature geometry on multi-modal 3d point clouds data registration. In *International Conference on 3D Vision (3DV)*, pages 202–209, 2014. 6
- [29] Y. Liu, C. Stoll, J. Gall, H.-P. Seidel, and C. Theobalt. Markerless motion capture of interacting characters using multi-view image segmentation. In *CVPR*, pages 1249–1256, 2011. 2
- [30] M. Loper, N. Mahmood, and M. J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics*, 33(6):220, 2014. 2
- [31] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6), 2015. 2, 3
- [32] A. Mustafa, H. Kim, J.-Y. Guillemaut, and A. Hilton. General dynamic scene reconstruction from multiple view video. In *ICCV*, 2015. 1, 2
- [33] A. Mustafa, H. Kim, J.-Y. Guillemaut, and A. Hilton. Temporally coherent 4d reconstruction of complex dynamic scenes. In *CVPR*, 2016. 1, 2
- [34] L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, and B. Schiele. Building statistical shape spaces for 3D human modeling. arXiv:1503.05860, 2015. 2
- [35] R. Plankers and P. Fua. Articulated soft objects for multi-view shape and motion capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:63–83, 2003. 2
- [36] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *ECCV*, 2016. 2, 3, 7
- [37] H. Rhodin, N. Robertini, C. Richardt, H.-P. Seidel, and C. Theobalt. A versatile scene model with differentiable visibility applied to generative pose estimation. In *Proceedings of the 2015 International Conference on Computer Vision (ICCV 2015)*, 2015. 2
- [38] N. Robertini, E. De Aguiar, T. Helten, and C. Theobalt. Efficient Multi-view Performance Capture of Fine Scale Surface Detail. In *International Conference on 3D Vision (3DV)*, 2014. 2, 3, 4, 5
- [39] L. Sigal, A. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1337–1344, 2007. 2
- [40] L. Sigal, M. Isard, H. Haussecker, and M. J. Black. Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. *International Journal of Computer Vision*, 98(1):15–48, 2012. 2
- [41] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt. Fast and robust hand tracking using detection-guided optimization. In *CVPR*, 2015. 2
- [42] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3):21–31, 2007. 1, 2, 3
- [43] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of gaussians body model. In *ICCV*, pages 951–958, 2011. 1, 2, 3, 4
- [44] R. Szeliski and P. Golland. Stereo matching with transparency and matting. In *ICCV*, pages 517–524, 1998. 2
- [45] A. Taneja, L. Ballan, and M. Pollefeys. Modeling dynamic scenes recorded with freely moving cameras. In *Asian Conference on Computer Vision*, pages 613–626, 2010. 1, 2
- [46] The Captury. Captury Studio. <http://www.thecaptury.com>. 4
- [47] R. Urtasun, D. J. Fleet, and P. Fua. Temporal motion models for monocular and multiview 3d human body tracking. *Computer vision and image understanding*, 104(2):157–177, 2006. 1

- [48] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics (TOG)*, 27(3), 2008. [1](#), [2](#)
- [49] D. Vlasic, P. Peers, I. Baran, P.Debevec, J. Popović, S. Rusinkiewicz, and W. Matusik. Dynamic shape capture using multi-view photometric stereo. *ACM Transactions on Graphics (TOG)*, 28(5), 2009. [1](#), [2](#)
- [50] T. Wang, J. Collomosse, and A. Hilton. Wide baseline multi-view video matting using a hybrid markov random field. In *ICPR*, pages 136–141, 2014. [2](#)
- [51] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfnder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997. [2](#)
- [52] C. Wu, C. Stoll, L. Valgaerts, and C. Theobalt. On-set performance capture of multiple actors with a stereo camera. *ACM Transactions on Graphics (TOG)*, 32(6), 2013. [2](#)
- [53] C. Wu, K. Varanasi, Y. Liu, H.-P. Seidel, and C. Theobalt. Shading-based dynamic shape refinement from multi-view video under general illumination. In *ICCV*, pages 1108–1115, 2011. [2](#)
- [54] C. Wu, K. Varanasi, and C. Theobalt. Full body performance capture under uncontrolled and varying illumination: A shading-based approach. In *ECCV*, pages 757–770, 2012. [2](#)
- [55] A. Zaharescu, E. Boyer, and R. Horaud. Transformesh: a topology-adaptive mesh-based approach to surface evolution. In *Asian Conference on Computer Vision*, pages 166–175. Springer Berlin Heidelberg, 2007. [7](#)