NRMVS: Non-Rigid Multi-View Stereo

Matthias Innmann[1,2*] Kihwan Kim[1] Jinwei Gu[1,3*] Matthias Nießner[4]
Charles Loop[1] Marc Stamminger[2] Jan Kautz[1]

[1]NVIDIA [2]Friedrich-Alexander University Erlangen-Nürnberg [3]SenseTime [4]Technical University of Munich

Abstract

Multi-view Stereo (MVS) is a common solution in photogrammetry applications for the dense reconstruction of a static scene from images. The static scene assumption, however, limits the general applicability of MVS algorithms, as many day-to-day scenes undergo non-rigid motion, e.g., clothes, faces, or human bodies. In this paper, we open up a new challenging direction: Dense 3D reconstruction of scenes with non-rigid changes observed from a small number of images sparsely captured from different views with a single monocular camera, which we call non-rigid multi-view stereo (NRMVS) problem. We formulate this problem as a joint optimization of deformation and depth estimation, using deformation graphs as the underlying representation. We propose a new sparse 3D to 2D matching technique with a dense patch-match evaluation scheme to estimate the most plausible deformation field satisfying depth and photometric consistency. We show that a dense reconstruction of a scene with non-rigid changes from a few images is possible, and demonstrate that our method can be used to interpolate novel deformed scenes from various combinations of deformation estimates derived from the sparse views.

1. Introduction

Surface reconstruction of dynamic objects from sparse spatial and temporal views is an unsolved and seemingly ill-posed problem. Existing work addresses related problems with additional constraints, e.g. static geometry, or with additional structure, e.g. depth maps or dense video frames. We demonstrate for the first time, that it is possible to jointly solve for depth and shape deformation from sparse spatial and temporal views. We believe this is crucial for everyday capture scenarios where a single high resolution still camera is used to image an object from multiple viewpoints, and the static behavior of the subject matter cannot be controlled.

For static scenes, multi-view stereo algorithms [5, 7, 17, 19, 48, 50] have played an important role in dense 3D reconstruction on top of estimated camera poses and sparse points from structure from motion (SfM) [47] algorithms. They are the keys to understand a scene for augmented reality, robotics, and autonomous driving. However, if a scene contains motion, such as non-stationary rigid objects or non-rigid surface deformations, the assumption of an epipolar constraint is violated [22], causing algorithms to fail in reconstructing most non-static regions. We observe that camera poses can be computed with sufficient static regions [44]. Scenes with rigidly-moving objects have been reconstructed by segmenting foreground objects from the background, and treating these regions independently [62, 29, 60]. For dynamic scenes with abundant non-rigid changes, to compute camera poses and (often) sparse points with deformation, various non-rigid structure from motion (NRSfM) methods have been introduced [25]. These methods often require either dense views (video frames) [3, 10] for the acquisition of dense tracks of correspondences, or prior information to constrain the problem [9]. Newcombe et al. [43] and Innmann et al. [24] recently demonstrated solutions for the 3D reconstruction of arbitrary, non-rigid, dynamic scenes using a dense stream of metric depths cap-

Figure 1: We take a small number of input images with non-rigid changes captured from different views (top row). We then reconstruct the 3D geometry of objects undergoing deformation. We first triangulate a canonical surface from a pair of views with minimal deformation (a). Then we compute the deformation from the canonical view to the other views, and reconstruct point clouds for both original canonical views (b) and other remaining views (c).

*The authors contributed to this work when they were at NVIDIA.

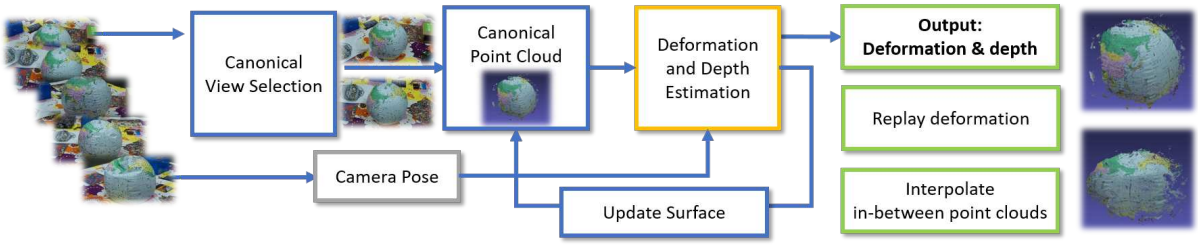


Figure 2: **Overview of our framework**: We first reconstruct the initial point cloud from the views with minimal deformation, which we call canonical views. Then, we estimate depths for the remaining views by estimating a plausible d eformation from a joint optimization between depth, deformation, and dense photometric consistency. With these computed depths and deformations, we also demonstrate an interpolation deformation between input views.

tured from a commercial depth camera.

In this paper, we are specifically interested in *dense 3D reconstruction of scenes with dynamic non-rigid deformations* from a small number of images sparsely captured from different views at different times. We refer to this as non-rigid multi-view stereo (NRMVS). This requires two solutions: (1) a method to compute the most plausible deformation from all potential deformations between a given baseline of the views, and (2) a dense 3D reconstruction algorithm that satisfies a photometric consistency constraint between images of surfaces undergoing non-rigid changes.

In our solution, we first compute a *canonical point cloud* from views that have minimal relative deformation (Fig. 1(a)), and then estimate the deformation between the canonical pose and other views by a joint optimization of depth and photometric consistency. During this process, the 3D point cloud of the canonical pose is continuously expanded (Fig. 1(b)). Then, through the individual deformation fields estimated from each view to the canonical point cloud, we can reconstruct a dense 3D point cloud of each single view (Fig. 1(c)). A brief overview of our entire framework is described in Fig. 2.

Our contributions are as follows:

- The first non-rigid MVS pipeline that densely reconstructs dynamic 3D scenes with non-rigid changes from sparse RGB views.
- A new formulation to model non-rigid motion using a deformation graph [55] and the approximation of the inverse-deformation used for the joint optimization to maximize photometric consistency along with a practical scheme to filter sparse feature matches in the presence of non-rigid scene motion.
- Patchmatch-based [5] dense sample propagation on top of an existing MVS pipeline [48], which allows flexible implementation depending on different MVS architectures.

2. Related Work

Dynamic RGB-D Scene Reconstruction. A prior step to full dynamic scene reconstruction is dynamic template tracking of 3D surfaces. The main idea is to track a shape template over time while non-rigidly deforming its surface [11, 2, 21, 23, 35, 32, 34, 18, 63]. Jointly tracking and reconstructing a non-rigid surface is significantly more challenging. In this context, researchers have developed an impressive line of works based on RGB-D or depth-only input [61, 41, 56, 6, 12, 14, 40, 59, 36]. DynamicFusion [43] optimizes a Deformation Graph [55] using dual-quaternion blending [27], then fuses the deformed surface with the current depth map. Innmann et al. [24] follows up on this work by using an as-rigid-as-possible regularizer to represent deformations [53], and incorporate RGB features in addition to a dense depth tracking term. Fusion4D [13] brings these ideas a level further by incorporating a high-end RGB-D capture setup, which achieves very impressive results. More recent RGB-D non-rigid fusion frameworks include KillingFusion [51] and SobolevFusion [52], which allow for implicit topology changes using advanced regularization techniques. This line of research has made tremendous progress in the recent years; but given the difficulty of the problem, all these methods either rely on depth data or calibrated multi-camera rigs.

Multi-View Stereo. Various MVS approaches for dense 3D scene reconstruction have been introduced in the last few decades [17, 19, 48, 50]. These methods commonly utilize camera poses and sparse 3D points estimated from SfM [47, 58, 26, 57] algorithms for further dense evaluation of the scene. A recent survey [49] on MVS shows that COLMAP [48] performs best among state-of-the-art methods. Therefore, in this paper, we adopt COLMAP's Patchmatch framework for dense photometric consistency. While these methods work well for static scenes, they often failed to reconstruct the regions that do not follow the epipolar geometry assumption [22]. Thus, even if there are sufficient static regions to estimate camera poses from SfM, generic MVS methods do not complete the reconstruction on the surfaces with non-rigid changes. Non-rigid registration is utilized to improve static 3D reconstructions [45].

Dynamic Scenes and NRSfM. Non-rigid structure from motion methods [25, 28, 46] tackle more challenging prob-

lems of estimating camera poses of the images mostly observing points with non-rigid motion. Hence, with a successful camera pose estimation and non-rigidly moving sparse points, our non-rigid dense reconstruction method could directly run on top of such methods. However, due to the inherent ill-posed property of the problem, most of the NRSfM methods require large number of video frames to use a dense track of correspondences [20, 3], prior shape/motion information [9, 4, 39] or multiple synchronized videos/images [42, 30, 8, 31].

In our work, we focus on the dense 3D reconstruction with joint optimization of depth and estimated deformation from only a few images. Thus, we assume that there are sufficient static regions that allow us to estimate camera poses even with standard SfM frameworks [1, 47, 15]. While more challenging dynamic scenes with insufficient static regions could be handled by advances in NRSfM for few images, it is beyond the scope of our approach.

3. Approach

The input to our NRMVS method is a set of images of a dynamic object taken from a single monocular camera in unique locations at different times. We do not assume any knowledge of temporal order, i.e., the images are an unorganized collection. However, we assume there are at least two images with minimal deformation, and the scene contains sufficient background in order to measure the ratio of non-rigidity and to recover the camera poses. We discuss details later in Sec. 3.3. The output of our method is an estimate of the deformation within the scene from the canonical pose to every other view, as well as a depth map for each view. After the canonical view selection, we reconstruct an initial canonical point cloud that serves as a template for the optimization. Given another arbitrary input image and its camera pose, we estimate the deformation between the canonical point cloud and the input. Furthermore, we compute a depth map for this processed frame using a non-rigid variant of PatchMatch. Having estimated the motion and the geometry for every input image, we recompute the depth for the entire set of images to maximize the growth of the canonical point cloud. Fig. 2 shows an overview of our method.

3.1. Modeling Deformation in Sparse Observations

To model the non-rigid motion in our scenario, we use the well known concept of deformation graphs [55]. Each graph node represents a rigid body transform, similar to the as-rigid-as-possible deformation model [53]. These transforms are locally blended to deform nearby space.

Given a point $\mathbf{v} \in \mathbb{R}^3$, its deformed version $\hat{\mathbf{v}}$ is:

$$\hat{\mathbf{v}} = \sum_{i=1}^{k} w_i(\mathbf{v}) \left[\mathbf{R}_i(\mathbf{v} - \mathbf{g}_i) + \mathbf{g}_i + \mathbf{t}_i\right],$$

Figure 3: **Deformation nodes and correspondences:** (Left) The deformation nodes at t_0 (orange), and another set of nodes at t_1 (red) overlaid in the canonical view. (Right) The relationship between deformation nodes from two views and sparse 3D matches (after lifting) in the context of a non-rigid change. Note that we only show the sparse matching for simpler visualization while there is also a dense term for photometric consistency that drives the displacement of deformation nodes with the sparse matches.

where $\mathbf{R}_i$ and $\mathbf{t}_i$ represent the rotation and translation of a rigid body transform about position $\mathbf{g}_i$ of the i-nearest deformation node, and k is the user-specified number of nearest neighbor nodes (we set $k = 4$ throughout our paper). The weights w_i are defined as:

$$w_i(\mathbf{v}) = \frac{1}{\sum_{j=1}^{k} w_j(\mathbf{v})} \left(1 - \frac{\|\mathbf{v} - \mathbf{g}_i\|_2}{\|\mathbf{v} - \mathbf{g}_{k+1}\|_2}\right)^2.$$

For a complete description of deformation graphs, we refer to the original literature [55].

When projecting points between different images, we also need to invert the deformation. The exact inverse deformation can be derived given known weights:

$$\mathbf{v} = \left(\sum_{i=1}^{k} w_i(\mathbf{v})\mathbf{R}_i\right)^{-1} \left[\hat{\mathbf{v}} + \sum_{i=1}^{k} w_i(\mathbf{v}) \left[\mathbf{R}_i\mathbf{g}_i - \mathbf{g}_i - \mathbf{t}_i\right]\right]$$

However, because we do not know the weights a priori, which requires the nearest neighbor nodes and their distances, this becomes a non-linear problem. Since this computationally expensive step is necessary at many stages of our pipeline, we introduce an approximate solution:

$$\mathbf{v} \approx \left(\sum_{i=1}^{k} \hat{w}_i(\hat{\mathbf{v}})\mathbf{R}_i\right)^{-1} \left[\hat{\mathbf{v}} + \sum_{i=1}^{k} \hat{w}_i(\hat{\mathbf{v}}) \left[\mathbf{R}_i\mathbf{g}_i - \mathbf{g}_i - \mathbf{t}_i\right]\right],$$

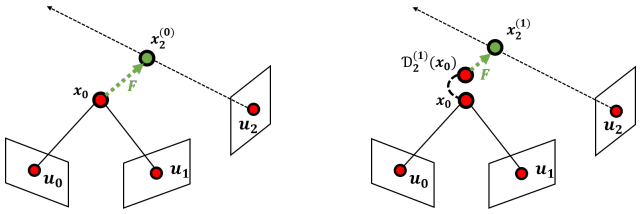
where the weights $\hat{w}_i$ are given by

$$\hat{w}_i(\hat{\mathbf{v}}) = \frac{1}{\sum_{j=1}^{k} \hat{w}_j(\hat{\mathbf{v}})} \left(1 - \frac{\|\hat{\mathbf{v}} - (\mathbf{g}_i + \mathbf{t}_i)\|_2}{\|\hat{\mathbf{v}} - (\mathbf{g}_{k+1} + \mathbf{t}_{k+1})\|_2}\right)^2.$$

Note that our approximation can be computed directly and efficiently, without leading to any error of observable influence in our synthetic experiments.

3.2. Non-rigid Photometric Consistency and Joint Optimization

With the deformation model in hand, we next estimate the depth of the other views by estimating deformations that



(a) Iteration 1 (b) Iteration 2

Figure 4: **Sparse correspondence association**: In iteration i, we transform the 3D point $\mathbf{x}_0$ according to the previous estimate of the deformation $\mathbf{D}_2^{(i-1)}$ and project $\mathbf{D}_2^{(i-1)}(\mathbf{x}_0)$ onto the ray defined by $\mathbf{u}_2$. The projection is used to define a force F pulling the point towards the ray.

are photometrically consistent with the collection images and subject to constraints on the geometry. This entire step can be interpreted as a non-rigid version of a multi-view stereo framework.

Canonical View Selection From the set of input images, we select two views with a minimal amount of deformation. We run COLMAP's implementation of PatchMatch [48] to acquire an initial temple model of the canonical pose. Based on this template, we compute the deformation graph by distributing a user-specified number of nodes on the point cloud. To this end, we start with all points of the point cloud as initial nodes. We iterate over all nodes, and for each node remove all its neighbors within a given radius. The process is repeated with a radius that is increased by 10%, until we have reached the desired number of nodes. In our experiments, we found that 100 to 200 nodes are sufficient to faithfully reconstruct the motion. Fig. 3(left) shows an example of the node distribution.

Correspondence Association For sparse global correspondences, we detect SIFT keypoints [37] in each image and match descriptors for every pair of images to compute a set of feature tracks $\{\mathbf{u}_i\}$. A *feature track* represents the same 3D point and is computed by connecting each keypoint with each of its matches. We reject inconsistent tracks, i.e., if there is a path from a keypoint $\mathbf{u}_i^{(j)}$ in image i to a different keypoint $\mathbf{u}_i^{(k)}$ with $j \neq k$ in the same image.

We lift keypoints $\mathbf{u}_i$ to 3D points $\mathbf{x}_i$, if there is a depth value in at least one processed view, compute its coordinates in the canonical pose $\mathbf{D}_i^{-1}(\mathbf{x}_i)$ and apply the current estimate of our deformation field $\mathbf{D}_j$ for frame j to these points. To establish a sparse 3D-3D correspondence $(\mathbf{D}_i^{-1}(\mathbf{x}_i), \mathbf{x}_j)$ between the canonical pose and the current frame j for the correspondence set S, we project $\mathbf{D}_j(\mathbf{D}_i^{-1}(\mathbf{x}_i))$ to the ray of the 2D keypoint $\mathbf{u}_j$ (see Fig. 4). To mitigate ambiguities and to constrain the problem, we also aim for dense photometric consistency across views. Thus, for each point of the template of the canonical pose, we also add a photometric consistency constraint with a mask $C_i \in \{0, 1\}$.

Deformation and Depth Estimation In our main iteration (see also Algorithm 1), we estimate the deformation $\hat{\mathbf{D}}$ between the canonical pose and the currently selected view by minimizing the joint optimization problem:

$$E = w_{\text{sparse}} E_{\text{sparse}} + w_{\text{dense}} E_{\text{dense}} + w_{\text{reg}} E_{\text{reg}} \tag{1}$$

$$E_{\text{sparse}} = \sum_{(i,j)\in S} \|\hat{\mathbf{D}}(\mathbf{x}_i) - \mathbf{x}_j\|_2^2$$

$$E_{\text{dense}} = \sum_r \sum_s \sum_i C_i \cdot (1 - \rho_{r,s}(\hat{\mathbf{D}}(\mathbf{x}_i), \hat{\mathbf{D}}(\mathbf{n}_i), \mathbf{x}_i, \mathbf{n}_i))^2$$

$$E_{\text{reg}} = \sum_{j=1}^{m} \sum_{k\in N(j)} \|\mathbf{R}_j(\mathbf{g}_k - \mathbf{g}_j) + \mathbf{g}_j + \mathbf{t}_j - (\mathbf{g}_k + \mathbf{t}_k)\|_2^2$$

To measure photometric consistency $\rho_{r,s}$ between a reference image r, i.e. the canonical pose, and a source view s, we use the bilaterally weighted adaption of normalized cross-correlation (NCC) as defined by Schoenberger et al. [48]. Throughout our pipeline, we employ COLMAP's default settings, i.e. a window of size 11×11. The regularizer E_{reg} as defined in [55] ensures a smooth deformation result. To ensure non-local convergence, we solve the problem in a coarse-to-fine manner using an image pyramid of 3 levels.

Both the sparse and dense matches are subject to outliers. In the sparse case, these outliers manifest as incorrect keypoint matches across images. For the dense part, outliers mainly occur due to occlusions, either because of the camera pose or because of the observed deformation.

To reject outliers in both cases, we reject correspondences with the highest residuals calculated from the result of the non-linear solution. We re-run the optimization until a user-specified maximum error is satisfied. This rejection is run in a 2-step process. First, we only solve for the deformation considering the sparse 3D-3D matches. Second, we fix the retained 3D-3D matches and solve the joint optimization problem, discarding only dense correspondences, resulting in a consistency map $C_i \in \{0, 1\}$.

We iterate this process (starting with the correspondence association) until we reach convergence. In our experiments, 3 to 5 iterations suffice to ensure a converged state.

To estimate the depth for the currently processed view, we then run a modified, non-rigid variant of COLMAP's PatchMatch [48]. Instead of simple homography warping, we apply the deformation to the point and its normal.

3.3. Implementation Details

In this section, we provide more details on the implementation of our framework (Fig. 2). Algorithm 1 shows the overall method, introduced in Sec. 3.1, and Sec. 3.2.

Given input RGB images, we first pre-process the input. To estimate the camera pose for the images, we use the SfM implementation of Agisoft Photoscan [1]. Our tests showed accurate results for scenes containing at least 60% static

Algorithm 1: Non-rigid multi-view stereo

Data: RGB input images $\{\mathbf{I}_k\}$
Result: Deformations $\{\mathbf{D}_k\}$, depth $\{d_k\}$

1 $P := \{1, \dots, k\}, Q := \emptyset$;
2 $\{\mathbf{C}_k\}$ = PhotoScanEstimateCameraPoses();
3 (i, j) = selectCanonicalViews();
4 $(d_i^{(0)}, \mathbf{n}_i^{(0)}, d_j^{(0)}, \mathbf{n}_j^{(0)})$ = ColmapPatchMatch($\mathbf{I}_i, \mathbf{I}_j$);
5 $\mathbf{D}_i^{(0)} = \mathbf{D}_j^{(0)}$ = initDeformationGraph($d_i^{(0)}, d_j^{(0)}$);
6 $\{\mathbf{u}_k\}$ = computeFeatureTracks();
7 $Q := Q \cup \{i, j\}$;
8 **while** $Q \neq P$ **do**
9 l = nextImage($P \setminus Q$);
10 $\{\mathbf{x}_k\}$ = liftKeyPointsTo3D($\{\mathbf{u}_k\}_{k \in Q}$) ;
11 $\{\mathbf{x}_i\} = \mathbf{D}_k^{-1}(\{\mathbf{x}_k\})$;
12 $\mathbf{D}_l^{(1)} = \mathbf{Id}$;
13 **for** $m = 1$ **to** N **do**
14 $\{\hat{\mathbf{x}}_l^{(m)}\} = \mathbf{D}_l^{(m)}(\{\mathbf{x}_i\})$;
15 $\{\mathbf{x}_l^{(m)}\}$ = projToRays($\{\hat{\mathbf{x}}_l^{(m)}\}, \{\mathbf{u}_l\}$) ;
16 $\{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_l^{(m)})\}$ = filter($\mathbf{D}_l^{(m)}, \{(\mathbf{x}_i, \mathbf{x}_l^{(m)})\}$);
17 $\mathbf{D}_l^{(m+1)}$ = solve($\mathbf{D}_l^{(m)}, \{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_l^{(m)})\}$, $\mathbf{I}_i, d_i^{(0)}, \mathbf{n}_i^{(0)}, \mathbf{I}_j, d_j^{(0)}, \mathbf{n}_j^{(0)}, \mathbf{I}_l$)
18 $\mathbf{D}_l = \mathbf{D}_l^{(m+1)}$;
19 $Q := Q \cup \{l\}$;
20 $(d_l^{(0)}, \mathbf{n}_l^{(0)})$ = NRPatchMatch($\{\mathbf{I}_k, \mathbf{D}_k\}_{k \in Q}$);
21 $\{(d_k, \mathbf{n}_k)\}_{k \in Q}$ = NRPatchMatch($\{\mathbf{I}_k, \mathbf{D}_k\}_{k \in Q}$);

background. A recent study [38] shows that 60∼90% static regions in a scene result in less than 0.02 degree RPE [54] error for standard pose estimation techniques (see more discussion in the supplementary material). Given the camera pose, we triangulate sparse SIFT matches [37], i.e., we compute the 3D position of the associated point by minimizing the reprojection error. We consider matches with a reprojection error of less than 1 pixel to be successfully reconstructed (static inliers). The ratio of static inliers to the number of total matches is a simple yet effective indication of the non-rigidity in the scene. We pick the image pair with the highest ratio to indicate the minimum amount of deformation and use these as the canonical views.

Two important aspects of our main iteration are described in more detail: We filter sparse correspondences (line 16 in Algorithm 1) by estimating the deformation without dense matches and rejecting the pairs with the largest residuals. This process is iterated, until the maximum residual is below a user defined threshold $d_{\max}$. The hierarchical optimization algorithm (line 17 in Algorithm 1) including filtering for dense correspondences is given in Algorithm 2.

The joint optimization in our framework is a computationally expensive task. The deformation estimation, which strongly dominates the overall run-time, is CPU intensive, while the depth computation runs on the GPU. Specifically, for the face example shown in Fig. 1 (6 images with 100 deformation nodes) the computation time needed is approximately six hours (Intel i7-6700 3.4 GHz, NVIDIA GTX 980Ti). More details on the computational expense are provided in the supplementary material.

Algorithm 2: Solving the joint problem

1 tbh **Data:** Threshold $\rho_{\max}$, Ratio $\tau \in (0, 1)$
2 **Function** `solve` $(\mathbf{D}_l, \{(\mathbf{x}_i, \mathbf{x}_l)\}, \mathbf{I}_i, d_i, \mathbf{n}_i, \mathbf{I}_j, d_j, \mathbf{n}_j, \mathbf{I}_l)$ **:**
3 $\hat{\mathbf{D}}_l = \mathbf{D}_l$;
4 **for** $m = 1$ **to** *levels* **do**
5 $\rho_{\text{cut}} := \tau \cdot (1 - \text{NCC}_{\min}) = \tau \cdot 2$;
6 $C_p := 1 \quad \forall p$;
7 **while** *true* **do**
8 $\mathbf{D}_l^*$ = solveEq1($\hat{\mathbf{D}}_l$) ;
9 $\{r_p\} = \{C_p \cdot (1 - \rho(\mathbf{D}_l^*(\mathbf{x}_p), \mathbf{D}_l^*(\mathbf{n}_p), \mathbf{x}_p, \mathbf{n}_p))\}$;
10 $e_{\max} = \max\{r_p\}$;
11 **if** $e_{max} < \rho_{max}$ **then**
12 $\hat{\mathbf{D}}_l := \mathbf{D}_l^*$;
13 break;
14 **if** $m = levels$ **then**
15 $\hat{\mathbf{D}}_l := \mathbf{D}_l^*$;
16 $C_p := 0 \quad \forall p : r_p > \rho_{\text{cut}}$;
17 $\rho_{\text{cut}} := \max\{\rho_{\max}, \tau \cdot \rho_{\text{cut}}\}$
18 **return** $\hat{\mathbf{D}}_l$

4. Evaluation

For existing non-rigid structure from motion methods, different types of datasets are used to evaluate sparse points [25, 9] and dense video frames with small baseline (including actual camera view variation) [3]. Since our problem formulation is intended for dense reconstruction of scenes with sufficient variation in both camera view and deformation, there are only few examples applicable to our scenario [36, 60, 24]. Unfortunately, these datasets are either commercial and not publicly available [36], or only exhibit rigid changes [60]. Only few depth-based approaches share input RGB data [24], but the quality of the images is not sufficient for our method (i.e., severe motion blur, low resolution (VGA) that does not provide sufficient detail to capture non-rigid changes). To quantitatively evaluate how our method can accurately capture a plausible deformation and reconstruct each scene undergoing non-rigid changes, we rendered several synthetic scenes with non-rigid deformations as shown in the first row of Fig. 5. We also captured several real-world scenes containing deforming surfaces from different views at different times. The examples (face, rubber globe, cloth and paper) in Fig. 6, and their other viewpoints are shown in the supplementary material.

4.1. Quantitative Evaluation with Synthetic Data

First, we evaluate the actual depth errors of the reconstructed depth of each time frame (i.e., propagated/refined to a specific frame), and of the final refined depth of the *canonical view*. Since we propose a challenging new problem, reconstructing non-rigid dynamic scenes from a small

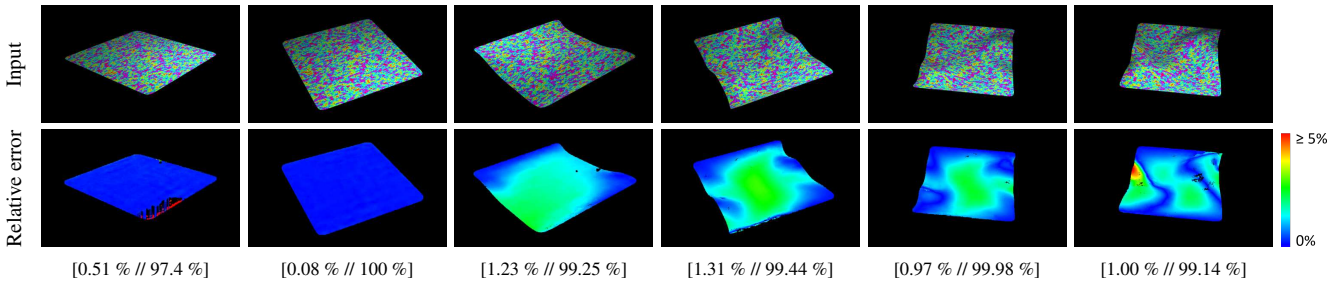


Figure 5: **Quantitative evaluation with synthetic data:** We created images of a deforming surface with 10 different views. For the evaluation, we randomly chose six examples from the set and reconstructed the point cloud. The first row shows the input images. The first two columns show the chosen canonical views. In the second row, we visualize the relative depth error compared to the ground truth. We also show the mean relative depth error value (%) and the completeness (%). The overall quantitative evaluation including a comparison to other baselines are shown in Table 1.

Table 1: **Evaluation with other baselines:** (a) COLMAP [48], (b) non-rigid ICP (NRICP) [33] with a dense photometric optimization, and (c) Our method with different settings. S denotes *sparse*, D denotes *dense*, photometric objective. N equals the number of iterations for sparse correspondence association (see Sec. 3.2). We compute the mean relative error (MRE) for all reconstructed and non-rejected depth values as well as the overall completeness. Note that we create additional synthetic data for multiple views per each deformation only for NRICP, which requires triangulated points per each deformation, as an upper bound (using ground truth depth images), whereas COLMAP and our methods *only use a single view per deformation*.

			(c) Our methods				
	(a) COLMAP	(b) NRICP	S ($N = 1$)	S ($N = 10$)	D	S ($N = 1$) + D	S ($N = 10$) + D
MRE	2.11 %	0.53 %	1.48 %	1.50 %	2.37 %	1.12 %	**1.11** %
Completeness	68.74 %	99.30 %	97.24 %	97.71 %	96.41 %	98.76 %	**98.99** %

set of images, it is not easy to find other baseline methods. Thus, we conduct our evaluation with an existing MVS method, COLMAP [48], as a lower bound, and use as an upper bound a non-rigid ICP method similar to Li et al. [33] based on the (in practice unknown) ground truth depth. The non-rigid ICP using the point-to-plane error metric serves as a geometric initialization. We refine the deformation using our dense photometric alignment (see Sec. 3.2).

To compare the influence of our proposed objectives for deformation estimation, i.e. sparse 3D-3D correspondences and dense non-rigid photometric consistency, we evaluate our algorithm with different settings. The relative performance of these variants can be seen as an ablation study. We perform evaluation on the following variants: 1) considering only the sparse correspondence association using different numbers of iterations (see Sec. 3.2), 2) considering only the dense photometric alignment, and 3) the combination of sparse and dense. The results of the quantitative evaluation can be found in Table 1. All methods/variants obtain a mean relative error $< 2.4\%$, overall resulting in a faithful reconstruction. Our joint optimization algorithm considerably improves the reconstruction result compared to COLMAP both in terms of accuracy (MRE reduction from 2.11% to 1.11%) and completeness (from 69% to 99%).

Importance of sparse and dense matching terms. In terms of overall accuracy, the quantitative evaluation shows that the sparse term is more important than the dense term (S vs. D). Sparse feature matches serve as global anchor points to guide the deformation, while the dense term provides local refinement ($S(N = 10)$ vs. $S(N = 10) + D$).

4.2. Qualitative Evaluation with Real Data

Fig. 6 shows results of our non-rigid 3D reconstruction. For each pair of rows, we show six input images and the corresponding deformed 3D point clouds. Note that the *deformed* point clouds belong to the collection of 3D reconstructed points propagated by the computed deformations using the other views as described in Sec. 3.2. The point cloud of each first column of Fig. 6 shows the first canonical point cloud (triangulated points from two views with minimal deformation). We visualize each reconstructed scene from similar viewpoints as the input views. More viewpoints of the results including visualizations of the deformation graphs can be found in the supplementary material.

4.3. Quantitative Evaluation with Real Data

To quantitatively evaluate our approach on real-world data, we generated different deformations of a rubber globe. For each deformation state, we took 16 pictures that allowed us to generate high-quality reconstructions [1] of that state,

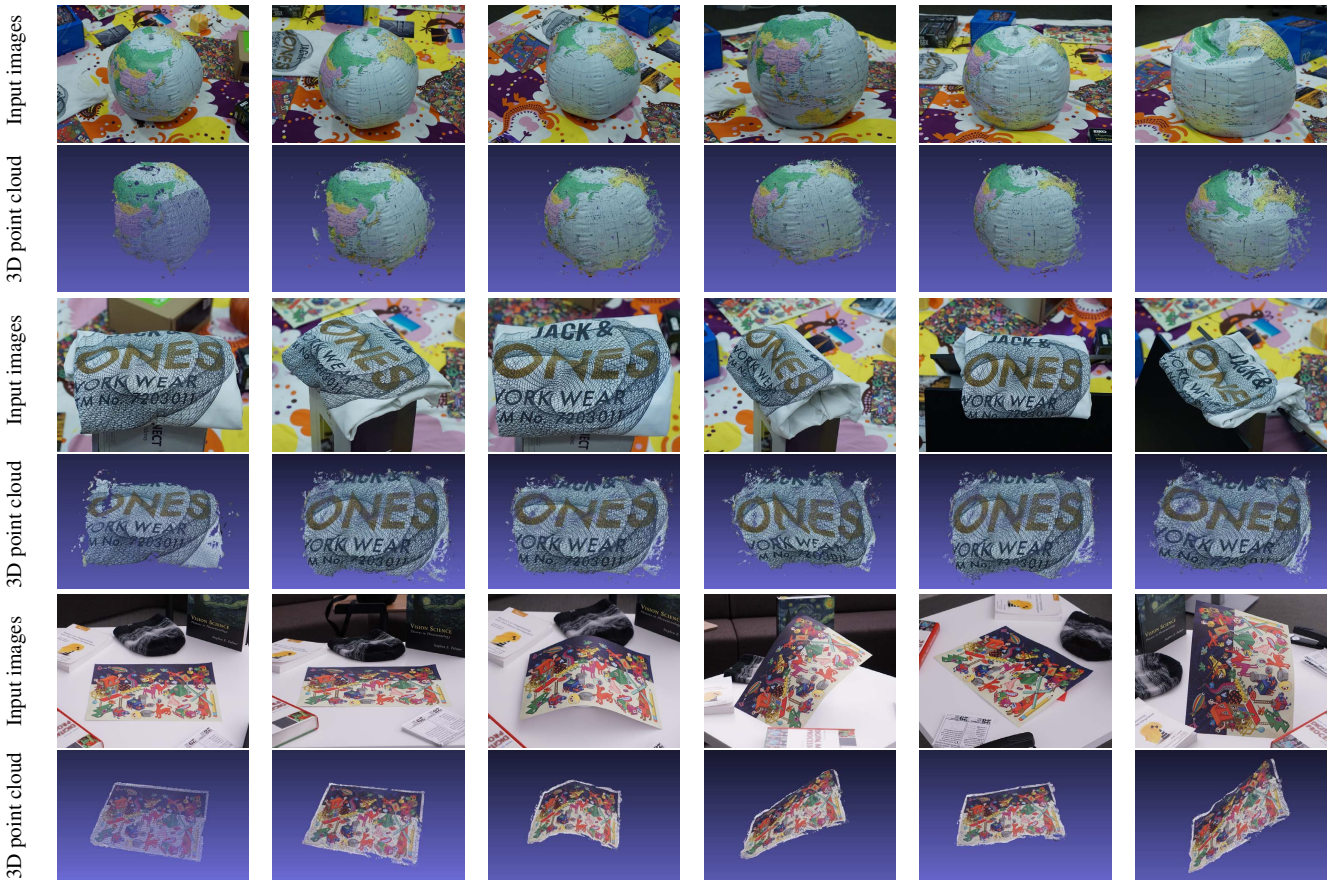


Figure 6: **Qualitative evaluation with real data**: In each row, the first two columns show the views used to create each canonical point cloud. The first column of each result row (even row) shows the original canonical point cloud. The remaining views from the second column of each result row shows the propagated version of reconstructed point clouds for each view.

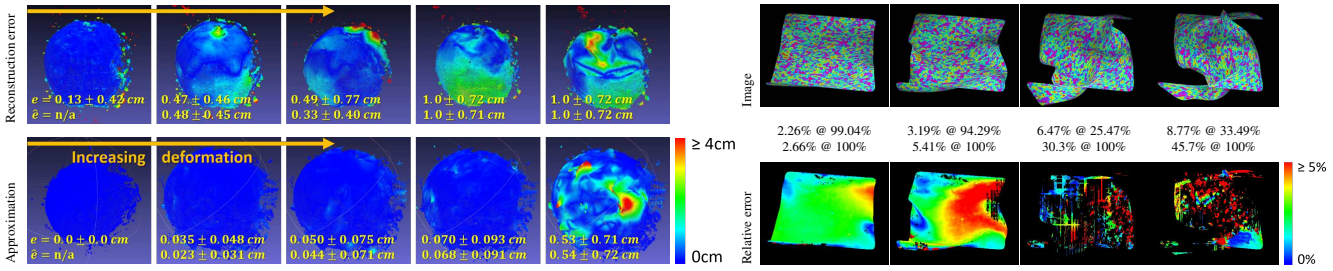


Figure 7: **Left: Quantitative evaluation with real data**: Reconstruction error (upper) and approximation error of the inverse deformation for the globe example (data as in Fig. 6). The average errors $\pm\sigma$ are given for the complete point cloud (e) and parts with deformations $\geq$ 4cm ($\hat{e}$). **Right: Stress test with extreme deformations**: *Relative depth error @ completeness*

which we consider as ground truth. We then apply our new approach using only one single image of each deformation state, with all images from different views. Fig. 7 (upper left) shows the distance to the ground truth, with an overall average error of 0.6cm (1.6% of the globe's diameter).

4.4. Inverse Deformation Approximation

We evaluate the error of our proposed approximation of the inverse deformation $\mathbf{D}^{-1}$ by $\|\mathbf{D}^{-1}(\mathbf{D}(\mathbf{p})) - \mathbf{p}\|_2$ for all points $\mathbf{p}$ of the reconstructed point cloud. Fig. 7 (lower left) shows an overall average error of 1.7mm. The canonical pose has an error of 0. With increasing amount of deformation, the accuracy slightly degrades due to the approximate nature of the weight computation.

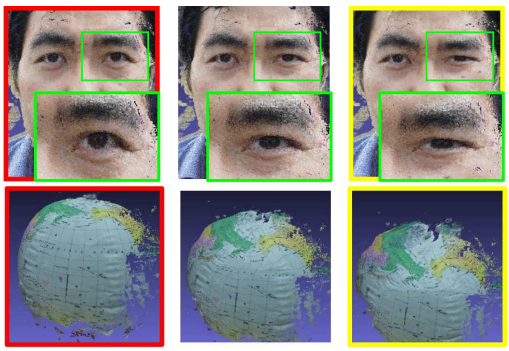


Figure 8: **Deformation interpolation with in-between views:** We interpolate a point cloud between two reconstructed views from their depth and deformation. We show two key-frames, source and target, denoted as red and yellow frames respectively, and then demonstrate the interpolated *intermediate point cloud* in the middle column. For the top row, zoomed in-set images of the eye region show how the deformation is applied to the intermediate point cloud. More examples of the interpolated 4D animations are shown in the supplementary materials.

4.5. Dynamic Scene Interpolation

Since we estimate deformations between each view and the canonical point cloud, once all deformation pairs have been created, we can easily interpolate the non-rigid structure. To blend between the deformations, we compute interpolated deformation graphs by blending the rigid body transform at each node using dual-quaternions [27].

In Fig. 8, we show interpolated results from reconstructed scenes of the face example and the globe example shown in Fig. 6. Note that even though the estimated deformation is defined between each view and the canonical pose, any combination of deformation interpolation is possible. More examples of the interpolated structures are shown in the supplementary material.

4.6. Limitations and Failure Cases

It is also important to point out the limitations of our approach (Fig. 9). We first assume that there is at least one pair of images that has minimal deformation for the initial canonical model. This can be interpreted as the first step used by many SLAM or 3D reconstruction algorithms for the initial triangulation. Fig. 9(a) shows an example of a canonical point cloud created from two views that contain too much deformation, only leading to a partial triangulation. Fig. 9(b) shows an example where the deformation occurs mostly along the view direction. While we successfully estimate the deformation and reconstruct a similar example shown in Fig. 6, depending on the view this can cause an erroneous estimation of the deformation.

To do a quantitative evaluation of failure cases, we stress-test our approach on a challenging synthetic sequence (Fig. 7 right). Despite large deformations, ours generates good results, except for extremely twisted geometry. We also evaluate the influence of the input resolution (scales 1, 0.5, 0.25, 0.125) for the rightmost instance of the globe example in Fig. 6. The avg. errors are 0.99cm, 0.91cm, 0.91cm, and 3.22cm, respectively. Up to a down scaling by a factor of 4, the reconstruction error remains stable at $\approx$ 1.0cm. With lower resolution images (scale ≥ 8), the algorithm is no longer able to accurately track the deformation.

Currently, our method assumes that there is only one deforming object in the scene. Extending it by an instance segmentation will allow to handle multiple objects.

(a) Bad canonical view selection (b) Ambiguity along view direction

Figure 9: **Failure cases:** (a) shows the result of canonical point cloud reconstruction from two views that are incorrectly selected (large deformation between two views: images in first and third column in top row of Fig. 6). While the camera pose is successfully computed, since there are large portions of non-rigid changes happening in the upper part of face and near the mouth, there are many holes on the face, which is not the best case if we choose this pair. (b) shows a failure case when deformation (red circles) occurs along the view direction, which causes the ambiguity.

5. Conclusion and Discussion

We propose a challenging new research problem for dense 3D reconstruction of scenes with non-rigid changes from a few images sparsely captured from different views of a single monocular camera. As a solution, we present a joint optimization technique that optimizes over depth, appearance, and the deformation field in order to model these non-rigid scene changes. We show that an MVS solution for non-rigid change is possible, and the estimated deformation field can be used to interpolate motion between views.

While our approach still shows limitations, e.g. due to incorrect canonical view selection (see 4.6), we believe that recent advances in deep learning-based approaches to estimate depth from single RGB input [16] or learning local rigidity [38] for rigid/non-rigid classification can play a key role for both the initialization and further mitigation of these ambiguities.

Acknowledgments

This research is partly funded by the Bayerische Forschungsstiftung (For3D).

References

[1] Agisoft. PhotoScan: MVS software, 2000–2004. 3, 4, 6

[2] B. Allain, J.-S. Franco, and E. Boyer. An efficient volumetric framework for shape tracking. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[3] M. D. Ansari, V. Golyanik, and D. Stricker. Scalable dense monocular surface reconstruction. *Intl. Conf. on 3D Vision*, 2017. 1, 3, 5

[4] A. Bartoli, Y. Gerard, F. Chadebecq, T. Collins, and D. Pizarro. Shape-from-template. *IEEE Trans. Pattern Anal. Machine Intell.*, 37(10):2099–2118, 2015. 3

[5] M. Bleyer, C. Rhemann, and C. Rother. Patchmatch stereo - stereo matching with slanted support windows. In *British Machine Vision Conf. (BMVC)*, pages 14.1–14.11, 2011. 1, 2

[6] M. Bojsen-Hansen, H. Li, and C. Wojtan. Tracking surfaces with evolving topology. *ACM Trans. on Graphics (TOG)*, 31(4):53, 2012. 2

[7] N. D. F. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conf. on Computer Vision (ECCV)*, pages 766–779, 2008. 1

[8] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. on Graphics (TOG)*, 34(4):69, 2015. 3

[9] Y. Dai, H. Deng, and M. He. Dense non-rigid structure-from-motion made easy - A spatial-temporal smoothness based solution. *CoRR*, abs/1706.08629, 2017. 1, 3, 5

[10] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, 107(2):101–122, 2014. 1

[11] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. *ACM Trans. on Graphics (SIGGRAPH)*, 27:1–10, 2008. 2

[12] M. Dou, H. Fuchs, and J.-M. Frahm. Scanning and tracking dynamic objects with commodity depth cameras. In *IEEE and ACM Intl. Sym. on Mixed and Augmented Reality (ISMAR)*, pages 99–106, 2013. 2

[13] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Trans. on Graphics (TOG)*, 35(4):114, 2016. 2

[14] M. Dou, J. Taylor, H. Fuchs, A. Fitzgibbon, and S. Izadi. 3D scanning deformable objects with a single RGBD sensor. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[15] O. Enqvist, F. Kahl, and C. Olsson. Non-sequential structure from motion. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 264–271. IEEE, 2011. 3

[16] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8

[17] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007. 1, 2

[18] J. Gall, B. Rosenhahn, and H. P. Seidel. Drift-free tracking of rigid and articulated objects. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008. 2

[19] S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Intl. Conf. on Computer Vision (ICCV)*, 2015. 1, 2

[20] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 3

[21] K. Guo, F. Xu, Y. Wang, Y. Liu, and Q. Dai. Robust non-rigid motion tracking and surface reconstruction using L0 regularization. *Intl. Conf. on Computer Vision (ICCV)*, 2015. 2

[22] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004. 1, 2

[23] C. Hernández, G. Vogiatzis, G. J. Brostow, B. Stenger, and R. Cipolla. Non-rigid photometric stereo with colored lights. In *Intl. Conf. on Computer Vision (ICCV)*, 2007. 2

[24] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger. VolumeDeform: Real-time volumetric non-rigid reconstruction. In *European Conf. on Computer Vision (ECCV)*, 2016. 1, 2, 5

[25] S. H. N. Jensen, A. Del Bue, M. E. B. Doest, and H. Aanæs. A benchmark and evaluation of non-rigid structure from motion. *Arxiv*, abs/1801.08388, 2018. 1, 2, 5

[26] T. Kanade and D. D. Morris. Factorization methods for structure from motion. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 356(1740):1153–1173, 1998. 2

[27] L. Kavan, S. Collins, J. Žára, and C. O'Sullivan. Skinning with dual quaternions. In *Proceedings of the 2007 symposium on Interactive 3D graphics and games*, pages 39–46. ACM, 2007. 2, 8

[28] S. Kumar, Y. Dai, and H. Li. Monocular dense 3D reconstruction of a complex dynamic scene from two perspective frames. *Intl. Conf. on Computer Vision (ICCV)*, 2017. 2

[29] A. Kundu, K. M. Krishna, and C. V. Jawahar. Real-time multibody visual SLAM with a smoothly moving monocular camera. In *Intl. Conf. on Computer Vision (ICCV)*, 2011. 1

[30] E. S. Larsen, P. Mordohai, M. Pollefeys, and H. Fuchs. Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In *Intl. Conf. on Computer Vision (ICCV)*, pages 1–8. IEEE, 2007. 3

[31] V. Leroy, J.-S. Franco, and E. Boyer. Multi-view dynamic shape refinement using local temporal integration. In *Intl. Conf. on Computer Vision (ICCV)*, pages 3094–3103, 2017. 3

[32] H. Li, B. Adams, L. J. Guibas, and M. Pauly. Robust single-view geometry and motion reconstruction. *ACM Trans. on Graphics (TOG)*, 28(5):175, 2009. 2

[33] H. Li, B. Adams, L. J. Guibas, and M. Pauly. Robust single-view geometry and motion reconstruction. In *ACM Trans. on Graphics (SIGGRAPH Asia)*, pages 175:1–175:10, 2009. 6

[34] H. Li, L. Luo, D. Vlasic, P. Peers, J. Popović, M. Pauly, and S. Rusinkiewicz. Temporally coherent completion of dynamic shapes. *ACM Trans. on Graphics (TOG)*, 31(1):2, 2012. 2

[35] H. Li, R. W. Sumner, and M. Pauly. Global correspondence optimization for non-rigid registration of depth scans. In *Computer Graphics Forum*, volume 27, pages 1421–1430, 2008. 2

[36] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev. 3d self-portraits. *ACM Trans. on Graphics (TOG)*, 32(6):187:1–187:9, Nov. 2013. 2, 5

[37] D. G. Lowe. Object recognition from local scale-invariant features. In *Intl. Conf. on Computer Vision (ICCV)*, 1999. 4, 5

[38] Z. Lv, K. Kim, A. Troccoli, D. Sun, J. Rehg, and J. Kautz. Learning rigidity in dynamic scenes with a moving camera for 3d motion field estimation. In *European Conf. on Computer Vision (ECCV)*, 2018. 5, 8

[39] M. Magnor and B. Goldlucke. Spacetime-coherent geometry reconstruction from multiple video streams. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, pages 365–372. IEEE, 2004. 3

[40] C. Malleson, M. Klaudiny, J. Y. Guillemaut, and A. Hilton. Structured representation of non-rigid surfaces from single view 3D point tracks. In *Intl. Conf. on 3D Vision*, volume 1, pages 625–632, 2014. 2

[41] N. J. Mitra, S. Flöry, M. Ovsjanikov, N. Gelfand, L. J. Guibas, and H. Pottmann. Dynamic geometry registration. In *Symposium on Geometry Processing (SGP)*, pages 173–182, 2007. 2

[42] J. Neumann and Y. Aloimonos. Spatio-temporal stereo using multi-resolution subdivision surfaces. *Intl. J. of Computer Vision*, 47(1):181–193, 2002. 3

[43] R. A. Newcombe, D. Fox, and S. M. Seitz. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2

[44] D. Nistér. Preemptive ransac for live structure and motion estimation. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, pages 199–, Washington, DC, USA, 2003. IEEE Computer Society. 1

[45] G. Palma, T. Boubekeur, F. Ganovelli, and P. Cignoni. Scalable non-rigid registration for multi-view stereo data. *ISPRS journal of photogrammetry and remote sensing*, 142:328–341, 2018. 2

[46] V. Rabaud and S. Belongie. Re-thinking non-rigid structure from motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008. 2

[47] J. L. Schönberger and J.-M. Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 3

[48] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 2, 4, 6

[49] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[50] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006. 1, 2

[51] M. Slavcheva, M. Baust, D. Cremers, and S. Ilic. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[52] M. Slavcheva, M. Baust, and S. Ilic. Sobolevfusion: 3d reconstruction of scenes undergoing free non-rigid motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[53] O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry Processing (SGP)*, volume 4, pages 109–116, 2007. 2, 3

[54] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Oct. 2012. 5

[55] R. W. Sumner, J. Schmid, and M. Pauly. Embedded deformation for shape manipulation. *ACM Trans. on Graphics (TOG)*, 26(3):80, 2007. 2, 3, 4

[56] A. Tevs, A. Berner, M. Wand, I. Ihrke, M. Bokeloh, J. Kerber, and H.-P. Seidel. Animation cartography—intrinsic reconstruction of shape and motion. *ACM Trans. on Graphics (TOG)*, 31(2):12, 2012. 2

[57] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *Intl. J. of Computer Vision*, 9(2):137–154, 1992. 2

[58] B. Triggs. Factorization methods for projective structure and motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1996. 2

[59] R. Wang, L. Wei, E. Vouga, Q. Huang, D. Ceylan, G. Medioni, and H. Li. Capturing dynamic textured surfaces of moving targets. In *European Conf. on Computer Vision (ECCV)*, 2016. 2

[60] T. Y. Wang, P. Kohli, and N. J. Mitra. Dynamic SfM: Detecting Scene Changes from Image Pairs. *Computer Graphics Forum*, 2015. 1, 5

[61] M. Zeng, J. Zheng, X. Cheng, and X. Liu. Templateless quasi-rigid shape modeling with implicit loop-closure. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[62] G. Zhang, J. Jia, and H. Bao. Simultaneous multi-body stereo and segmentation. In *Intl. Conf. on Computer Vision (ICCV)*, 2011. 1

[63] M. Zollhöfer, M. Nießner, S. Izadi, C. Rehmann, 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 Trans. on Graphics (TOG)*, 33(4), 2014. 2