KillingFusion: Non-rigid 3D Reconstruction without Correspondences

Miroslava Slavcheva [1,2] Maximilian Baust [1] Daniel Cremers [1] Slobodan Ilic [1,2]
[1] Technische Universität München [2] Siemens Corporate Technology

Abstract

We introduce a geometry-driven approach for real-time 3D reconstruction of deforming surfaces from a single RGB-D stream without any templates or shape priors. To this end, we tackle the problem of non-rigid registration by level set evolution without explicit correspondence search. Given a pair of signed distance fields (SDFs) representing the shapes of interest, we estimate a dense deformation field that aligns them. It is defined as a displacement vector field of the same resolution as the SDFs and is determined iteratively via variational minimization. To ensure it generates plausible shapes, we propose a novel regularizer that imposes local rigidity by requiring the deformation to be a smooth and approximately Killing vector field, i.e. generating nearly isometric motions. Moreover, we enforce that the level set property of unity gradient magnitude is preserved over iterations. As a result, KillingFusion reliably reconstructs objects that are undergoing topological changes and fast inter-frame motion. In addition to incrementally building a model from scratch, our system can also deform complete surfaces. We demonstrate these capabilities on several public datasets and introduce our own sequences that permit both qualitative and quantitative comparison to related approaches.

1. Introduction

The growing markets of virtual and augmented reality, combined with the wide availability of inexpensive RGB-D sensors, are perpetually increasing the demand for various applications capable of capturing the user environment in real time. While many excellent solutions for the reconstruction of static scenes exist [5, 12, 23, 31, 33, 34, 43, 54], the more common real-life scenario - where objects move and interact non-rigidly - is still posing a challenge.

The difficulty stems from the high number of unknown parameters and the inherent ambiguity of the problem, since various deformations can yield the same shape. These issues can be alleviated through additional constraints, thus solutions for multi-view surface tracking [4, 8, 9, 10, 18, 22, 50] and template-based approaches [1, 28, 57] have been developed. DynamicFusion [32] is the pioneering work that addresses the general case of incrementally building a 3D model from a single Kinect stream in real time, which is also the objective of our work. VolumeDeform [20] tackles the same problem, combining depth-based correspondences with SIFT features to increase robustness to drift. While both systems demonstrate results of impressive visual quality, they may suffer under larger inter-frame motion due to the underlying mesh-based correspondence estimation.

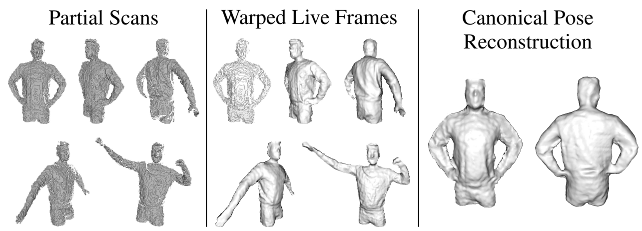


Figure 1. Non-rigid reconstruction from a single noisy Kinect depth stream: KillingFusion builds a complete model under large deformations, rapid inter-frame motion and topology changes.

Many recent works on deformable 3D reconstruction use a signed distance field (SDF) to accumulate the recovered geometry [10, 20, 32], benefiting from its ability to smooth out errors in the cumulative model [7]. However, they intermittently revert back to a mesh representation in order to determine correspondences for non-rigid alignment [20, 32], thereby losing accuracy, computational speed and the capability to conveniently capture topological changes. On the other hand, an SDF inherently tackles situations when surfaces are merging or splitting, *e.g.* a man puts hands on his hips or takes his hat off (Fig. 1, 2), a dog bites its tail, *etc*.

In this paper we propose a non-rigid reconstruction pipeline where the deformation field, the data explanation and regularization are operating on a single shape representation: the SDF. We formulate the problem of interest as building a 3D model in its canonical pose by estimating a 3D deformation field from each new depth frame to the global model and subsequently fusing its data. To this end, we incrementally evolve the projective SDF of the current frame towards the target SDF following a variational framework. The main energy component is a data term which

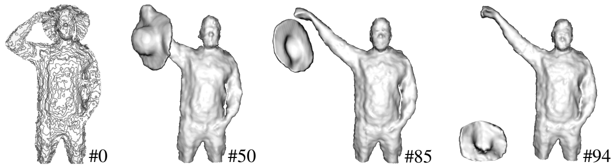


Figure 2. Warped live frames from two-object topological changes.

aligns the current frame to the cumulative model by minimizing their voxel-wise difference of signed distances - thus without explicit correspondence search and suitable for parallelization. In order to handle noise and missing data, we impose smoothness both on the deformation field and on the SDFs, and require a certain level of rigidity. This is done by enforcing the deformation field to be approximately Killing [3, 41, 46] so that it generates locally nearly isometric motions - in analogy to as-rigid-as-possible constraints on meshes [42]. Furthermore, we ensure that the SDF evolution is geometrically correct by conserving the level set property of unity gradient magnitude [26, 35].

To sum up, we contribute a novel variational non-rigid 3D reconstruction system that handles topological changes inherently and circumvents expensive correspondence estimation. Due to the generality of the representation, it can be directly applied to evolving complete meshed models. Last but not least, we propose a methodology for quantifying reconstruction error from a single RGB-D stream.[1]

2. Related Work

Here we discuss existing approaches on level set evolution, vector field estimation and deformable surface tracking in RGB-D data, identifying their limitations in the context of our problem of interest and suggesting remedies.

Level set methods Deformable reconstruction systems commonly rely on meshes for correspondence estimation, making them highly susceptible to errors under larger deformations or topology changes [24]. On the contrary, level sets inherently handle such cases [35]. They have been used for surface manipulation and animation in graphics [6, 14, 47, 53] where models are complete and noise-free, while our goal is incremental reconstruction from noisy partial scans. In medical imaging, where high fidelity shape priors for various organs are available [13, 16], level set methods have been applied to segmentation [2, 17] and registration [25, 30], usually guided by analytically defined evolution equations [36]. However, as we have no template or prior knowledge of the scene, we propose an energy that is driven by the geometry of the SDF and deformation field.

In computer vision, Paragios *et al.* [37] use distance functions for non-rigid registration driven by a vector field, but are limited to synthetic 2D examples. Fujirawa *et al.* [15] discuss extensions of their locally rigid globally non-rigid registration to 3D, but demonstrate only few tests on full surfaces. Instead, we define the energy in 3D and impose rigidity constraints so that 2.5D scans can be fused together from scratch.

Scene flow Determining a vector field that warps 2.5D/3D frames is the objective of works on scene flow [19, 21, 39, 48, 51, 52]. They are typically variational in nature, combining a data alignment term with a smoothness term that ensures that nearby points undergo similar motion. However, this is not sufficient for incremental reconstruction where new frames exhibit previously unseen geometry that has to be overlaid on the model in a geometrically consistent fashion. This is why we include another rigidity prior that requires the field to be approximately Killing - generating nearly isometric motions [3, 41, 46]. In this way we conveniently impose local rigidity through the deformation field, without need for a control grid as in embedded deformation [44] and as-rigid-as-possible modelling [42].

Multiview and template-based surface tracking External constraints help to alleviate the highly unconstrained nature of non-rigid registration. The system of Zollhöfer *et al.* [57] deforms a template to incoming depth frames in real time, but requires the subject to stay absolutely still during the template generation, which cannot be guaranteed when scanning animals or kids. Multi-camera setups are another way to avoid the challenging task of incrementally building a model. Fusion4D [10] recently demonstrated a powerful real-time performance capture system using 24 cameras and multiple GPUs, which is a setup not available to the general user. Moreover, Section 8 of [10] states that even though Fusion4D deals with certain topology changes, the algorithm does not address the problem intrinsically.

Incremental non-rigid reconstruction from a single RGB-D stream The convenience of using a single sensor makes incremental model generation highly desirable. Dou *et al.* [11] proposed a pipeline that achieves impressive quality thanks to a novel non-rigid bundle adjustment, which may last up to 9-10 hours. DynamicFusion [32] was the first approach to simultaneously reconstruct and track the surface motion in real time. VolumeDeform [20] extended the method, combining dense depth-based correspondences with matching of sparse SIFT features across all frames in order to reduce drift and handle tangential motion in scenes of poor geometry. While both works demonstrate compelling results, the shown examples suggest that only relatively controlled motion can be recovered. We aim to utilize the properties of distance fields in order to achieve full evolution under free general motion.

[1] Our data is publicly available at http://campar.in.tum.de/personal/slavcheva/deformable-dataset/index.html.

3. Preliminaries

In the following we define our mathematical notation and outline the non-rigid reconstruction pipeline.

3.1. Notation

Our base representation is a signed distance field (SDF), which assigns to each point in space the signed distance to its closest surface location. One of its characteristic geometric properties is that its gradient magnitude equals unity everywhere where it is differentiable [35]. It is widely used since it can be easily converted to a mesh via marching cubes [29] - the surface is the zero-valued interface between the negative inside and positive outside.

SDF generation is done in a pre-defined volume of physical space, discretized into voxels of a chosen side length. The function $\phi : \mathbb{N}^3 \mapsto \mathbb{R}$ maps grid indices (x, y, z) to the signed distance calculated from the center of the respective voxel. We follow the usual creation process [40, 56], where additionally a confidence weight counting the number of observations is associated with each voxel. We also apply the standard practice of truncating the signed distances. In our case, voxels further than 10 voxels away from the surface are clamped to ± 1. This also serves the purpose of a narrow-band technique, as we only estimate the deformation field over the near-surface non-truncated voxels.

In the given discrete setting, all points in space that belong to a certain voxel obtain the same properties. Thus an index $(x, y, z) \in \mathbb{N}^3$ refers to the whole voxel.

Our goal is to determine a vector field $\Psi\colon \mathbb{N}^3 \mapsto \mathbb{R}^3$ that aligns a pair of SDFs. It assigns a displacement vector (u, v, w) to each voxel (x, y, z). This formulation is similar to VolumeDeform [20] where the deformation field is of the same resolution as the cumulative SDF, while DynamicFusion [32] only has a coarse sparse control grid. However, both require a 6D motion to be estimated per grid point, while a 3D flow field is sufficient in our case due to the dense smooth nature of the SDF representation and the use of alignment constraints directly over the field. Moreover, this makes the optimization process less demanding.

3.2. Rigid Component of the Motion

Although the whole motion from target to reference can be estimated as a deformation, singling out the rigid part of the motion serves as a better initialization. The deformation field is initialized from the previous frame, so we determine frame-to-frame rigid camera motion. We use the SDF-2-SDF registration energy [40] which registers pairs of voxel grids by direct minimization. We prefer this over ICP where the search for point correspondences can be highly erroneous under larger deformation. Nevertheless, any robust rigid registration algorithm of choice can be used instead.

3.3. Overview

We accumulate the model ϕ_{global} in its canonical pose via the weighted averaging scheme of Curless and Levoy [7]. Given a new depth frame D_n, we register it to the previous one and obtain an estimate of its pose relative to the global model. Next, we generate a projective SDF ϕ_n from this pose. The remaining task is to estimate the deformation field Ψ which will best align ϕ_{global} and $\phi_n(\Psi)$, explained in detail in the next section. The field is estimated iteratively and after each step the increment is applied on ϕ_n, updating its values using trilinear interpolation. Once the minimization process converges, we fuse the fully deformed $\phi_n(\Psi)$ into the model via weighted averaging.

The choice to deform the live frame towards the canonical model and not vice versa is based on multiple reasons. On the one hand, this setting is easier for data fusion into the cumulative model. On the other hand, the global SDF has achieved a certain level of regularity after sufficiently many frames have been fused, while a single Kinect depth image is inevitably noisy. Thus, if the model is deformed towards the live frame without imposing enough rigidity, there is a high risk that it would grow into the sensor noise.

4. Non-rigid Reconstruction

In this section we describe our model for determining the vector field Ψ that aligns $\phi_n(\Psi)$ with ϕ_{global}.

4.1. Energy

Our level-set-based, and thus correspondence-free, non-rigid registration energy is defined as follows:

$$E_{\substack{non\\rigid}}(\Psi) = E_{data}(\Psi) + \omega_k E_{Killing}(\Psi) + \omega_s E_{\substack{level\\set}}(\Psi)\,. \quad (1)$$

It consists of a data term and two regularizers whose influence is controlled by the factors ω_k and ω_s.

Data term The main component of our energy follows the reasoning that under perfect alignment, the deformed SDF and the cumulative one would have the same signed distance values everywhere in 3D space. Therefore the flow vector (u, v, w) applied at each voxel (x, y, z) of the current frame's SDF ϕ_n will align it with ϕ_{global}. For brevity we omit the dependence of u, v, w on location:

$$E_{data}(\Psi) = \frac{1}{2}\sum_{x,y,z}\big(\phi_n(x+u, y+v, z+w) - \\ -\phi_{global}(x, y, z)\big)^2\,. \quad (2)$$

Motion regularization To prevent uncontrolled deformations, *e.g.* in case of spurious artifacts caused by sensor noise, we impose rigidity over the motion. Existing approaches typically employ an as-rigid-as-possible [42] or an embedded deformation [44] regularization, which ensures that the vertices of a latent control graph move in an approximately rigid manner. We take a rather different strategy and impose local rigidity directly through the deformation field.

A 3D flow field generating an isometric motion is called a *Killing vector field* [3, 41, 46], named after the mathematician Wilhelm Killing. It satisfies the *Killing condition* $J_\Psi + J_\Psi^\top = \mathbf{0}$, where J_Ψ is the Jacobian of Ψ.

A Killing field is divergence-free, *i.e.* it is volume-preserving, but does not regularize angular motion. A field which generates only nearly isometric motion and thus balances both volume and angular distortion is an *approximately Killing vector field (AKVF)* [41]. It minimizes the Frobenius norm of the Killing condition:

$$E_{\text{AKVF}}(\Psi) = \frac{1}{2} \sum_{x,y,z} ||J_\Psi + J_\Psi^\top||_F^2 \,. \tag{3}$$

However, as we are handling deforming objects, this constraint might be too restrictive. Thus, we propose to damp the Killing condition. In order to do so, we rewrite Eq. 3 using the column-wise stacking operator $vec(\cdot)$:

$$\begin{aligned} E_{\text{AKVF}}(\Psi) &= \frac{1}{2} \sum_{x,y,z} vec(J_\Psi + J_\Psi^\top)^\top vec(J_\Psi + J_\Psi^\top) = \\ &= \sum_{x,y,z} vec(J_\Psi)^\top vec(J_\Psi) + vec(J_\Psi^\top)^\top vec(J_\Psi) \,. \end{aligned} \tag{4}$$

Next, we notice that the first term can be written as:

$$vec(J_\Psi)^\top vec(J_\Psi) = |\nabla u|^2 + |\nabla v|^2 + |\nabla w|^2 \,, \tag{5}$$

which is the typical motion smoothness regularizer used in scene and optical flow [19, 45, 52]. It only encourages that nearby points move in a similar manner, but does not explicitly impose rigid motion. Based on this observation, we devise the damped Killing regularizer

$$\begin{aligned} &E_{Killing}(\Psi) = \\ &= \sum_{x,y,z} \left(vec(J_\Psi)^\top vec(J_\Psi) + \gamma vec(J_\Psi^\top)^\top vec(J_\Psi) \right) \,, \end{aligned} \tag{6}$$

where γ controls the trade-off between Killing property and volume distortion penalization, so that non-rigid motions can also be recovered. A value of $\gamma = 1$ corresponds to the pure Killing condition. We refer the interested reader to the supplementary material for a more detailed derivation.

Level set property To ensure geometric correctness during the evolution of ϕ_n, the property that the gradient magnitude in the non-truncated regions of an SDF is unity has to be conserved [35]:

$$E_{\substack{level\\set}}(\Psi) = \frac{1}{2} \sum_{x,y,z} \left(|\nabla \phi_n(x{+}u, y{+}v, z{+}w)| {-} 1 \right)^2 . \tag{7}$$

It is important to note that a subsequent work of the same authors proposes an improved regularizer for maintaining the level set property [27]. However, it is only useful when the function to be evolved is initialized with a piecewise constant function, and not a signed distance one. As we are initializing ϕ_n with an SDF, the regularizer of Eq. 7 is absolutely sufficient for the considered application.

4.2. Energy Minimization

One of the main benefits of our energy formulations is that it can be applied to each voxel independently, as each term only contains values of the current estimates for the deformation field and SDFs or their derivatives. Therefore the displacement vector updates can be computed in parallel.

We follow a gradient descent scheme. It is variational since Ψ is a function of coordinates in space. Only final results of the Euler-Lagrange equations are presented here, with full derivations given in the supplementary material.

We separate the 3D vector field Ψ into its spatial components, each of which is a scalar field. This allows us to calculate partial derivatives of the energy terms in each spatial direction and to combine them into vectors in order to execute the gradient descent steps.

To ease notation, we will no longer specify summation over voxel indices. Further, we will write $\phi(\Psi)$ instead of $\phi(x{+}u, y{+}v, z{+}w)$ to refer to the value of ϕ after the deformation field has been applied. Note that the summation of integer- and real-valued indices is not problematic, since interpolation is done after every step. We thus obtain the following derivatives with respect to the deformation field:

$$E'_{data}(\Psi) = \left(\phi_n(\Psi) - \phi_{global}\right) \nabla\phi_n(\Psi) \,, \tag{8}$$

$$E'_{Killing}(\Psi) = 2 H_{uvw} \left(vec(J_\Psi^\top) \;\; vec(J_\Psi) \right) \begin{pmatrix} 1 \\ \gamma \end{pmatrix} \,, \tag{9}$$

$$E'_{\substack{level\\set}}(\Psi) = \frac{|\nabla\phi_n(\Psi)| - 1}{|\nabla\phi_n(\Psi)|_\epsilon} \, H_{\phi_n(\Psi)} \, \nabla\phi_n(\Psi) \,. \tag{10}$$

Here $\nabla\phi_n(\Psi) \in \mathbb{R}^{3\times 1}$ is the spatial gradient of the deformed SDF of frame number n and $H_{\phi_n(\Psi)} \in \mathbb{R}^{3\times 3}$ is its Hessian matrix, composed of second-order partial derivatives. Similarly, $H_{uvw} = \begin{pmatrix} H_u & H_v & H_w \end{pmatrix}$ is a 3×9 matrix consisting of the 3×3 Hessians of each component of the deformation field. To avoid division by zero we use $|\cdot|_\epsilon$, which equals the norm plus a small constant $\epsilon = 10^{-5}$.

Finally, we obtain the new state of the deformation field Ψ^{k+1} as a gradient descent step of size α starting from Ψ^k:

$$\Psi^{k+1} = \Psi^k - \alpha \, E'_{\substack{non\\rigid}}(\Psi^k) \,. \tag{11}$$

The field of each incoming frame is initialized with that of the previous frame. Naturally, for the very first frame the initial state is without deformation. Registration is terminated when the magnitude of the maximum vector update in Ψ falls below a threshold of 0.1 mm.

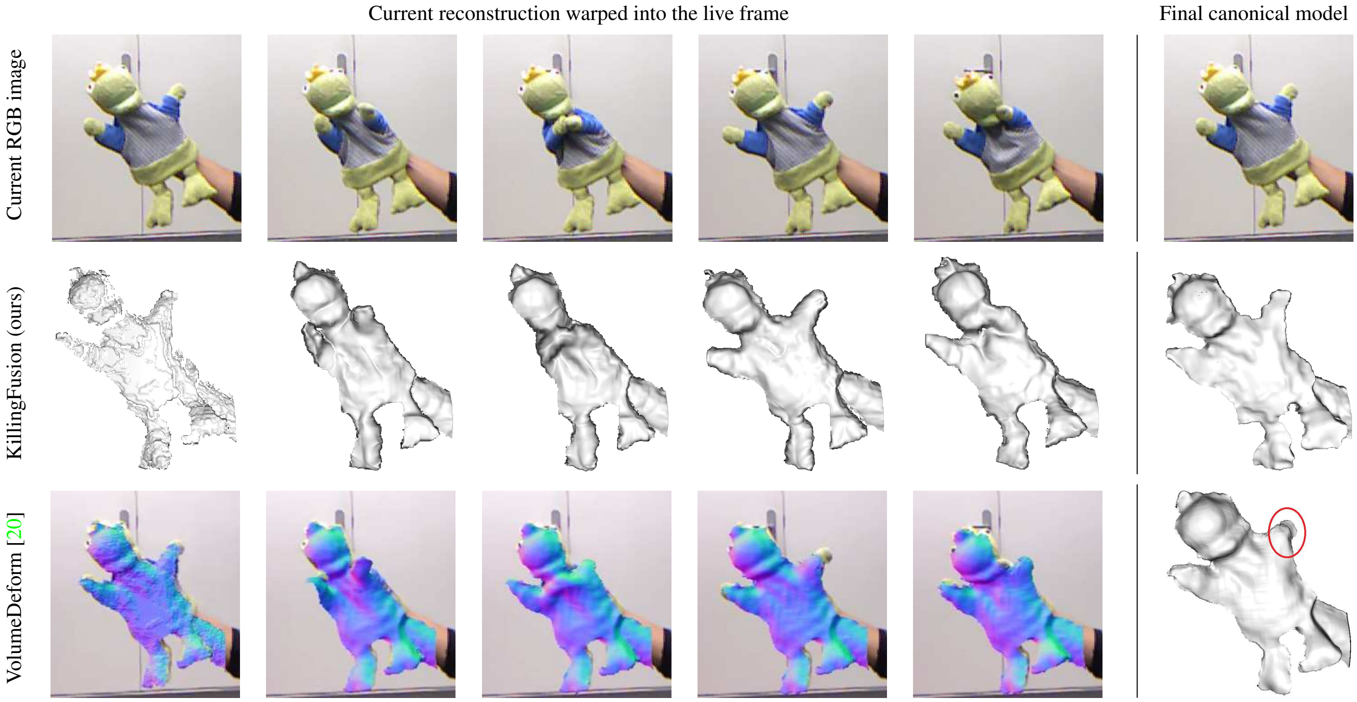


Figure 3. Comparison under **topological changes**. Our level-set-based KillingFusion fully evolves into the correct geometric shape between frames, while VolumeDeform [20] does so only partially (3rd and 5th live frames), which is reflected as artifacts in the final reconstruction.

4.3. Implementation Details

Equations 8-10 are highly suitable for parallelization as the update for each voxel depends only on its immediate neighbourhood. Thus we opted for a GPU implementation, which we tested on an NVIDIA Quadro K2100M. It runs at 3-30 frames per second for all shown examples. In particular, it takes 33 ms for a grid consisting of approximately 80^3 voxels. Naturally, speed decreases with increasing grid resolution. However, the slowdown is not cubic, since only the near-surface voxels contribute for the deformation field estimation, which typically constitute less than 10% of all.

5. Results

This section contains qualitative and quantitative evaluation of the proposed non-rigid reconstruction framework. The parameters were fixed as follows: gradient descent step $\alpha = 0.1$, damping factor for the Killing energy $\gamma = 0.1$, weights for the motion and level set regularization respectively $\omega_k = 0.5$, $\omega_s = 0.2$. The choice of values for ω_s and ω_k not only balances their influence, but also acts as normalization since signed distances are truncated to the interval $[-1; 1]$, while the deformation field contains vectors spanning up to several voxels. We used a voxel size of 8 mm for human-sized subjects and 4 mm for smaller-scale ones.

Changing topology and large inter-frame motion The first experiments that we carried out focus on highlighting the strengths of our KillingFusion compared to other single-stream deformable reconstruction pipelines: changing topology and rapid motion between frames. To be able to quantify results, we used mechanical toys that can both deform and move autonomously. We first reconstructed them in their static rest pose using a markerboard for external ground-truth pose estimation. Then we recorded their non-rigid movements starting from the rest pose, which lets us evaluate the error in the canonical-pose reconstruction.

We shared our recordings with the authors of VolumeDeform [20], who kindly run the *Frog*, *Duck* and *Snoopy* sequences and gave us their final canonical-pose reconstructions and videos of the model warped onto the live images.

Figures 3 and 4 juxtapose our results. Note that the reconstructions are partial because these objects do not complete 360° loops. Both approaches perform well under general motion. However, the third and fifth *Frog* live frames demonstrate that VolumeDeform, as an example of a method that determines mesh-based correspondences, does not track topological changes. Similarly, the latter three *Snoopy* live frames show that it cannot recover once a topological change occurs when the feet touch. Furthermore, the rapid ear motion, making a full revolution from horizontal to vertical position and back within 5 frames, cannot be captured and causes artifacts in the final reconstruction, while our level-set based KillingFusion fully evolves the surface even in such cases. Thus SDFs are better suited for overcoming large inter-frame motion and changing topology.

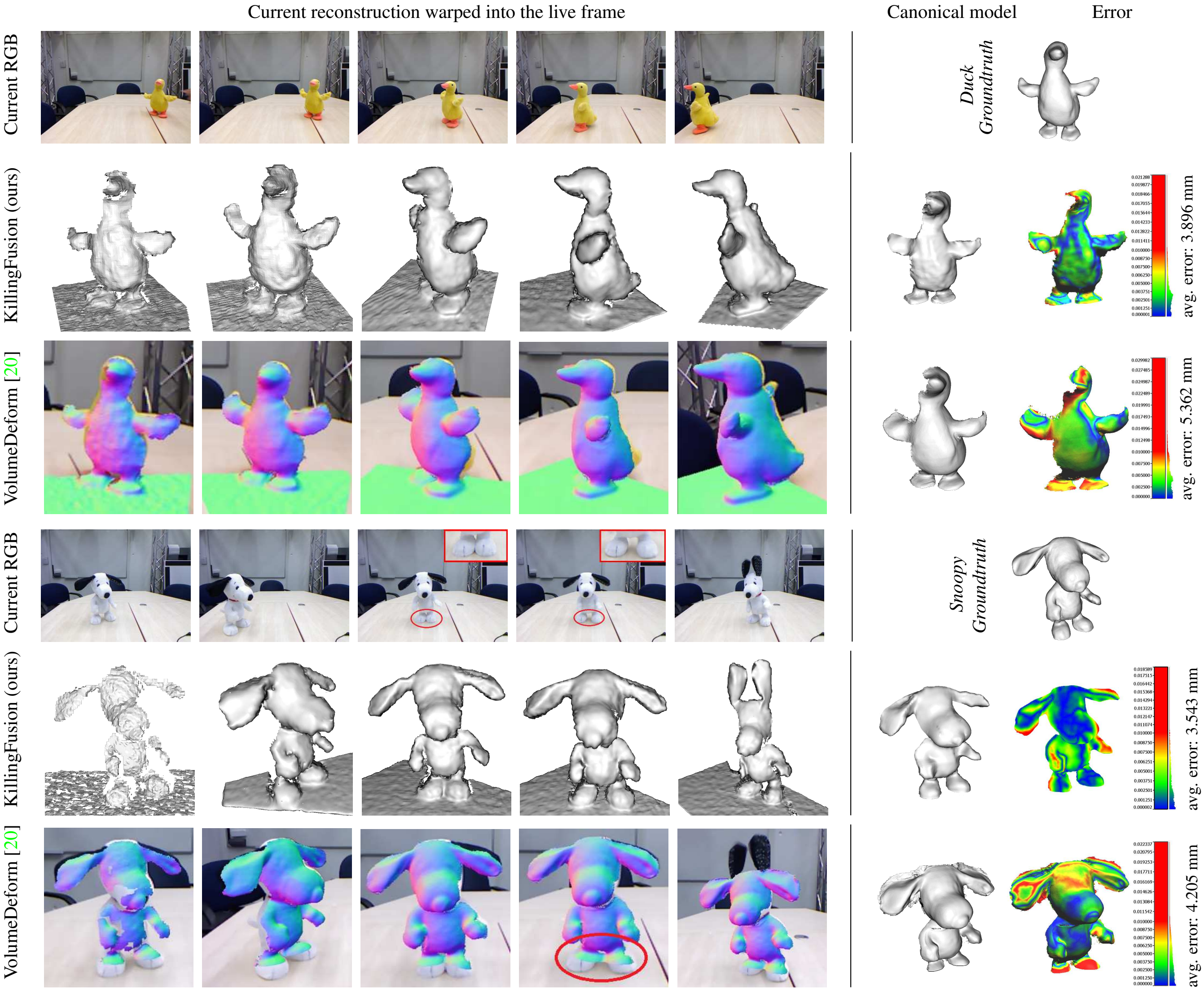


Figure 4. Comparison of KillingFusion to VolumeDeform [20] under **rapid motion** and **topological changes**. *Duck*'s wings and *Snoopy*'s ears make a complete up-down revolution within 5 frames, and *Snoopy*'s feet touch and separate several times. While a mesh-based method does not handle such motions, our SDF-based approach fully captures the deformations. This is reflected in less artifacts and lower error in the final model. Live frames are in chronological order, the objects do not complete 360° loops. Red is saturated at 1 cm in all error plots.

The last column of Figure 4 contains snapshots from the evaluation of the canonical-pose outputs against the groundtruth in *CloudCompare*[2]. Our models tend to be less detailed than those of VolumeDeform due to the coarse voxel resolution. However, we achieve higher geometric consistency: our average errors are 3.5 mm on *Snoopy* and 3.9 mm on *Duck*, while those of VolumeDeform are 4.2 mm and 5.4 mm respectively. Note that the voxel size we used is 4 mm, indicating that our accuracy stays within its limits. As expected, KillingFusion is closer to the groundtruth model in the areas of fast motion, while VolumeDeform has accumulated artifacts there.

Figure 5. Canonical-pose result on a 360° sequence: KillingFusion reconstructs a complete, geometrically consistent model.

Finally, in Fig. 5 we scanned another object, which completes a full 360° loop while moving non-rigidly, in order to demonstrate our capabilities to incrementally build a complete water-tight model from scratch. The reconstruction error remained of the same order as for the partial view scans.

[2]CloudCompare - 3D Point Cloud and Mesh Processing Software, http://www.danielgm.net/cc/.

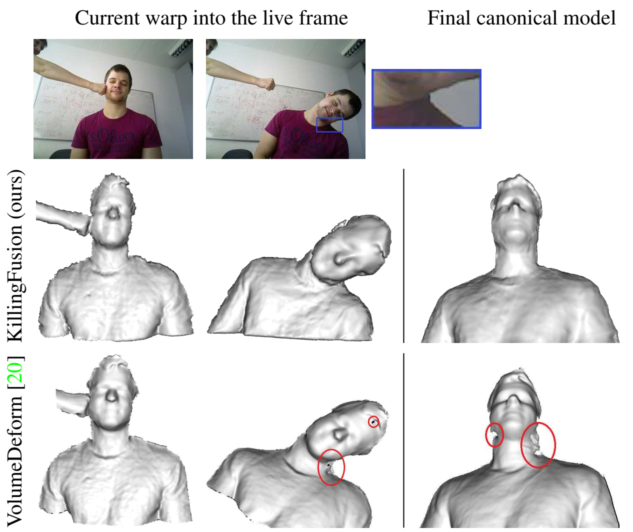


Figure 6. Comparison of our depth-only KillingFusion to VolumeDeform [20] which additionally relies on the color frames for SIFT matching: our reconstructions are of comparable fidelity. In particular, our canonical model exhibits less artifacts where larger motion occurred, *e.g*. around the neck which bends over $90°$. Moreover, our live frames show that KillingFusion follows the folds of the neck more naturally (see marked regions).

Public single-stream RGB-D datasets Next, we tested KillingFusion on the datasets used in related single-stream non-rigid reconstruction works. We chose the sequences that we identify as most challenging, *i.e*. exhibiting large deformations and completing a full loop in front of the camera, where available.

First, we tested KillingFusion on data from the VolumeDeform publication [20]. The authors have also made publicly available their canonical-pose and warped reconstructions for every 100^{th} frame. The comparison in Figure 6 shows that KillingFusion achieves similar quality. Notably, the second warped frame demonstrates that our SDFs deform to the geometry more naturally: our warped model replicates the skin folding around the neck, while the model of VolumeDeform does not bend further than a certain extent, causing artifacts in the final reconstruction as well. This is similar to the behaviour we observed on our own rapid motion recordings. In conclusion, another dataset also indicates that level set evolution allows to capture larger motion better than mesh-based techniques.

Next, we run KillingFusion on $360°$ sequences used in Dou *et al*.'s offline non-rigid bundle adjustment paper [11] and DynamicFusion [32]. As we do not have the authors' resulting meshes, we show snapshots available from the publications. KillingFusion manages to recover a complete model of comparable fidelity to the other techniques. In particular, despite the coarse voxel resolution, it preserves fine-scale details such as noses, ears and folds on shirts after a full loop around the subject.

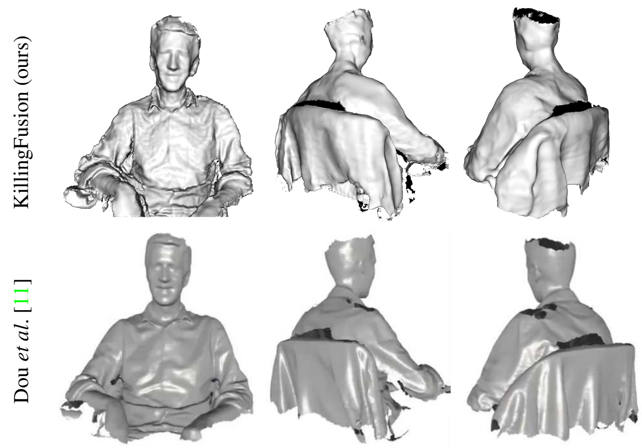


Figure 7. Comparison to the offline bundle adjustment method of Dou *et al*. [11]: our KillingFusion achieves similar quality at real time, preserving fine structures, such as shirt folds and the nose, after a full loop around the subject.

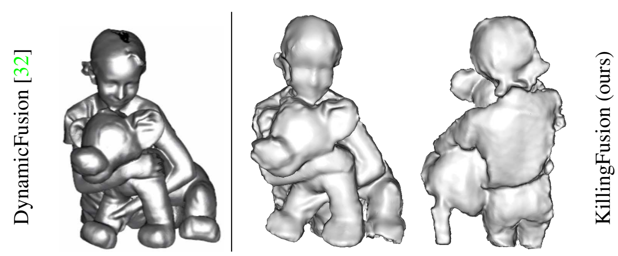
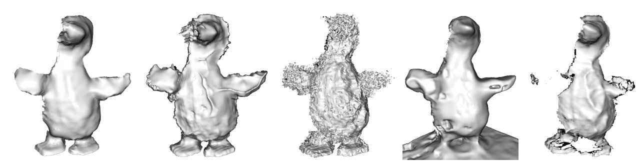


Figure 8. KillingFusion result on the full-loop *Squeeze* sequence from DynamicFusion [32], showing front and back of the canonical-pose reconstruction.



(a) all terms (b) $\omega_s = 0$ (c) $\omega_k = 0$ (d) $\gamma = 0$ (e) $\gamma = 1$

Figure 9. Evaluation of energy component effects. (a) Standard parameter setting. (b) No level set property preservation. (c) No motion regularization. (d) Conventional motion smoothness without a Killing component. (e) Pure Killing condition.

Contributions of energy components In order to confirm that all regularizers from our non-rigid energy formulation are essential, we studied their effects in Fig. 9. The model is not smooth and fine artifacts, visible as small holes, appear without the level set property (Fig. 9b), because it has been violated in places during the SDF evolution. Without motion regularization (Fig. 9c), the moving parts of the object, such as the wings and head, get destroyed as more frames are fused. In case of applying standard motion smoothness, without enforcing divergence-free Killing behaviour (Fig. 9d), the model is somewhat

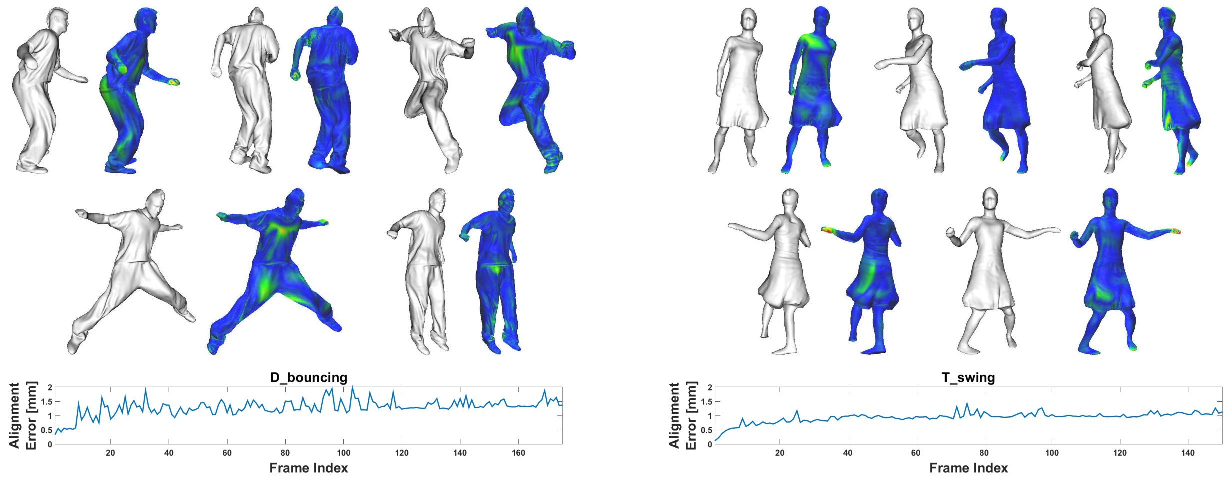


Figure 10. Non-rigid registration of complete 3D shapes from the MIT dataset [49]. Starting with an initial SDF, we gradually evolve it to match every next model in the sequence. Each pair shows our reconstruction along with its corresponding error plot (scale same as before).

smoother, but in several regions the geometry between different frames is inconsistent, resulting in holes. Conversely, if we do not damp the Killing condition (Fig. 9e) and thus the energy steers towards completely rigid motion, the non-rigidly moving wings almost vanish. We empirically determined favourable values for γ to be between 0.05 and 0.3.

Multiview mesh datasets To show the generality of our SDF-based approach, we run KillingFusion on the MIT multiview mesh dataset [49], as done by Zollhöfer *et al*. [57]. It contains several sequences of 150-200 meshes, fused from multiview captures around people who are executing movements with considerably large deformation. Therefore it also permits another quantitative evaluation.

Figure 10 shows our reconstructions throughout the sequences, together with the alignment error indicating the deviation from the ground truth. We started with an SDF initialized from the first mesh and continuously evolve it towards the SDF corresponding to every next frame. While the error tends to slightly increase over time, the effects of drift accumulation are not severe. The model error remains below 2 mm throughout both sequences, with an average of 1.3 mm in *D_bouncing* and 0.9 mm in *T_swing*. We included one of the dancing girl sequences, as they are typically used in literature to demonstrate problems with topology changes when the dress touches the legs [11] - but do not cause a problem for KillingFusion. In particular, we notice no larger artifacts near the dress edge than other areas of the model. The biggest errors are, in fact, typically near the hands of the subjects. This is because the used voxel size of 8 mm does not always manage to recover fine structures like the fingers with absolute accuracy. Last but not least, we noticed that if instead we deform the first SDF to every frame, more iterations are required to converge, but the errors do not change significantly.

6. Limitations and Future Work

The primary aim of our non-rigid reconstruction system is to recover the 3D shape of the deforming object. As this is done via level set evolution rather than by determining the new position of each point, applications which require explicit point correspondences, such as texture mapping, fall out of the scope of our approach. Thus we plan to integrate backward tracking of point correspondences in level sets [38] in order to open up further possibilities. Moreover, we plan to explore representing the flow field at a coarser resolution grid using interpolation of radial basis functions [55], so that a larger volume can be covered.

7. Conclusion

We have presented a novel framework for non-rigid 3D reconstruction that inherently handles changing topology and is able to capture rapid motion. Our lightweight energy formulation allows to determine dense deformation flow field updates without correspondence search, based on a combination of a newly introduced damped Killing motion constraint and level set validity regularization. A variety of qualitative and quantitative examples have shown that KillingFusion can recover the geometry of objects undergoing diverse kinds of deformations. We believe our contribution is a step forward towards making real-time recovery of unconstrained motion truly available to the general user.

Acknowledgements We thank Matthias Innmann for running VolumeDeform on our data; Mohamed Souiai, Gabriel Peyré and Chun-Hao Huang for valuable discussions; and Alexander Seeber for creative assistance in the recordings. M. Baust acknowledges the support of DFG-funded Collaborative Research Centre SFB824-Z2 *Imaging for Selection, Monitoring and Individualization of Cancer Therapies*.

References

[1] B. Allain, J. Franco, and E. Boyer. An Efficient Volumetric Framework for Shape Tracking. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[2] E. Angelini, Y. Jin, and A. Laine. State of the Art of Level Set Methods in Segmentation and Registration of Medical Imaging Modalities. *Handbook of Biomedical Image Analysis: Registration Models*, III, 2005. 2

[3] M. Ben-Chen, A. Butscher, J. Solomon, and L. Guibas. On Discrete Killing Vector Fields and Patterns on Surfaces. *Computer Graphics Forum (CGF)*, 29(5), 2010. 2, 4

[4] C. Cagniart, E. Boyer, and S. Ilic. Iterative Deformable Surface Tracking in Multi-View Setups. In *5th International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, 2010. 1

[5] S. Choi, Q.-Y. Zhou, and V. Koltun. Robust Reconstruction of Indoor Scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[6] D. Cohen-Or, A. Solomovic, and D. Levin. Three-dimensional Distance Field Metamorphosis. *ACM Transactions on Graphics (TOG)*, 17(2):116–141, 1998. 2

[7] B. Curless and M. Levoy. A Volumetric Method for Building Complex Models from Range Images. In *23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pages 303–312, 1996. 1, 3

[8] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H. Seidel, and S. Thrun. Performance Capture from Sparse Multi-view Video. *ACM Transactions on Graphics (TOG)*, 27(3), 2008. 1

[9] M. Dou, H. Fuchs, and J. Frahm. Scanning and tracking dynamic objects with commodity depth cameras. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2013. 1

[10] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. Fanello, A. Kowdle, S. Escolano, C. Rhemann, D. Kim, J. Taylor, P. Kohli, V. Tankovich, and S. Izadi. Fusion4D: Real-time Performance Capture of Challenging Scenes. *ACM Transactions on Graphics (TOG)*, 35(4):114, 2016. 1, 2

[11] M. Dou, J. Taylor, H. Fuchs, A. Fitzgibbon, and S. Izadi. 3D Scanning Deformable Objects with a Single RGBD Sensor. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 7, 8

[12] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard. An Evaluation of the RGB-D SLAM System. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2012. 1

[13] A. Frangi, D. Rueckert, J. Schnabel, and W. Niessen. Automatic Construction of Multiple-Object Three-Dimensional Statistical Shape Models: Application to Cardiac Modeling. *IEEE Transactions on Medical Imaging (TMI)*, 21(9):1151–1166, 2002. 2

[14] S. F. Frisken and R. N. Perry. Designing with Distance Fields. In *ACM SIGGRAPH 2006 Courses*, SIGGRAPH '06, pages 60–66, 2006. 2

[15] K. Fujiwara, K. Nishino, J. Takamatsu, B. Zheng, and K. Ikeuchi. Locally Rigid Globally Non-rigid Surface Registration. In *IEEE International Conference on Computer Vision (ICCV)*, 2011. 2

[16] B. v. Ginneken, A. Frangi, J. Staal, B. Romeny, and M. Viergever. Active Shape Model Segmentation with Optimal Features. *IEEE Transactions on Medical Imaging (TMI)*, 21(8):924–933, 2002. 2

[17] S. Ho, E. Bullitt, and G. Gerig. Level-Set Evolution with Region Competition: Automatic 3-D Segmentation of Brain Tumors. In *16th International Conference on Pattern Recognition (ICPR)*, 2002. 2

[18] C. Huang, C. Cagniart, E. Boyer, and S. Ilic. A Bayesian Approach to Multi-view 4D Modeling. *International Journal of Computer Vision (IJCV)*, 116(2):115–135, 2016. 1

[19] F. Huguet and F. Devernay. A Variational Method for Scene Flow Estimation from Stereo Sequences. In *IEEE International Conference on Computer Vision (ICCV)*, 2007. 2, 4

[20] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger. VolumeDeform: Real-time Volumetric Non-rigid Reconstruction. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 2, 3, 5, 6, 7

[21] M. Jaimez, M. Souiai, J. Gonzalez-Jimenez, and D. Cremers. A Primal-Dual Framework for Real-Time Dense RGB-D Scene Flow. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2015. 2

[22] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[23] O. Kähler, V. A. Prisacariu, C. Y. Ren, X. Sun, P. Torr, and D. Murray. Very High Frame Rate Volumetric Integration of Depth Images on Mobile Devices. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 21(11):1241–1250, 2015. 1

[24] O. Karpenko, J. Hughes, and R. Raskar. Free-form Sketching with Variational Implicit Surfaces. *Computer Graphics Forum (CGF)*, 21(3):585–594, 2002. 2

[25] T. Lee and S. Lai. 3D Non-rigid Registration for MPU Implicit Surfaces. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2008. 2

[26] C. Li, C. Xu, C. Gui, and M. D. Fox. Level Set Evolution Without Re-initialization: A New Variational Formulation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 2

[27] C. Li, C. Xu, C. Gui, and M. D. Fox. Distance Regularized Level Set Evolution and Its Application to Image Segmentation. *IEEE Transaction on Image Processing (TIP)*, 19(12):3243–3254, 2010. 4

[28] H. Li, B. Adams, L. Guibas, and M. Pauly. Robust Single-View Geometry and Motion Reconstruction. *ACM Transactions on Graphics (TOG)*, 28(5), 2009. 1

[29] W. E. Lorensen and H. E. Cline. Marching Cubes: A High Resolution 3D Surface Construction Algorithm. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '87, pages 163–169, 1987. 3

[30] J. Maintz and M. Viergever. A Survey of Medical Image Registration. *Medical Image Analysis*, 2(1):1–36, 1998. 2

[31] M. Meilland and A. I. Comport. On Unifying Key-frame and Voxel-based Dense Visual SLAM at Large Scales. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013. 1

[32] R. A. Newcombe, D. Fox, and S. M. Seitz. DynamicFusion: Reconstruction and Tracking of Non-rigid Scenes in Real-Time. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 3, 7

[33] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: Real-Time Dense Surface Mapping and Tracking. In *10th International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011. 1

[34] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3D Reconstruction at Scale using Voxel Hashing. *ACM Transactions on Graphics (TOG)*, 2013. 1

[35] S. Osher and R. Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*, volume 153 of *Applied Mathematical Science*. Springer, 2003. 2, 3, 4

[36] S. Osher and J. Sethian. Fronts Propagating with Curvature-dependent speed: Algorithms based on Hamilton-Jacobi Formulations. *Journal of Computational Physics*, 79(1):12–49, 1988. 2

[37] N. Paragios, M. Rousson, and V. Ramesh. Non-rigid Registration Using Distance Functions. *Computer Vision and Image Understanding (CVIU)*, 89(2-3):142–165, 2003. 2

[38] J. Pons, G. Hermosillo, R. Keriven, and O. Faugeras. How to Deal with Point Correspondences and Tangential Velocities in the Level Set Framework. In *9th IEEE International Conference on Computer Vision (ICCV)*, 2003. 8

[39] J. Quiroga, T. Brox, F. Devernay, and J. Crowley. Dense Semi-rigid Scene Flow Estimation from RGBD Images. In *European Conference on Computer Vision (ECCV)*, 2014. 2

[40] M. Slavcheva, W. Kehl, N. Navab, and S. Ilic. SDF-2-SDF: Highly Accurate 3D Object Reconstruction. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[41] J. Solomon, M. Ben-Chen, A. Butscher, and L. Guibas. As-Killing-As-Possible Vector Fields for Planar Deformation. *Computer Graphics Forum (CGF)*, 30(5), 2011. 2, 4

[42] O. Sorkine and M. Alexa. As-Rigid-As-Possible Surface Modeling. In *Fifth Eurographics Symposium on Geometry Processing (SGP)*, 2007. 2, 3

[43] F. Steinbrücker, C. Kerl, J. Sturm, and D. Cremers. Large-Scale Multi-Resolution Surface Reconstruction from RGB-D Sequences. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. 1

[44] R. W. Sumner, J. Schmid, and M. Pauly. Embedded Deformation for Shape Manipulation. *ACM Transactions on Graphics (TOG)*, 26(3), 2007. 2, 4

[45] D. Sun, S. Roth, and M. J. Black. Secrets of Optical Flow Estimation and Their Principles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 4

[46] M. Tao, J. Solomon, and A. Butscher. Near-Isometric Level Set Tracking. *Computer Graphics Forum (CGF)*, 35(5), 2016. 2, 4

[47] G. Turk and J. O'Brien. Shape Transformation Using Variational Implicit Functions. In *26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, 1999. 2

[48] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-Dimensional Scene Flow. In *IEEE International Conference on Computer Vision (ICCV)*, 1999. 2

[49] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated Mesh Animation from Multi-view Silhouettes. *ACM Transactions on Graphics (TOG)*, 27(3), 2008. 8

[50] 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

[51] C. Vogel, K. Schindler, and S. Roth. Piecewise Rigid Scene Flow. In *IEEE International Conference on Computer Vision (ICCV)*, 2013. 2

[52] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, U. Franke, and D. Cremers. Efficient Dense Scene Flow from Sparse or Dense Stereo Data. In *10th European Conference on Computer Vision (ECCV)*, 2008. 2, 4

[53] Y. Weng, M. Chai, W. Xu, Y. Tong, and K. Zhou. As-Rigid-As-Possible Distance Field Metamorphosis. *Computer Graphics Forum (CGF)*, 32(7):381–389, 2013. 2

[54] T. Whelan, S. Leutenegger, R. F. Salas-Moreno, B. Glocker, and A. J. Davison. ElasticFusion: Dense SLAM Without A Pose Graph. In *Robotics: Science and Systems (RSS)*, 2015. 1

[55] X. Xie and M. Mirmehdi. Radial Basis Function Based Level Set Interpolation and Evolution for Deformable Modelling. *Image and Vision Computing (IVC)*, 29(2-3):167–177, 2011. 8

[56] C. Zach, T. Pock, and H. Bischof. A Globally Optimal Algorithm for Robust TV-L^1 Range Image Integration. In *Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007. 3

[57] M. Zollhöfer, M. Nießner, S. Izadi, C. Rhemann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt, and M. Stamminger. Real-time Non-rigid Reconstruction using an RGB-D Camera. *ACM Transactions on Graphics (TOG)*, 33(4), 2014. 1, 2, 8