2015 IEEE International Conference on Computer Vision

# A Versatile Scene Model with Differentiable Visibility Applied to Generative Pose Estimation

Helge Rhodin[1] Nadia Robertini[1, 2] Christian Richardt[1, 2] Hans-Peter Seidel[1] Christian Theobalt[1]

[1] MPI Informatik [2] Intel Visual Computing Institute

## Abstract

*Generative reconstruction methods compute the 3D configuration (such as pose and/or geometry) of a shape by optimizing the overlap of the projected 3D shape model with images. Proper handling of occlusions is a big challenge, since the visibility function that indicates if a surface point is seen from a camera can often not be formulated in closed form, and is in general discrete and non-differentiable at occlusion boundaries. We present a new scene representation that enables an analytically differentiable closed-form formulation of surface visibility. In contrast to previous methods, this yields smooth, analytically differentiable, and efficient to optimize pose similarity energies with rigorous occlusion handling, fewer local minima, and experimentally verified improved convergence of numerical optimization. The underlying idea is a new image formation model that represents opaque objects by a translucent medium with a smooth Gaussian density distribution which turns visibility into a smooth phenomenon. We demonstrate the advantages of our versatile scene model in several generative pose estimation problems, namely marker-less multi-object pose estimation, marker-less human motion capture with few cameras, and image-based 3D geometry estimation.*

## 1. Introduction

Many vision algorithms employ a generative approach to estimate the configuration $\boldsymbol{\theta}$ of a 3D shape that optimizes a function measuring the similarity of the projected 3D model with one or more input camera views of a scene. In rigid object tracking, for example, $\boldsymbol{\theta}$ models the global pose and orientation of an object, whereas in generative marker-less human motion capture, $\boldsymbol{\theta}$ instead models the skeleton pose, and optionally surface geometry and appearance.

The ideal objective function for optimizing similarity has several desirable properties that are often difficult to satisfy: it should have analytic form, analytic derivative, exhibit few local minima, be efficient to evaluate, and numerically well-behaved, *i.e.* smooth. Many approaches already fail to satisfy the first condition and use similarity functions that cannot be expressed or differentiated analytically. This necessitates the use of computationally expensive particle-based optimization methods or numerical gradient approximations that may cause instability and inaccuracy.

A major difficulty in achieving the above properties is the handling of occlusions when projecting from 3D to 2D. Only those parts of a 3D model visible from a camera view should contribute to the similarity. In general, this can be handled by using a visibility function $\mathcal{V}(\boldsymbol{\theta})$ in the similarity measure that describes the visibility of a surface point in pose $\boldsymbol{\theta}$ when viewed from a certain direction. For many shape representations, this function is unfortunately not only hard to formulate explicitly, but it is also binary for solid objects, and hence non-differentiable at points along occlusion boundaries. This renders the similarity function non-differentiable.

In this paper, we introduce a 3D scene representation and image formation model that holistically addresses visibility within a generative similarity energy. It is the first model that satisfies all the following properties:

1. It enables an analytic, continuous and smooth visibility function that is differentiable everywhere in the scene.

2. It enables similarity energies with rigorous visibility handling that are differentiable everywhere in the model and camera parameters.

3. It enables similarity energies that can be optimized efficiently with gradient-based techniques, and which exhibit favorable and more robust convergence in cases where previous visibility approximations fail, such as disocclusions or multiple nearby occlusion boundaries.

Our method approximates opaque objects by a translucent medium with a smooth density distribution defined via a collection of Gaussian functions. This turns occlusion and visibility into smooth phenomena. Based on this representation, we derive a new rigorous image formation model that

1550-5499/15 $31.00 © 2015 IEEE
DOI 10.1109/ICCV.2015.94

is inspired by the principles of light transport in translucent media common in volumetric rendering [4], and which ensures the advantageous properties above.

Although visibility of solid objects is non-differentiable by nature, we demonstrate experimentally in Section 6 that introducing approximations on the scene level is advantageous compared to state-of-the-art methods that employ binary visibility and use spatial image smoothing. We demonstrate the advantages of our approach in several scenarios: marker-less human motion capture with a low number of cameras compared to state-of-the-art methods that lack rigorous visibility modeling [25, 9], body shape and appearance estimation from silhouettes, and more robust multi-object pose optimization compared to methods using local visibility approximations [16].

## 2. Related work

For numerical optimization of generative model-to-image similarity, the objective function needs to consider surface visibility, and needs to be differentiated. The problem is that the (discrete) visibility function $\mathcal{V}$ is generally non-differentiable at occlusion boundaries of solid objects, and often hard to express in analytic form. Some approaches avoid explicit construction of $\mathcal{V}$ by heuristically fixing occlusion relations at the beginning of iterative numerical optimization, which can easily lead to convergence to erroneous local optima, or by re-computation of visibility before each iteration, which can become computationally prohibitive. Commonly the object's silhouette boundaries are handled as special cases, different from the shape interior [17, 24, 26, 29]. These approaches optimize the model configuration (*e.g.* pose and/or shape) such that the projected model boundaries align with multi-view input silhouette boundaries (and possible additional features away from the silhouette), *e.g.* [8, 21, 27, 22, 1].

The integration of binary visibility into the similarity function is more complex. Analytic visibility gradients can be obtained for implicit shapes [11] and mesh surfaces [7], by resorting to distributional derivatives [30, 11, 7], and by geometric considerations on the replacement of a surface with another with respect to motions of the occlusion boundary [13, 6, 16]. For multi-view reconstruction of convex objects, visibility can be inferred efficiently from surface orientation [15]. While these approaches yield similarity functions that are mathematically differentiable almost everywhere, non-differentiability is resolved only locally, which still leads to abrupt changes of object visibility, as illustrated in Figure 1. Efficient gradient-based numerical optimization of the similarity does not fare well under such abrupt and localized changes, leading to frequent convergence to erroneous local optima.

*OpenDR* uses binary visibility, provides an open-source renderer that models illumination and appearance of arbitrary mesh objects, and computes numerically approximated derivatives with respect to the model parameters for perspective projection [16]. Finite differences are used for the spatial derivatives of pixel colors, as proposed for faces by Jones and Poggio [14]. To attain smooth visibility at single occlusion boundaries, spatial smoothing by convolution of the model projection with a smooth kernel [14, 30], and coarse-to-fine pyramid representations [14, 2, 16] are used. Some global dependencies in pose energy between distant scene elements are also handled by coarse-to-fine approaches. Our scene and visibility model handles such global effects by design and is the only model that handles the important case of double occlusion boundaries well, *e.g.* at the point of complete occlusion of an object, see Figure 1.

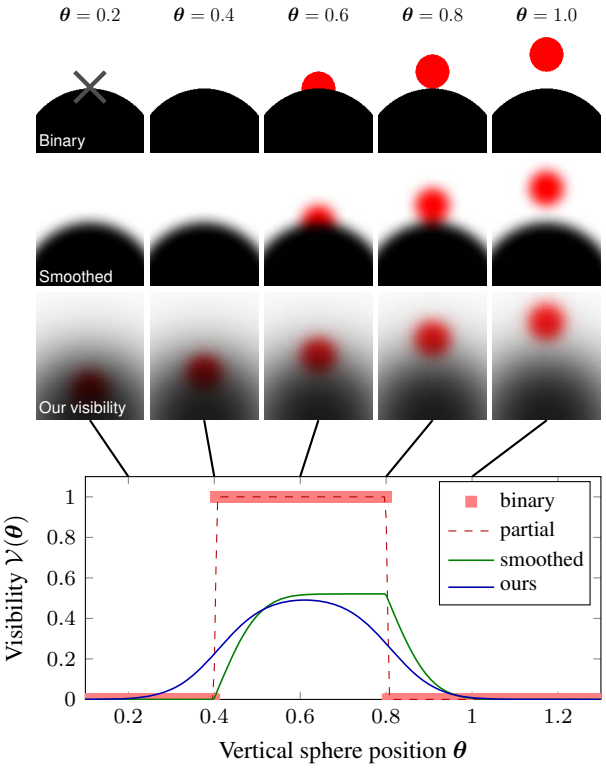


Figure 1. Visibility comparison on a vertically moving sphere. Top to bottom: solid scene with binary visibility, spatial image smoothing, and our visibility model for positions $\boldsymbol{\theta} \in \{0.2, 0.4, 0.6, 0.8, 1\}$. Bottom: plot of the red sphere's visibility at the central pixel (marked by the gray cross in the first image) versus sphere position $\boldsymbol{\theta}$ for the different visibilities. Only our method is smooth at the double occlusion boundaries at $\boldsymbol{\theta} = 0.4$ and $\boldsymbol{\theta} = 0.8$.

Some recent methods abandon the use of surface representation and instead employ an implicit 3D shape model for performing tracking of full-body motion [19, 18, 25, 9], hand motion [3, 23] and object poses [20]. The implicit surfaces can be considered smooth at reprojection boundaries and are therefore well-suited for modeling a differentiable, well-behaved visibility function. Stoll *et al.* [25] do marker-

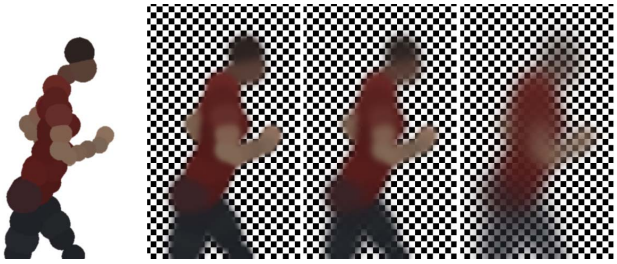

Figure 2. From left to right: A solid sphere actor model, our representation by a translucent medium with Gaussian density visualized on a checkerboard background for increasing smoothness levels ($m = \{0.0001, 0.01, 0.5\}$). Note the proper occlusion, *e.g.* of the left arm and the torso.

less skeletal pose optimization from multi-view video, and use a collection of volumetric 3D Gaussians to represent the human body, as well as 2D Gaussians to model the images. A coarse occlusion heuristic thresholds the overlap between model and image Gaussians. This design allows for long-range effects between model and observation, and avoids expensive occlusion tests, but leads to a problem formulation that is merely piecewise differentiable. Follow-up work empowered tracking with a lower number of cameras by augmenting generative pose tracking with part detections in images [9]. None of these approaches uses a rigorous visibility model, as we do in this paper.

## 3. Scene model

We propose a scene model that approximates solid objects by a smooth density representation, resulting in a visibility function that is well-behaved and differentiable everywhere. In this section, we introduce our scene representation (§3.1), give a physically-based intuition of the resulting visibility function in terms of a translucent medium (§3.2), and present the corresponding image formation model (§3.3). Results of our scene model applied to rigid pose tracking and marker-less motion capture from sparse cameras are shown in Sections 5 and 6.

### 3.1. Smooth scene approximation

Hard object boundaries cause discontinuities of visibility at occlusion boundaries. To obtain a smooth visibility function, we propose a smooth scene representation. We diffuse objects to a smooth translucent medium – with high density at the inside of the original object and a smooth falloff to the outside. In our model, the density defines the extinction coefficient which models how opaque a point in space is, and thus how much it occludes [4]. To obtain an analytic form and for performance reasons, we use a parametric representation for the density $D(\mathbf{x})$ at position $\mathbf{x}$ as the sum of scaled isotropic Gaussians $\mathcal{G} = \{G_q\}_q$, defined as

$$D(\mathbf{x}) = \sum_{G_q \in \mathcal{G}} G_q(\mathbf{x}), \text{ where each Gaussian}$$
$$G_q(\mathbf{x}) = c_q \cdot \exp\left(-\frac{\left\|\mathbf{x} - \boldsymbol{\mu}_q\right\|^2}{2\sigma_q^2}\right) \quad (1)$$

has a magnitude $c_q$, center $\boldsymbol{\mu}_q$ and standard deviation $\sigma_q$. Appearance is modeled by annotating each Gaussian with an albedo attribute $\mathbf{a}_q$. Figure 2 shows an example of the colored density representation for a human actor, consisting of Gaussians of varying size and albedo.

Our model leads to a low-dimensional scene representation parametrized by $\boldsymbol{\gamma} = \{c_q, \boldsymbol{\mu}_q, \sigma_q, \mathbf{a}_q\}_q$. For readability, we use $G_q(\mathbf{x})$ for Gaussians and omit the dependence on $\boldsymbol{\gamma}$.

The degree of opaqueness and smoothness is adjustable by tuning the magnitudes $c_q$ and standard deviations $\sigma_q$ of the Gaussians. We discuss the conversion of a general scene to our Gaussian density representation in Section 4. While other smooth basis functions are conceivable, Gaussians lead to simple analytic expressions for the visibility that work well in practice. While our Gaussian representation is similar to Stoll *et al.*'s [25], its semantics of a translucent medium is fundamentally different and our image formation model with rigorous visibility is phrased in entirely new ways, as explained in the following sections.

### 3.2. Light transport and visibility

Our computation of the visibility $\mathcal{V}$ of a 3D point from a given camera position is inspired by the physical laws of light transport in translucent media, and based on simulation techniques from computer graphics [4]. As the translucent medium is only used as a tool to model continuous visibility, we assume a medium with uniform absorption of all colors without scattering. According to the Beer-Lambert law of light attenuation, the transmittance (the percentage of light transmitted between two points in space) decays exponentially with the optical thickness of a medium, *i.e.* the accumulated density, as visualized in Figure 3. Specifically, the transmittance $T$ of a 3D point at distance $s$ along a ray from a camera position $\mathbf{o}$ in direction $\mathbf{n}$ is

$$T(\mathbf{o}, \mathbf{n}, s, \boldsymbol{\gamma}) = \exp\left(-\int_0^s D(\mathbf{o} + t\mathbf{n})\,\mathrm{d}t\right). \quad (2)$$

Note that for a specific camera, $\mathbf{n}(u, v)$ is uniquely defined for each pixel location $(u, v)$; from now on, we use the short notation $\mathbf{n}$ that is implicitly dependent on the pixel position. With our Gaussian density representation, the density at any point on a line through a sum of 3D Gaussians is in turn the sum of 1D Gaussians. Specifically, inserting the line equation $\mathbf{x} = \mathbf{o} + s\mathbf{n}$ into the 3D Gaussian $G_q$ (1) results in a scaled 1D Gaussian of form $\bar{c}\exp\left(-\frac{(x-\bar{\mu})^2}{2\bar{\sigma}^2}\right)$,

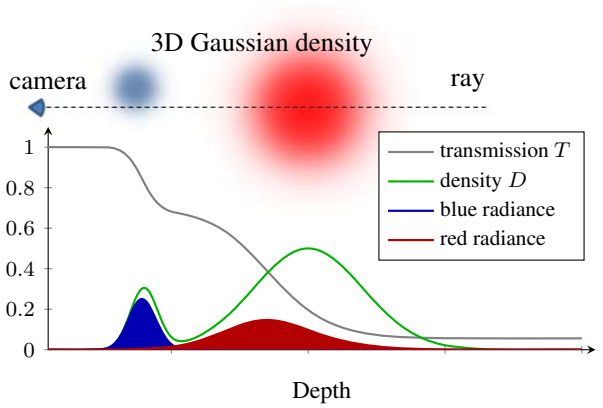


Figure 3. Top: Raytracing of a Gaussian density. Bottom: Light transport along the ray. The density along a ray is a sum of 1D Gaussians (green), and transmittance (gray) falls off from one for increasing optical depth. The radiance is the fraction of reflected light that reaches the camera (red and blue areas). We use it to compute the visibility of a particular Gaussian.

with $\bar{\mu} = (\boldsymbol{\mu} - \mathbf{o})^\top \mathbf{n}$ and $\bar{\sigma} = \sigma$. The updated magnitude is $\bar{c} = c \cdot \exp\left(-\frac{(\boldsymbol{\mu}-\mathbf{o})^\top(\boldsymbol{\mu}-\mathbf{o})-\bar{\mu}^2}{2\bar{\sigma}^2}\right)$. Using the Gaussian form of the density, we can rewrite the transmittance function (2) in analytic form in terms of the error function, $\operatorname{erf}(s) = \frac{2}{\sqrt{\pi}} \int_0^s \exp(-t^2)\, \mathrm{d}t$, as

$$T(\mathbf{o}, \mathbf{n}, s, \boldsymbol{\gamma}) = \exp\left(-\int_0^s \sum_q G_q(\mathbf{o} + t\mathbf{n})\, \mathrm{d}t\right) \tag{3}$$

$$= \exp\left(\sum_q \frac{\bar{\sigma}_q \bar{c}_q}{\sqrt{\frac{2}{\pi}}} \left(\operatorname{erf}\left(\frac{-\bar{\mu}_q}{\sqrt{2}\bar{\sigma}_q}\right) - \operatorname{erf}\left(\frac{s-\bar{\mu}_q}{\sqrt{2}\bar{\sigma}_q}\right)\right)\right). \tag{4}$$

Similar formulations are used for cloud rendering [31, 12]. The transmittance of a medium is symmetric, it also measures the *fractional visibility* of a point $\mathbf{x} = \mathbf{o} + s\mathbf{n}$ from position $\mathbf{o}$, which we denote by $\mathcal{V}(\mathbf{x}, \boldsymbol{\gamma}) := T(\mathbf{o}, \mathbf{n}, s, \boldsymbol{\gamma})$.

### 3.3. Image formation and Gaussian visibility

For image formation, we assume that all scene elements emit an equal amount of light $L_e$. To produce a discrete image from the proposed Gaussian density model, we shoot a ray through each pixel of a virtual pinhole camera. The pixel color is the fraction of source radiance that is emitted along the ray and reaches the camera (color and radiance are related by the camera transfer function; we assume a linear camera response and use pixel color and radiance interchangeably). For the defined medium with pure absorption, the received radiance is the product of transmittance $T$, density $D$, albedo $\mathbf{a}$ and ambient radiance $L_e$, integrated along the ray $\mathbf{x} = \mathbf{o} + s\mathbf{n}$,

$$L(\mathbf{o}, \mathbf{n}) = \int_0^\infty T(\mathbf{o}, \mathbf{n}, s, \boldsymbol{\gamma}) D(\mathbf{x}(s)) \mathbf{a}(\mathbf{x}(s)) L_e\, \mathrm{d}s. \tag{5}$$

This is a special form of the integrated *radiative transfer equation* [4, 5], and it models the fact that each point in space emits light proportional to its density $D(\mathbf{x})$ and illumination $L_e$. For our Gaussian density with parameters $\boldsymbol{\gamma}$ and fixed $L_e = 1$, we obtain

$$L(\mathbf{o}, \mathbf{n}, \boldsymbol{\gamma}) = \int_0^\infty T(\mathbf{o}, \mathbf{n}, s, \boldsymbol{\gamma}) \sum_q G_q(\mathbf{o} + s\mathbf{n}) \mathbf{a}_q\, \mathrm{d}s. \tag{6}$$

To obtain an analytic form, we approximate the infinite integral by sampling a compact interval $S_q = \{\bar{\mu}_q + k\lambda_q \,|\, k \in K \subset \mathbb{Z}\}$ around the mean of each $G_q$:

$$\hat{L}(\mathbf{o}, \mathbf{n}, \boldsymbol{\gamma}) = \sum_q \mathbf{a}_q \sum_{s \in S_q} \lambda_q T(\mathbf{o}, \mathbf{n}, s, \boldsymbol{\gamma}) G_q(\mathbf{o} + s\mathbf{n}), \tag{7}$$

where $\lambda_q \sim \bar{\sigma}_q$ is the sampling step length, which is adaptive to the Gaussian's size.

Gaussians have infinite support ($G_q(\mathbf{x}) > 0$ everywhere), but each Gaussian's contribution vanishes exponentially with the distance from its mean, so local sampling is a good approximation. In practice, we found that five samples with $K = \{-4, -3, \ldots, 0\}$ and $\lambda = \bar{\sigma}$ suffice. Importance sampling could further enhance accuracy.

A final insight is that the inner sum in the radiance equation (7), the sum of the product of source radiance and transmission, measures the contribution of each Gaussian to the pixel color, and therefore computes the *Gaussian visibility*

$$\mathcal{V}_q(\mathbf{o}, \mathbf{n}, \boldsymbol{\gamma}) := \sum_{s \in S_q} \lambda_q T(\mathbf{o}, \mathbf{n}, s, \boldsymbol{\gamma}) G_q(\mathbf{o} + s\mathbf{n}), \tag{8}$$

of $G_q$ from camera $\mathbf{o}$ in direction $\mathbf{n}$. The Gaussian visibility from pixel $(u, v)$, $\mathcal{V}_q((u, v), \boldsymbol{\gamma})$, is equivalent to $\mathcal{V}_q(\mathbf{o}, \mathbf{n}(u, v), \boldsymbol{\gamma})$. The radiance $\hat{L}$ and Gaussian visibility $\mathcal{V}_q$ depend on the set of all Gaussians in the scene. However, our model enables us to represent most scenes with a moderate number of Gaussians (§4), such that the analytic forms of $\hat{L}$ and $\mathcal{V}_q$ can be evaluated efficiently. For increased performance, we also exclude Gaussians with magnitude $\bar{c}_q < 10^{-5}$ for the given ray direction, which does not impair tracking quality (see supplemental document).

## 4. Model creation

In principle, arbitrary shapes can be approximated using a sufficiently large number of small, localized Gaussians. We convert an existing mesh model to our representation by first filling the object's volume with spheres. In our experiments, like the actor in Figure 2, we place spheres manually, but automatic sphere packing could be used instead [28].

We then replace the spheres by Gaussians of 'equal perceived extent'. A translucent object forms its boundary at the point of strongest transmittance change. To find suitable parameters $c_q$ and $\sigma_q$ that approximate a sphere of radius $r$,

we place a Gaussian $G_q$ at its center and analyze the visibility $\mathcal{V}_q(\mathbf{o}, \mathbf{n}, \{c_q, \sigma_q\})$ viewed from an orthographic camera (*i.e.* $\mathbf{n}$ is fixed in view direction and $\mathbf{o}$ is the pixel location). We solve for magnitude $c_q$ and standard deviation $\sigma_q$ such that the transparency at the Gaussian's center, $1 - \mathcal{V}_q$, is equal to a constant $m$, and the inflection point of $\mathcal{V}_q$ lies at distance $r$ from the center. Here, $m$ is a free parameter determining the level of smoothness and translucency, see Figure 2. This is a useful tool to tune robustness versus specificity, as we demonstrate in Section 6.3.2. This procedure aligns the perceived Gaussian size with the reference sphere outline while maintaining a consistent opacity across Gaussians of different size. An example is Figure 1, where the inflection point is aligned with the binary occlusion boundary. The optimization is necessary, as the inflection point of visibility deviates from the density's inflection point, and parameters $c_q$ and $\sigma_q$ jointly influence its location.

For generative tracking, the configuration of the tracked model $\boldsymbol{\theta}$ needs to be mapped to our scene representation using a function $\boldsymbol{\gamma}(\boldsymbol{\theta})$. In rigid object tracking, $\boldsymbol{\gamma}(\boldsymbol{\theta})$ is a single rigid transform that determines the position $\boldsymbol{\mu}_q$ of all Gaussians $G_q$; the sizes $\sigma_q$ and densities $c_q$ are then fixed. For skeletal motion capture, each Gaussian is rigidly attached to a bone in the skeleton, and $\boldsymbol{\theta}$ represents global pose and joint angles. Other mappings for non-rigidly deforming shapes can be easily used, too.

## 5. Pose optimization

Reconstruction methods based on our representation will compute $\boldsymbol{\theta}$ from a set of image observations $\{\mathbf{I}_o\}_o$ captured from different camera positions $\mathbf{o}$, by minimizing an objective function of the general form

$$\mathrm{F}(\boldsymbol{\theta}, \{\mathbf{I}_o\}_o) = \sum_o \mathrm{D}(\boldsymbol{\gamma}(\boldsymbol{\theta}), \mathbf{I}_o) + \mathrm{P}(\boldsymbol{\theta}), \tag{9}$$

where $\mathrm{D}(\boldsymbol{\gamma}, \mathbf{I})$ is a data term, *e.g.* photo-consistency, and $\mathrm{P}(\boldsymbol{\theta})$ is a prior on configurations, *e.g.* a general regularization terms. With our image formation model (§3.3), we can formulate a photo-consistency-based model-to-image overlap in a fully visibility-aware, yet analytic and analytically differentiable way as:

$$\mathrm{D}_{\mathrm{pc}}(\boldsymbol{\gamma}, \mathbf{I}) := \sum_{(u,v)\in I} \left\| \hat{L}(\mathbf{o}, \mathbf{n}(u,v), \boldsymbol{\gamma}) - \mathbf{I}(u,v) \right\|_2^2, \tag{10}$$

where $\mathbf{I}(u,v)$ is the image color at pixel $(u,v)$. In Section 6.1, we use the photo-consistency energy $\mathrm{F}_{\mathrm{pc}}(\boldsymbol{\theta}, \{\mathbf{I}_o\}_o) = \sum_o \mathrm{D}_{\mathrm{pc}}(\boldsymbol{\gamma}(\boldsymbol{\theta}), \mathbf{I}_o)$ without prior for rigid object tracking and body shape and appearance estimation.

We also demonstrate our approach for marker-less human motion capture. The generative method by Stoll *et al.* [25] uses a Gaussian representation for the skeletal body model, and transforms the input image into a collection of Gaussians using color clustering. Their data term sums the color-weighted overlap of all image and projected model Gaussians using a scaled orthographic projection and without rigorous visibility handling, see their paper for details.

For visibility-aware motion capture, we define a new pose energy $\mathrm{F}_{\mathrm{mc}}$ with a perspective camera model and a new visibility-aware data term that accumulates the color dissimilarity $d(\mathbf{I}(u,v), a_q)$ over all pixels $(u,v)$ in image $\mathbf{I}$ and Gaussians $G_q$, weighted by the Gaussian visibility $\mathcal{V}_q$:

$$\mathrm{D}_{\mathrm{mc}}(\boldsymbol{\gamma}, \mathbf{I}) = \sum_{(u,v)} \sum_q d(\mathbf{I}(u,v), a_q) \mathcal{V}_q(\mathbf{o}, \mathbf{n}(u,v), \boldsymbol{\gamma}). \tag{11}$$

To analyze the influence of our new visibility function in isolation, we adopt the remaining model components from the baseline method of Elhayek *et al.* [10]. To compensate for illumination changes, colors are represented in HSV space and the value channel is scaled by $0.2$. To ensure temporal smoothness and anatomical joint limits, accelerations and joint limit violations are quadratically penalized in the prior term $\mathrm{P}(\boldsymbol{\theta})$. Motion capture with the new visibility-aware energy leads to significantly improved results, as we demonstrate in Section 6.3.

For all our experiments on rigid and articulated tracking, we utilize a conditioned conjugate gradient descent solver to minimize the objective function. The analytic derivatives of the objective functions $\mathrm{F}_{\mathrm{pc}}$ and $\mathrm{F}_{\mathrm{mc}}$ with respect to all parameters are listed in the supplemental document.

## 6. Results

We first validate the advantageous properties of our model in general (§6.1), and then show how our scene representation and image formation model lead to improvements over the state of the art in rigid object tracking (§6.2), shape estimation, and marker-less human motion capture (§6.3).

### 6.1. General validation

We validate the smoothness and global support of our visibility handling using a scene with simple occlusions: a red sphere, initially hidden by an occluder, moves up vertically and becomes visible (Figure 1). In our model, the visibility $\mathcal{V}$ of the red sphere (a single Gaussian) is smooth, and hence differentiable with respect to position $\boldsymbol{\gamma}$ (blue line). This is in contrast to surface representations which are only piecewise differentiable: binary visibility functions have discontinuities (red line), and visibility with partial pixel coverage is continuous but non-differentiable at occlusion boundaries (dashed red line). Methods that smooth pixel intensities spatially as a post-process obtain smoothness at single occlusion boundaries. However, when an object occludes or disoccludes behind another occlusion boundary, like the red sphere becoming visible behind the black sphere, and thus two or more occlusion boundaries are in spatial vicinity, visibility is non-differentiable and localized (green line).

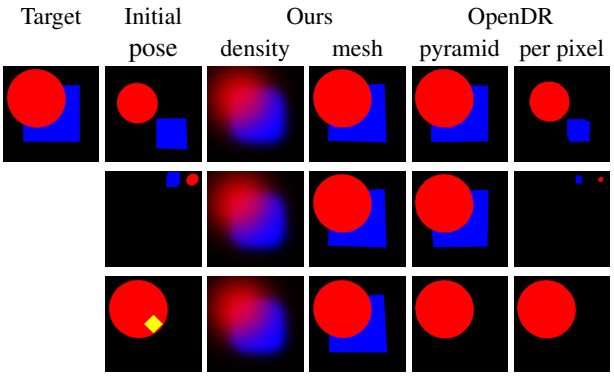


Figure 4. 3D reconstruction of a red sphere (position) and blue cube (position and orientation) using photo-consistency energy $E_{pc}$ from one image for three different initializations. Top to bottom: initialization with overlap to final pose, distant initialization, occluded initialization (the occluded cube is shown in yellow). OpenDR does not find the right solution for initializations without overlap or when far from the solution, and fails under full occlusions. Our method finds the correct pose in all three cases.

Improper handling of this case is a major limitation in practical applications, for instance in motion capture where an arm may disocclude from behind the body. Our approach handles these cases by considering near-visible objects, it 'peeks' behind occlusion boundaries.

### 6.2. Object tracking

We show the advantages of our representation for gradient-based multi-object pose optimization from a single view under photo-consistency and compare against OpenDR [16]. The nine parameters in $\boldsymbol{\theta}$ for the synthetic test scene are the 3D position of a red sphere, and position and orientation of the blue cube. Both objects shall reach the pose shown in the *Target* image of Figure 4. We compare the optimization of $E_{pc}$ with $m = 0.1$ using our model, OpenDR with per-pixel photo-consistency, and OpenDR with a Gaussian pyramid of 6 levels. The optimizer is initialized with 100 random and three manual cases (rows in Figure 4). Without smoothing, OpenDR fails in all cases as object and observation boundary do not overlap sufficiently (last column). With smoothing, OpenDR captures 70% of all random initializations, when one object is fully occluded it fails (fifth column, last row). Our solution captures the correct pose in 88% of all random initializations, even under full occlusion if the occluded object is in the vicinity of the occlusion boundary (forth column, last row). Only in few cases an erroneous local minimum is reached. Averaged over all successful optimizations, the 3D Euclidean positional error of the sphere is $7 \times 10^{-3}$ times its diameter (ours) vs. $4 \times 10^{-5}$ (OpenDR) and for the cube $1.7 \times 10^{-2}$ (ours) vs. $1.3 \times 10^{-2}$ (OpenDR). In essence, the density approximation (one Gaussian for the sphere, 27 Gaussians for the cube) increases robustness and

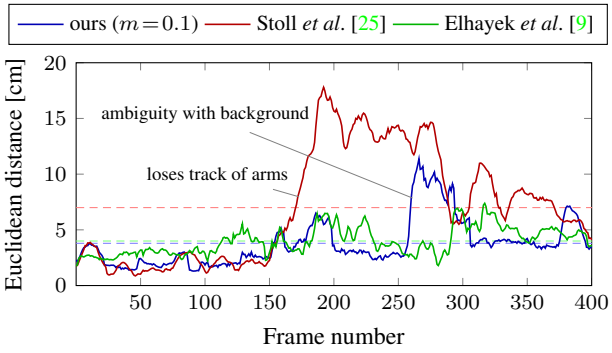


Figure 5. Reconstruction accuracy against marker-based ground truth. Stoll *et al.* looses track of the arms after frame 150, and Elhayek *et al.* lacks accuracy during the first half of the sequence. Our method has a 3.7 cm average joint position error – 45% better than Stoll *et al.* with 7 cm and best overall (dashed lines).

is essential for certain scene configurations. The inaccuracy due to the approximation of sharp edges is of small scale, $\approx 10^{-2}$ compared to OpenDR, model and observation align very well when visualized as meshes (fourth column).

We show in the supplemental document that our approach also enables accurate generative geometry and appearance estimation of non-trivial 3D shapes from images.

### 6.3. Marker-less human motion capture

We now show the benefits of our approach for marker-less human motion capture on three multi-view video sequences with single and multiple actors[1]. Our approach optimizes $E_{mc}$ in the skeletal joint parameters (see Section 5). We compare against the purely generative approach by Stoll *et al.* [25], and the recent combination of their generative method with a ConvNet-based joint detection [9], which was previously the only approach capable of marker-less skeletal motion capture in outdoor scenes with only 2–3 cameras.

We first quantitatively compare against both methods using the average Euclidean reconstruction error against ground-truth 3D joint positions (from a concurrently run marker-based system) using two cameras of the indoor sequence *Marker* [9], see Figure 5. All three algorithms use the same skeleton with 44 pose parameters, 72 Gaussians and data terms using HSV color space [9] to be comparable. The implicit model needed for our approach is created as described in Section 4. Our method improves over the baseline [25] by a 45% lower average error (3.7 cm versus 7 cm), as the baseline cannot track large parts of the sequence at all from only two views. Compared to Elhayek *et al.*, who use a discriminative component together with generative tracking, our method with visibility-aware, purely

[1]A table describing each scene and the relevant parameters, such as number, type and resolution of cameras, how many actors, pose parameters, run time etc., is given in the supplemental document.

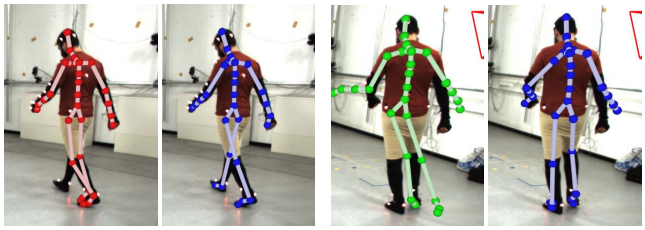


Figure 6. Pose estimates for the *Marker* sequence, using two views for reconstruction. Our method properly handles occlusion of the legs in frame 38 (left), and has much higher accuracy for frame 131 (right), here viewed from a third camera not used for tracking.

generative tracking is more precise for the first 250 frames and comparable for the last frames, where all three methods show errors due to ambiguities with the black background. We thus achieve similarly robust marker-less captured with only two cameras as their more complex method. These quantitative improvements also manifest as clear qualitative pose improvements, as shown in Figure 6. The proper visibility handling overcomes failures of previous techniques when arms and legs occlude. Please see the supplemental material for videos. We also show that the qualitative accuracy of our new marker-less approach captured with only four cameras on the *Walker* sequence from [25] is comparable to their 12-camera result.

We repeat the same comparison on the outdoor sequence *Soccer* with two actors, strong occlusions and fast actions, from only three views. Again, we obtain significantly better accuracy than Stoll *et al.* in terms of 3D joint position, in particular for the limb joints (see Figure 7), as their results quickly show severe failures with so few cameras. To analyze the impact of occlusions, we run our method once for a single actor, and in a second run we jointly track both actors (in total 84 parameters and 182 Gaussians). Simultaneous optimization not only handles self-occlusions but also the mutual occlusion of both actors. This improves by 10.6% and demonstrates the strength of precise and differentiable occlusion handling (see also Figure 8).

In the supplemental document we also analyze the performance of our approach when evaluating our data term for cells of a quad tree that clusters pixels of similar color, as in Stoll *et al.*, instead of for all pixels of the input images .

### 6.3.1 Radius of convergence

Our improved scene model with rigorous visibility handling leads to more *well-behaved* similarity energies with a large *radius of convergence*, *i.e.* they converge for points further away from the global minimum, and a *smooth energy landscape* with few local minima (already observed in OpenDR comparison). We now validate these properties for the motion capture energy $\mathrm{F_{mc}}$, see Figure 9 left. For frame 83

Avg. error [cm]

| | ours | Stoll *et al.* [25] | Elhayek *et al.* [9] |
|---|---|---|---|
| all joints | 7.18 | 10.72 | 4.53 |
| limbs | 4.81 | 9.39 | 4.80 |

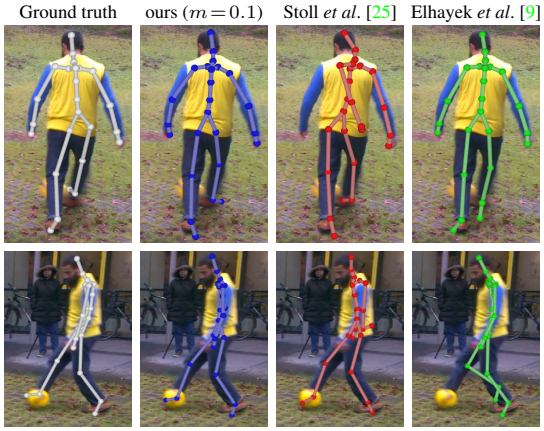


Figure 7. Reconstruction accuracy for the *Soccer* sequence against manually annotated ground truth and comparison to [25] and [9]. The figure illustrates a case of ambiguity across two views, where the generative approach [25] looses track of the person's arm. Our approach and the approach from Elhayek *et al.* [9] instead keep good average tracking with low 3D reprojection error for all joint positions, see the table above.

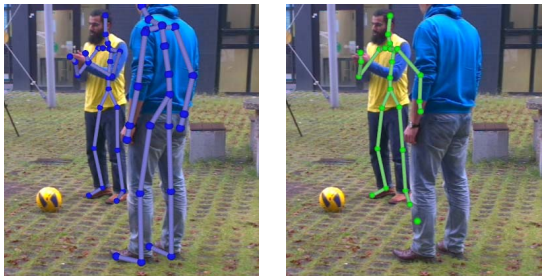


Figure 8. Reconstruction for the *Soccer* sequence with comparison to tracking and modeling all versus a single actor. As shown in the figure, tracking of both subjects at the same time is advantageous as occluding regions are effectively handled by our approach.

of the *Marker* sequence, we initialized the shoulder joint with $\alpha \in [-127°, 43°]$ and analyzed different choices of $m$. As expected, the energy $\mathrm{F_{mc}}$ is smoother and contains fewer local minima for smaller values of $m$. We measure the convergence radius by optimizing from 100 initializations with $\alpha$ equally spaced over the shown interval and count successful convergences. While all configurations succeed for initializations $\alpha \in [-100°, -20°]$ close to the minimum $\alpha = -58°$, only smoother versions ($m \geq 0.1$) converge for distant initializations, see Figure 9 right. The case $m = 0.0001$ models very sharp object boundaries, and hence gives results similar to methods with binary visibility.

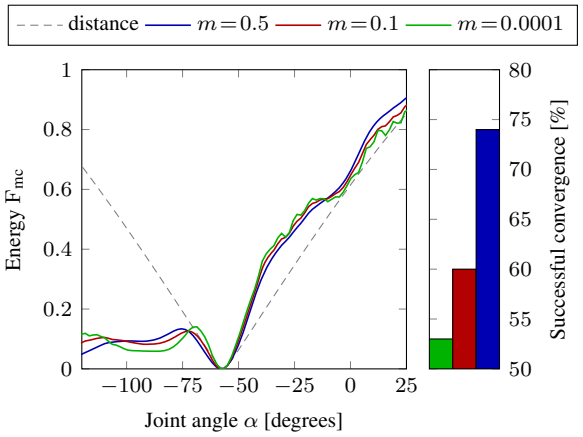


Figure 9. A 1D slice through the energy landscape (for the shoulder joint angle) for different smoothness values $m$. Higher values lead to a smoother energy with fewer local minima, and larger peaks further from the occlusion boundary (at $\alpha = -75$). The global minimum of all configurations aligns well with the Euclidean distance to the ground truth (at $\alpha = -58$). Right: Convergences from 100 initializations within the shown interval.

#### 6.3.2 Visibility gradient and smoothness level

In our final experiment, we show that our new differentiable and well-behaved visibility function is essential for the success of our approach in marker-less human motion capture with very few cameras. For the *Marker* sequence, we fix the visibility for each Gaussian and each camera prior to each iteration of the gradient-based optimizer, *i.e.* changes in occlusion are ignored during optimization. This setup quickly looses track of the limbs and fails completely after 110 frames, see Figure 10. Moreover, to analyze the behavior of our method for different degrees of smoothness in our scene model, we compare multiple fixed smoothness levels. The best trade-off between smoothness and specificity is attained for $m = 0.1$. Which we use for all our experiments unless otherwise specified.

#### 6.3.3 Computational complexity and efficiency

Our implementation of functions $F_{mc}$ and $F_{pc}$ and their gradients has complexity $O(N_I N_q^2 N_K + N_{\boldsymbol{\theta}} N_q)$ for $N_I$ input pixels (summed over all views), and a scene of $N_q$ Gaussians, $N_K$ radiance samples and $N_{\boldsymbol{\theta}}$ parameters. The quadratic complexity in terms of the number of Gaussians originates from the handling of multiple occlusion boundaries (occlusion test for each pair of Gaussians). Our energy is nevertheless efficient to evaluate as the Gaussian density allows to model even complex objects, such as a human, by few primitives. For higher accuracy, coarse-to-fine approaches could be applied. For the *Marker* sequence the performance is 8.1 gradient iterations per second for $F_{mc}$, 10K input pixels (pixels far from the model do not contribute

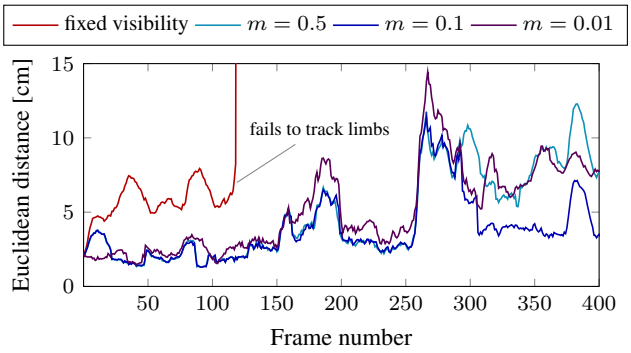


Figure 10. Reconstruction accuracy of joints for different smoothness levels $m$, and for pre-computed fixed visibility per Gaussian.

and are excluded), 72 Gaussians, and 44 pose parameters. The experiments are executed for 200 iterations on a quad-core CPU with 3.6 GHz. As the visibility evaluation of each pixel is independent, further speedups could be obtained by stochastic optimization and parallel execution on GPUs.

### 6.4. Discussion and conclusion

We presented a new scene model and corresponding image formation model that approximates a scene by a translucent medium defined by Gaussian basis functions. This intentionally smoothes out shape and appearance. While this may introduce some uncertainty of shape models, it enables a visibility function and an image formation model that are differentiable everywhere, and efficient to evaluate. Analytic pose optimization energies were already used for motion capture [25, 9], but visibility was only approximated. Our new approach advances the state of the art by enabling analytic, smooth and differentiable pose energies with analytic and differentiable visibility. It also leads to larger convergence radii of these similarity energies. This not only enables us to perform purely generative motion capture at the same accuracy but with far fewer cameras than Stoll *et al.* [25], but also to achieve comparable accuracy with only 2–3 cameras as the more complex method by Elhayek *et al.* [9] which combines generative and discriminative approaches.

OpenDR and other surface models may more accurately represent shape and texture, and also integrate light sources into the scene model. This allows for higher alignment precision for some shapes, but it comes at the cost of a smaller convergence radius, failure under full occlusion, and lower computational efficiency than with our model.

## Acknowledgements

We thank Oliver Klehm, Tobias Ritchel, Carsten Stoll and Nils Hasler for valuable discussions and feedback. This research was funded by the ERC Starting Grant project CapReal (335545).

## References


[1] L. Ballan, A. Taneja, J. Gall, L. V. Gool, and M. Pollefeys. Motion capture of hands in action using discriminative salient points. In *ECCV*, 2012. 2

[2] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, pages 187–194, 1999. 2

[3] L. Bretzner, I. Laptev, and T. Lindeberg. Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering. In *Automatic Face and Gesture Recognition*, pages 423–428, 2002. 2

[4] E. Cerezo, F. Pérez, X. Pueyo, F. J. Seron, and F. X. Sillion. A survey on participating media rendering techniques. *The Visual Computer*, 21(5):303–328, 2005. 2, 3, 4

[5] S. Chandrasekhar. *Radiative Transfer*. Dover Publications Inc, 1960. 4

[6] M. de La Gorce, N. Paragios, and D. J. Fleet. Model-based hand tracking with texture, shading and self-occlusions. In *CVPR*, 2008. 2

[7] A. Delaunoy and E. Prados. Gradient flows for optimizing triangular mesh-based surfaces: Applications to 3D reconstruction problems dealing with visibility. *IJCV*, 95(2):100–123, 2011. 2

[8] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *IJCV*, 61(2):185–205, 2005. 2

[9] A. Elhayek, E. 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*, 2015. 2, 3, 6, 7, 8

[10] A. Elhayek, C. Stoll, K. I. Kim, and C. Theobalt. Outdoor human motion capture by simultaneous optimization of pose and camera parameters. *Computer Graphics Forum*, 2014. 5

[11] P. Gargallo, E. Prados, and P. Sturm. Minimizing the reprojection error in surface reconstruction from images. In *ICCV*, 2007. 2

[12] W. Jakob, C. Regg, and W. Jarosz. Progressive expectation–maximization for hierarchical volumetric photon mapping. *Compututer Graphics Forum*, 30(4):1287–1297, 2011. 4

[13] A. Jalobeanu, F. O. Kuehnel, and J. C. Stutz. Modeling images of natural 3D surfaces: Overview and potential applications. In *CVPR Workshops*, pages 188–188, 2004. 2

[14] M. J. Jones and T. Poggio. Model-based matching by linear combinations of prototypes. A.I. Memo 1583, MIT, 1996. 2

[15] V. Lempitsky, Y. Boykov, and D. Ivanov. Oriented visibility for multiview reconstruction. In *ECCV*, 2006. 2

[16] M. Loper and M. J. Black. OpenDR: An approximate differentiable renderer. In *ECCV*, 2014. 2, 6

[17] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan. Image-based visual hulls. In *SIGGRAPH*, pages 369–374, 2000. 2

[18] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *CVIU*, 104(2):90–126, 2006. 2

[19] R. Plankers and P. Fua. Articulated soft objects for multiview shape and motion capture. *PAMI*, 25(9):1182–1187, 2003. 2

[20] C. Y. Ren, V. Prisacariu, O. Kaehler, I. Reid, and D. Murray. 3D tracking of multiple objects with identical appearance using RGB-D input. In *3DV*, pages 47–54, 2014. 2

[21] B. Rosenhahn, C. Perwass, and G. Sommer. Pose estimation of 3D free-form contours. *IJCV*, 62(3):267–289, 2005. 2

[22] L. Sigal, A. O. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1–2):4–27, 2010. 2

[23] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt. Fast and robust hand tracking using detection-guided optimization. In *CVPR*, 2015. 2

[24] J. Starck and A. Hilton. Surface capture for performance based animation. *IEEE Computer Graphics and Applications*, 27(3):21–31, 2007. 2

[25] 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. 2, 3, 5, 6, 7, 8

[26] T. Tung, S. Nobuhara, and T. Matsuyama. Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *ICCV*, pages 1709–1716, 2009. 2

[27] R. Urtasun, D. J. Fleet, and P. Fua. 3D people tracking with Gaussian process dynamical models. In *CVPR*, pages 238–245, 2006. 2

[28] R. Wang, K. Zhou, J. Snyder, X. Liu, H. Bao, Q. Peng, and B. Guo. Variational sphere set approximation for solid objects. *The Visual Computer*, 22(9–11):612–621, 2006. 4

[29] 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*, 2011. 2

[30] A. Yezzi and S. Soatto. Stereoscopic segmentation. *IJCV*, 53(1):31–43, 2003. 2

[31] K. Zhou, Q. Hou, M. Gong, J. Snyder, B. Guo, and H.-Y. Shum. Fogshop: Real-time design and rendering of inhomogeneous, single-scattering media. In *Pacific Graphics*, pages 116–125, 2007. 4