Intrinsic Dynamic Shape Prior for Dense Non-Rigid Structure from Motion*

Vladislav Golyanik[1] André Jonas[2] Didier Stricker[2,3] Christian Theobalt[1]

[1]MPI for Informatics, SIC [2]University of Kaiserslautern [3]DFKI

Abstract

While dense non-rigid structure from motion (NRSfM) has been extensively studied from the perspective of the reconstructability problem over the recent years, almost no attempts have been undertaken to bring it into the practical realm. The reasons for the slow dissemination are the severe ill-posedness, high sensitivity to motion and deformation cues, and the difficulty to obtain reliable point tracks in the vast majority of practical scenarios.

To fill this gap, we propose a new framework that first extracts prior knowledge from an input image sequence with NRSfM. Our Dynamic Shape Prior Reconstruction (DSPR) approach then uses the obtained 3D reconstructions as a dynamic shape prior for sequential surface recovery in scenarios with recurrence. DSPR can be combined with existing dense NRSfM techniques while its energy functional is optimised with multi-start gradient descent at real-time frame rates for new incoming point tracks. The proposed versatile framework with a new core NRSfM approach outperforms several other methods in the ability to handle inaccurate and noisy point tracks, provided we have access to a representative (in terms of the deformation variety) image sequence. Comprehensive experiments highlight convergence properties and the accuracy of DSPR under different disturbing effects. We also perform a joint study of tracking and reconstruction and show applications to shape compression and heart reconstruction under occlusions. We achieve state-of-the-art metrics (accuracy and compression ratios) in different scenarios.

1. Introduction

Dynamic non-rigid 3D reconstruction from monocular image sequences relying exclusively on motion and deformation cues and weak prior assumptions is known as *non-rigid structure from motion* (NRSfM) [9, 8, 53]. Despite advances over recent years in the reconstruction accuracy and variety of scenarios which can be handled by NRSfM [41, 23, 15, 18, 30], there is a gap between results achieved in a controlled environment and real scenarios. Often, it is difficult to obtain reliable dense correspondences across input views. Due to the high ill-posedness of NRSfM, there is no universal set of prior constraints that works equally well across different scenarios.

The **main contribution** of this paper is a new fast and sequential technique for dense monocular non-rigid reconstruction with an *intrinsic* dynamic shape prior (DSP), *i.e.,* a sequence-specific set of ordered and gradually changing 3D states obtained on a representative image sequence (Sec. 3). In the vast majority of real-world cases, not deformations but rather different angles of view (camera poses) cause different 2D measurements. It is assumed that the representative sequence provides a sufficient variety of deformations as they are likely to occur in a given scene, whereas there are no strong requirements for poses except that those must be nondegenerate. While the DSP generation is offline, the reconstruction of new frames with DSP is light-weight and well parallelisable. It implicitly assumes temporally-disjoint rigidity, *i.e.,* the situation when a newly observed 3D state is reoccurring with respect to the DSP.

For every new incoming measurement, the proposed shape-from-DSP or *Dynamic Shape Prior Reconstruction* (DSPR) approach finds a globally optimal 3D state corresponding to the 2D measurements and rigidly transforms it to the pose as observed in the measurements by alternating between multi-start gradient descent (MSGD) and camera pose estimation. Note that the pose in the incoming frames can be arbitrary and differ significantly from poses observed during the DSP generation in the generative sequence, due to the decoupling property of shapes and poses in NRSfM. Thus, our framework can be considered as a variant of incremental NRSfM, since we decouple the basis estimation from the weights and camera poses. See Fig. 6 for an example of monocular non-rigid reconstruction with DSPR.

As a **further contribution**, we propose a new light-weight dense per-point extension of [19] which we call *Dense Consolidating Monocular Dynamic Reconstruction* (D-CMDR) approach for the DSP recovery from a representative sequence, even though any accurate existing dense

*This work is supported by the ERC Consolidator Grant 4DReply (770784) and the BMBF projects DYNAMICS (01IW15003) and VIDETE (01IW18002).

NRSfM method can be employed for this task (Sec. 3.1). Thus, *the focus of this paper is towards making NRSfM applicable in real-world scenarios*, and not improving the accuracy of NRSfM from the perspective of the reconstructability problem *per se*. Apart from real-time monocular reconstruction from noisy data, our main idea can also be applied to several related problems. Since DSP represents a compact footprint of the geometry carrying a learned sequence-specific deformation model, it suggests the suitability of DSPR for **geometry compression** (Sec. 4.6).

We thoroughly evaluate our DSPR framework and the D-CMDR approach for the DSP generation (Sec. 4). Apart from the standard NRSfM datasets and the NRSfM challenge [24] covering more than fifteen methods (Sec. 4.2), we synthesise a new *actor mocap* dataset for joint evaluation of dense point tracking and reconstruction (Sec. 4.5). Moreover, compared to the prevalent evaluation policy of dense NRSfM in the literature, we evaluate our framework with perturbed point tracks and missing data (Sec. 4.3).

2. Related Work

Some recent works on NRSfM focus on dense [15] and scalable methods [5, 29, 28] as well as approaches for complex non-linear deformations [56, 26]. A distinct tendency is investigating new, often simple and, at the same time, overlooked ideas [12, 33, 27] and models for NRSfM [50, 2, 21]. More attention is paid to hybrid approaches which make stronger assumptions than classical NRSfM but fewer assumptions than template-based counterparts [38, 55, 25] or domain-specific approaches which expect a known object class [7, 51, 40]. A hybrid NRSfM technique with occlusion handling [18] obtains a static shape prior on several unoccluded frames of a sequence. Some methods with a trained deformation model rely on a representative training dataset [39, 44]. Our algorithm has thrived on the ideas proposed in the works mentioned above. The most closely related methods to DSPR are [33] and [18].

The Method of Li *et al.* [33]. Li *et al.* [33] propose to exploit state recurrency in sparse NRSfM. While a local rigidity method rapidly reaches its lower bound on the number of views necessary for the rigid reconstruction to produce meaningful results [42], the method of Li *et al.* [33] does not rely on connected temporal windows and is agnostic to the deformation intensity over a short period. The number of rigid clusters has to be set in [33] in advance. Besides, if some states are unique or degenerate (are not observed in other poses), they are assigned to some non-empty clusters and treated as noise. Thus, non-reoccurring states are reconstructed less accurately. Moreover, the method of Li *et al.* [33] requires computationally costly graph clustering and works for a few sparse points. In contrast, DSPR fits an instance from DSP which is related to a given dense 2D measurement by rigidity. We do not explicitly cluster dense point tracks into bins relating the underlying 3D states by rigidity. Instead, we find a subsequence providing as diverse deformations as possible in as few views as possible.

Shape Priors and Degenerate Data Handling. Del Bue [10] proposed an NRSfM factorisation with a supportive pre-computed shape basis. The method was shown to handle degeneracies in the sparse point tracks robustly. Golyanik *et al.* [18] included a static shape prior into dense variational NRSfM. Compared to them, we extract multiple states from a representative sequence reflecting the entire deformation model. While the aim of [18] is the stability under large occlusions, their method also tends to overconstrain the reconstructions. Our primary goal is a lightweight sequential scheme with recurrent state identification, and still, it is remarkably robust under occlusions.

Several methods for sparse NRSfM address missing data [35, 22, 23, 2, 32]. Gotardo and Martinez [22, 23] rely on a pre-defined trajectory basis and the smooth deformations constraint while recovering a low-rank approximation of the measurement matrix with estimated missing entries. The approach of Lee *et al.* [32] is robust to moderate portions of missing data as the shape likelihoods are influenced only by available entries in their method. Our DSPR approach is robust to moderate portions of missing entries. Note that we treat those as erroneous measurements, which is a more realistic assumption in the dense setting.

Sequential NRSfM. The majority of NRSfM methods operate globally on frame batches [37, 15, 56, 2, 26, 5, 29]. Paladini *et al.* [36] proposed the seminal sequential method which incrementally updates deformation modes upon the data availability. Agudo and coworkers [1] introduced a probabilistic model with physics-based constraints for dense sequential NRSfM. Once DSP is obtained, our DSPR switches to the sequential reconstruction and requires only a single measurement and the latest regressed surface as an input. In contrast to [36, 1], it is explicitly designed with the handling of inaccurate correspondences in mind. Moreover, our optimisation is very fast and highly parallelisable. It is possibly faster than most of the NRSfM algorithms in the literature so far, considering the methods [1, 20, 3].

Recovery of the Dynamic Shape Prior. In the proposed D-CMDR for the reconstruction of a representative sequence, up to several millions of parameters are optimised with non-linear least squares (NLLS). D-CMDR is tailored for the dense per-point case and is a variant of the segment-wise CMDR [19]. The most closely related approach to D-CMDR is the template-based method of Yu *et al.* [55], with several differences: 1) instead of using a multiview reconstruction to obtain a template, we initialise shapes and camera poses with the rigid factorisation [52]; 2) we use trajectory regularisation instead of as-rigid-as-possible regulariser [45], and 3) the fitting term operates on point tracks and not directly on images. An NRSfM technique with simultaneous constraints in metric and trajectory spaces is Column Space Fitting [23]. Our trajectory smoothness term was rarely used in energy-based NRSfM so far. It allows integration of subspace constraints on point trajectories and originates from [4]. We demand smoothness of neighbouring trajectories by optimising the total variation of trajectory coefficients. A similar regulariser was previously applied in multi-frame optical flow (MFOF) [16, 49]. Olsen and Bartoli [35] proposed one of the first spatial regularisers with a related principle, *i.e.,* a surface continuity prior term imposing similarity constraint on neighbouring point trajectories for the enhanced robustness against missing data.

3. The Proposed DSPR Approach

Our objective is the 3D reconstruction of a current 3D state $\mathbf{S}_f \in \mathbb{R}^{3\times N}$ given incoming measurements $\mathbf{W}_f \in \mathbb{R}^{2\times N}$, $f \in \{1,\dots,F\}$ and a DSP $\mathbf{D} = \{\mathbf{D}_i\}$, $i \in \{1,\dots,Q\}$ with Q temporal rigidity bases. F is the total number of frames and N is the number of points per frame. We formulate dense sequential NRSfM as a per-frame energy minimisation problem of finding $\mathbf{D}_i$ related to $\mathbf{S}_f$ by a rigid transformation and camera pose $\mathbf{R}_f \in \mathbb{R}^{3\times 3}$ ($\mathbf{R}_f^\mathsf{T} = \mathbf{R}_f^{-1}$, $\det(\mathbf{R}_f) = 1$) so that the product $\mathbf{R}_f\mathbf{D}_i$ explains the current observation $\mathbf{W}_f$:

$$\begin{aligned}\mathbf{E}(\mathbf{S}_f = \mathbf{D}_i, \mathbf{R}_f) = \alpha\,\|\mathbf{W}_f - \mathbf{I}_{2\times 3}\mathbf{R}_f\,\mathbf{D}_{i:\lambda_i=1}\|_\mathcal{F} + \\ \beta\,\|\mathbf{D}_{i:\lambda_i=1} - \mathbf{S}_{f-1}\|_\mathcal{F} + \gamma\,(\|\boldsymbol{\lambda}\|_0 - 1)^2,\end{aligned} \tag{1}$$

where $\|\cdot\|_0$ and $\|\cdot\|_\mathcal{F}$ denote a zero-norm of a vector and Frobenius norm, respectively, $\mathbf{I}_{2\times 3}$ models orthographic projection and $\boldsymbol{\lambda} = [\lambda_i]$ is the indicator function for DSP. The energy functional (1) contains a data term weighted by α, temporal smoothness term weighted by β and a DSP regularisation term weighted by γ. The data term ensures that the factorisation $\mathbf{R}_f\mathbf{S}_f$ is accurately projected to $\mathbf{W}_f$. The smoothness term expresses the assumption of the gradual character of changes in the states as well as helps to converge faster. The regulariser ensures that a single $\mathbf{D}_i$ is required to explain observations upon our model. This practice contrasts to some other methods, where every shape is encoded as a linear combination of basis shapes (recovered during the reconstruction or known in advance) [37]. In our model, DSP is assumed to provide a sufficient variety to cover the entire space of reoccurring deformations, and we use the decoupling property of the shape and pose.

The energy functional (1) is minimised iteratively, by alternatingly fixing $\mathbf{R}_f$ and releasing $\mathbf{D}_{i:\lambda_i=1} = \mathbf{S}_f$, and vice versa, in every iteration. When $\mathbf{S}_f$ is fixed, the only term dependent on $\mathbf{R}_f$ is the data term. For the sake of low computational latency per frame, $\mathbf{R}_f$ can be updated in the closed form by projecting its affine update to the $\mathrm{SO}(3)$ group or by linear least squares with quaternion parametrisation. When $\mathbf{R}_f$ is fixed, an optimal $\mathbf{D}_i$ is found by taking the partial derivative of the energy subspace with the fixed $\mathbf{R}_f$, denoted by $\mathbf{E}_{\mathbf{R}_f}$, w.r.t. $\boldsymbol{\lambda}$ and equating it to zero:

$$\frac{\partial \mathbf{E}_{\mathbf{R}_f}(\mathbf{S})}{\partial \mathbf{S}}\frac{\partial \mathbf{S}}{\partial \boldsymbol{\lambda}} = 0. \tag{2}$$

The optimality criterion (2) defines a state when a small change in the shape caused by a small change in the prior state does not change the energy. We minimise the energy functional (1) — when $\mathbf{R}_f$ is fixed — by the *multi-start gradient descent* (MSGD) method. Starting from multiple regularly sampled values of $\boldsymbol{\lambda}$, we compute differences in $\mathbf{E}_{\mathbf{R}_f}$ and update $\boldsymbol{\lambda}$ in the direction of the energy decrease. Multiple starting points are required to obtain a globally optimal solution since $\mathbf{E}$ is non-convex. The global minimum is obtained by comparing locally minimal energy values. MSGD is well parallelisable as every thread can converge or finish upon a boundary condition (*e.g.,* when leaving the assigned range of values) independently from other threads. Thanks to MSGD, DSPR executes with three-five frames per second on our hardware without parallelisation (see Sec. 4).

3.1. Obtaining Dynamic Shape Prior (DSP)

DSP generation includes an accurate 3D reconstruction of a representative image sequence with a general-purpose NRSfM method. In principle, we are free to choose any dense scalable NRSfM technique for the initial reconstruction. In the quantitative experiments, we use two accurate existing methods, *i.e.,* Garg *et al.* [15] and Ansari *et al.* [5]. Additionally, we propose a new energy-based NRSfM method which outperforms the approaches mentioned above in a subset of evaluation scenarios.

3.1.1 Our Core NRSfM Approach for DSP Acquisition

For notational consistency, we denote the measurements of the representative sequence and the corresponding 3D shapes in this section by $\mathbf{W}_{2F\times N} = [\mathbf{W}_f]$ and $\mathbf{S}_{3F\times N} = [\mathbf{S}_f]$ respectively, with N denoting the number of points in every frame. The new method minimises the following energy functional with the Gauss-Newton algorithm:

$$\begin{aligned}\mathbf{E}_{\text{D-CMDR}}(\mathbf{R}, \mathbf{S}, \mathbf{A}) = \alpha\,\mathbf{E}_{\text{fit}}(\mathbf{R}, \mathbf{S}) + \beta\,\mathbf{E}_{\text{temp}}(\mathbf{S}) + \\ + \lambda\,\mathbf{E}_{\text{linking}}(\mathbf{S}, \mathbf{A}) + \rho\,\mathbf{E}_{\text{reg}}(\mathbf{A}),\end{aligned} \tag{3}$$

where $\mathbf{A}$ is a matrix with trajectory coefficients explained below. The data term constrains projections of the recovered shapes to agree with the 2D measurements:

$$\mathbf{E}_{\text{fit}}(\mathbf{R}, \mathbf{S}) = \sum_f \|\mathbf{W}_f - \mathbf{I}_{2\times 3}\mathbf{R}_f\,\mathbf{S}_f\|_\epsilon^2\,, \tag{4}$$

where $\|\cdot\|_\epsilon$ is Huber norm ($\epsilon = 0.1$). The temporal smoothness term imposes similarity on adjacent reconstructions:

$$\mathbf{E}_{\text{temp}}(\mathbf{S}) = \sum_{f=2}^{F} \|\mathbf{S}_f - \mathbf{S}_{f-1}\|_\epsilon^2\,. \tag{5}$$

The linking term expresses our assumptions about the complexity of deformations (deformation model). Here, we rely on K known basis trajectories Θ sampled from discrete cosine transform (DCT) at regular intervals:

$$\mathbf{E}_{\text{linking}}(\mathbf{S}, \mathbf{A}) = \|\mathbf{S} - (\Theta \otimes \mathbf{I}_3)_{3F\times 3K}\ \mathbf{A}_{3K\times N}\|_\epsilon^2, \text{ where} \tag{6}$$

$$\begin{gathered}\Theta = \left([\theta_{11} \quad \dots \quad \theta_{1K}] \quad \dots \quad [\theta_{F1} \quad \dots \quad \theta_{FK}]\right)^\mathsf{T}, \\ \theta_{tk} = \frac{\sigma_k}{\sqrt{2}} \cos\left(\frac{\pi}{2F}(2t-1)(k-1)\right) \text{ and} \\ \sigma_k = \begin{cases} 1 & \text{if } k = 1, \\ \sqrt{2} & \text{otherwise.}\end{cases}\end{gathered} \tag{7}$$

In Eq. (6), $\mathbf{A}$ holds coefficients of linear combinations which approximate reconstructed 3D trajectories. $\mathbf{E}_{\text{linking}}$ connects or *links* these trajectories to unknown though valid combinations of basis trajectories. Depending on the linking strength, the calculated trajectories will more or less accurate resemble valid combinations of basis trajectories.

Finally, the regularisation term imposes a temporal coherence constraint on 3D trajectories of adjacent points. Since the recovered 3D trajectories are parameterised by $\mathbf{A}_k$, the regularisation term can be expressed as

$$\mathbf{E}_{\text{reg}}(\mathbf{A}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \|\nabla \mathbf{A}_{k,n}\|_\epsilon^2. \tag{8}$$

To calculate gradients of trajectory coefficients, Eq. (8) requires a point adjacency lookup table which is derived from the spatial arrangement of the points in the reference frame.

Our core NRSfM approach is called *Dense Consolidating Monocular Dynamic Reconstruction* (D-CMDR), as it unifies constraints in the metric and trajectory spaces into a single energy functional. In the beginning, $\mathbf{S}$ and $\mathbf{R}$ are initialised under rigidity assumption with [52] on the unaltered point tracks $\mathbf{W}$. α, λ and ρ are usually equivalued, while β is set an order of magnitude lower.

3.1.2 Postprocessing of DSP

After the reconstruction of the representative sequence, we obtain L shapes $\mathbf{S}_l^\sharp$, $l \in \{1, \dots, L\}$. The recovered poses are not applied to $\mathbf{S}_l^\sharp$ and discarded but a single global arbitrary pose for all $\mathbf{S}_l^\sharp$ can be chosen. Next, we build a map of pairs $\chi = (||\mathbf{S}_l^\sharp||_\mathcal{F}, \mathbf{S}_l^\sharp)$ where the shapes are arranged in the increasing order of $||\mathbf{S}_l^\sharp||_\mathcal{F}$. Starting from $\mathbf{S}_1^\sharp$, we iteratively include $\mathbf{S}_l^\sharp$ into DSP if the norm difference between the current $\mathbf{S}_l^\sharp$ and the latest included $\mathbf{D}_i$ exceeds some μ. By varying μ, we can control Q, *i.e.,* the cardinality of $\mathbf{D}$. Similarly to [21], we observe a strong correlation between $||\mathbf{S}_l^\sharp||_\mathcal{F}$ values and the corresponding shapes , *i.e.,* if Frobenius norms are similar, the shapes are close likewise[1].

[1] if shapes are related by reflection around the yz-plane (which is rarely the case in practice), they will have the same Frobenius norm

TB [4]	**MP** [37]	**VA** [15]	**DSTA** [12]	**CDF** [21]	**SMSR** [5]	**GM** [29]	**JM** [28]	**CMDR (ours)**
0.1252	0.0611	0.0346	0.0374	0.0886	0.0304	0.0294	0.280	0.0324
0.1348	0.0762	0.0379	0.0428	0.0905	0.0319	0.0309	0.327	0.0369

Table 1: Mean RMSE on *seq. A* (the first row) and *seq. B* (the second row).

4. Experimental Evaluation

This section outlines the evaluation methodology and summarises the results. We implement DSPR in C++ for a single thread. All values are reported for a system with 32 Gb RAM and Intel Core i7-6700K processor with cores running at 4.00GHz under Ubuntu 16.04.3.

4.1. Evaluation Methodology

We develop several tests with synthetic and real data for the evaluation of the convergence, accuracy and runtime aspects of DSPR. For the DSP reconstruction, we use several NRSfM methods based on different principles , *i.e.,* Variational Approach (VA) [15], Scalable Monocular Surface Reconstruction (SMSR) [5] and the proposed D-CMDR.

Depending on the evaluation scenario, we report different metrics characterising the accuracy of geometry and camera pose estimation. Let $\mathbf{S}'_f$ and $\mathbf{R}'_f$, $f \in \{1, \dots, F\}$, be the ground truth geometries and camera poses respectively. As a shape fidelity metric, we report a *mean root-mean-square error* (RMSE) for a set of views defined as $e_{3D} = \frac{1}{F}\sum_f \frac{||\mathbf{S}'_f - \mathbf{S}_f||_\mathcal{F}}{||\mathbf{S}'_f||_\mathcal{F}}$, where $||\cdot||_\mathcal{F}$ is Frobenius norm. Since the camera poses are recovered up to an arbitrary rotation, we find a single optimal corrective rotation $\mathbf{R}^\sharp$ aligning the recovered poses and the ground truth camera poses. Thus, for the evaluation purposes we solve the energy minimisation problem $\min_{\mathbf{R}^\sharp} \sum_f \left\|\mathbf{R}'_f - \mathbf{R}^\sharp\,\mathbf{R}_f\right\|_\epsilon$, with $\|\cdot\|_\epsilon$ denoting Huber norm with the threshold value $\epsilon = 1.0$. After applying $\mathbf{R}^\sharp$ to all $\mathbf{R}_f$, we compute a *mean quaternionic error* (QE) defined as $e_q = \frac{1}{F}\sum_f |\mathbf{q}'_f - \mathbf{q}_f|$, with $|\cdot|$ standing for the quaternion norm. $\mathbf{q}'_f$ and $\mathbf{q}_f$ are the quaternions[2] corresponding to $\mathbf{R}'_f$ and $\mathbf{R}^\sharp\mathbf{R}_f$ respectively.

Next, we evaluate the core D-CMDR approach individually (Sec. 4.2) and jointly with DSPR (Secs. 4.3-4.4). We perform self- and cross-convergence tests of DSPR with perturbed and missing data (Sec. 4.3), MSGD convergence tests (Sec. 4.4) and joint evaluation of flow and DSRP (Sec. 4.5). In the *self-convergence test,* DSP is reconstructed on ground truth point tracks, and the same tracks are used for the evaluation, whereas in the *cross-convergence test*, the point tracks for the reconstruction of the shape prior and the DSP are different. In Secs. 4.2–4.4, we use two 99 frames long synthetic face sequences with known geometry and dense point tracks from [15]. Both sequences A and B originate from the same set of facial

[2] here, the quaternions are guaranteed to have a positive sign

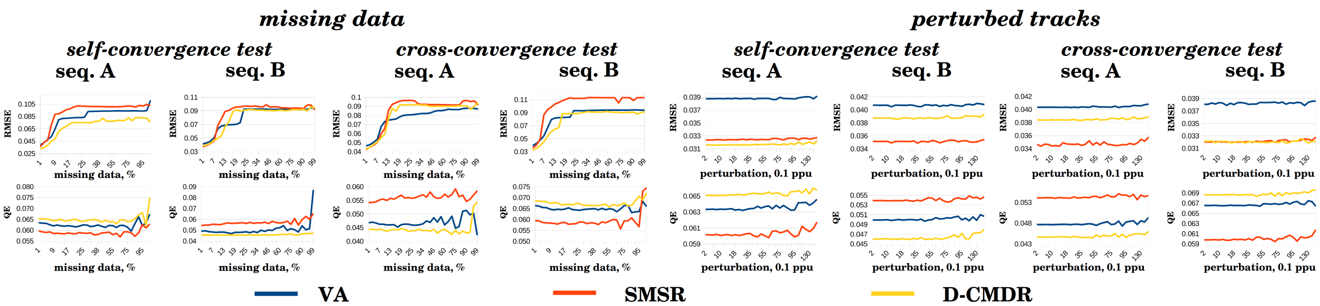


Figure 1: Experimental results of the *self-* and *cross-convergence* tests with missing and perturbed data. Three core approaches are tested, *i.e.,* VA [15], SMSR [5] and D-CMDR (ours). In all experiments, we report mean RMSE and QE as the functions of missing data ratio (in %) and perturbation (measured in 0.1 pixels per unit or *ppu*). Missing data is varied in the range $[0; 99]\%$, and the perturbation is varied in the range $[0; 15]$ pixels.

METHOD		PERTURBED DATA							MISSING DATA				
		0.4 px	**1.2 px**	**1.6 px**	**2.0 px**	**3.0 px**	**4.0 px**	**5.0 px**	**1%**	**3%**	**11%**	**17%**	**23%**
SMSR [5]	RMSE	0.0455	0.0962	0.1243	0.1536	0.2232	0.2956	0.3885	0.1001	0.1778	0.3365	0.4143	0.4849
	QE	0.2434	0.2999	0.3287	0.2450	0.3068	0.2280	0.3510	0.2972	0.2973	0.2968	0.2968	0.2975
D-CMDR (ours)	RMSE	0.0646	0.1918	0.2541	0.2867	0.3571	0.4056	0.4522	0.1001	0.1777	0.3365	0.4143	0.4849
	QE	0.0689	0.1077	0.1514	0.1711	0.4617	0.4578	0.4506	0.0663	0.0663	0.0662	0.0663	0.0663
DSPR (ours)	RMSE	**0.0324**	**0.0324**	**0.0324**	**0.0324**	**0.0324**	**0.0324**	**0.0325**	**0.0327**	**0.03578**	**0.0754**	**0.0962**	**0.0994**
	QE	**0.0602**	**0.0601**	**0.0600**	**0.0601**	**0.0603**	**0.0600**	**0.0603**	**0.0602**	**0.05984**	**0.0584**	**0.0585**	**0.0581**

Table 2: Mean RMSE and mean QE for SMSR [5] and D-CMDR (our method for DSP reconstruction) on perturbed tracks and tracks with missing entries.

expressions. The difference lies in the series of camera poses applied to the interpolated expressions. Due to the different camera pose patterns, the sequences are of varying difficulty for NRSfM and reconstructed differently (in many cases close to each other but not exactly in the same way). Thus, they offer an optimal testbed for the cross-convergence test. Finally, we show applications of DSPR on real data and report shape compression ratios (Sec. 4.6).

4.2. Evaluation of D-CMDR Separately from DSPR

Although D-CMDR is evaluated jointly with DSPR in the following, we report mean RMSE for it on synthetic faces, see Table 1. The errors for Trajectory Basis (TB) [4], Metric Projections (MP) [37], VA [15] and Dense Spatio-Temporal Approach (DSTA) [12] are replicated from [12], and the numbers for Coherent Depth Fields (CDF) [21], Grassmannian Manifold (GM) [29] and Jumping Manifolds (JM) [28] are taken from the original papers. We compute RMSE for SMSR [5] as the authors reported another metric.

Our approach is ranked fourth out of nine, and the gap between the most accurate methods is far less than 10^{-2}, which does not allow to generalise this result with confidence. Our RMSE is remarkably close to the currently most accurate GM/JM methods on these sequences, even though we design D-CMDR based on simpler principles.

By submitting our results to the NRSfM challenge [24], we additionally compare the proposed D-CMDR against more than fifteen methods, including TB, MP and SMSR on five evaluation scenarios. DSTA, CDF, GM and JM are not compared on the NRSfM challenge yet. As this dataset targets sparse and semi-dense reconstructions, we disable $\mathbf{E}_{\text{reg.}}$ and achieve the overall RMSE of $50.19\,mm$ outperforming multiple recent methods [13, 26, 11, 31] and coming close to [14] ($48.79\,mm$). For the *tricky* camera path, we obtain the RMSE across all sequences of $46.74\,mm$, which is among the best four results [24]. The most accurate camera trajectory for us is *zigzag* with RMSE of $36.69\,mm$ (ranked average across all methods). See [24] for more numbers.

4.3. Self- and Cross-Convergence Tests

The results of self- and cross-convergence tests with missing data and perturbed tracks are summarised in Fig.1. We ascertain that — due to the decoupling property of shapes and poses — DSP can be retrieved on a sequence with different shape poses compared to poses of the incoming measurements in the online mode, with virtually no influence on the reconstruction accuracy.

Missing Data. The amount of missing data is varied in the range $[0; 99]\%$. We observe that at 30%, RMSE largely stabilises, and QE is very stable even with up to 75% of missing data for three cases out of four. This shows that much fewer points are often sufficient to recover the camera pose. In the cross-convergence test for *seq.* A, a 50% threshold is identifiable for two DSP generation methods (VA and SMSR). After surpassing the threshold, the standard deviation of QE gradually increases, up to the exception of D-CMDR. In the latter case, QE is stable across the range of missing data patterns up to 90%.

Perturbed Tracks. In the case of the perturbed data, DSPR is stable and accurate in the whole tested range of $[0.1; 15]$ pixels of uniform perturbance per pixel. Across all experi-

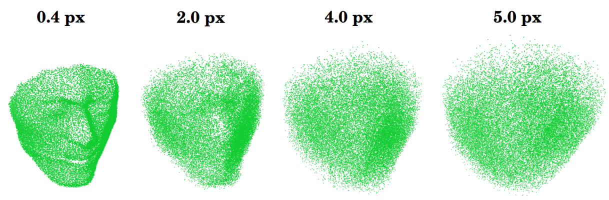


Figure 2: Reconstruction results of SMSR [5] on the perturbed point tracks (four different perturbation magnitudes), for frame 11 of *seq. A*.

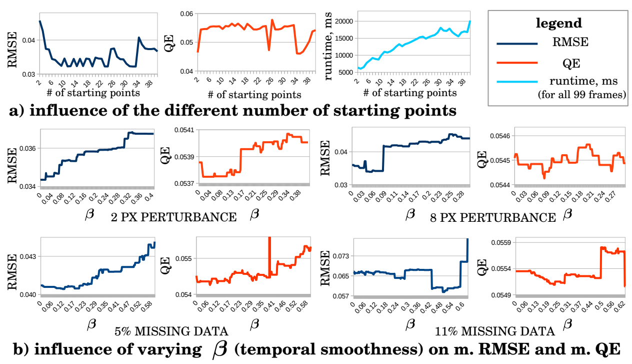


Figure 3: Results of the experiments with MSGD parameters: (a) the influence of the different number of starting points is evaluated on the measurements without noise; (b) the influence of β is evaluated on perturbed tracks and tracks with missing data with 20 starting points. In both cases, the shapes of *seq. B* are taken as a DSP and the clean tracks of *seq. A* are taken as the incoming dense point tracks. Mind the scaling of the y-axis.

ments and test cases, RMSE is kept on the same level of accuracy and is nearly uninfluenced by the perturbations. On the contrary, QE is slightly affected by the increasing perturbation amplitude. Still, there is no observable qualitative difference in the estimation of the camera poses.

Altogether, this is a notable result. Fig. 2 congregates selected outcomes of SMSR on perturbed point tracks of *seq. A* arranged in ascending order of deterioration. As the perturbation magnitude increases, the point scattering effects become more distinct. Already at 3-4 pixels, the structure is barely recognisable. Next, the appearance obtained on the tracks with missing data is reasonable but contains missing entries. Suddenly, with 23-25% of missing entries, no meaningful structure can be reconstructed by SMSR.

Table 2 summarises the metrics. Error patterns of the plain D-CMDR and SMSR are comparable. In contrast, DSPR operates on the tracks with amounts of missing data exceeding 25%. Even though the accuracy drops by the factor of 2-3, the structure remains recognisable, and the accuracy of camera pose estimation is only marginally affected.

4.4. Influence of the MSGD Parameters

The goal of this test is to examine the influence of the number of MSGD seeds, and verify that the temporal smoothness term affects reconstructions while optimising (1). Therefore, we fix α and vary β (in the range $[0.0; 0.63]$ with the step $2 \cdot 10^{-3}$) under a different number of MSGD starting points (in the range $[2; 40]$), see Fig. 3.

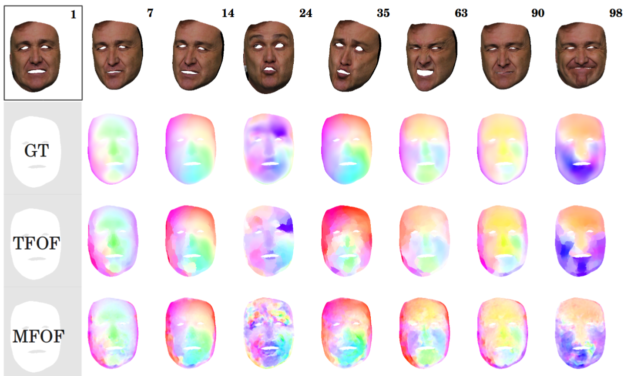
Varying Number of MSGD Starting Points. As expected, the runtime increases with the increasing number of seeds, and the dependency is close to a linear, see Fig. 3-(a). Starting from 6 seconds for two points, the runtime increases to 20 seconds for 40 points (for all 99 frames). For 10 and 25 starting points, RMSE is the smallest. In this region, we observe oscillations of the growing period and amplitude caused by regular shifts of the starting points and different convergence due to different camera poses. M. QE, on the contrary, does not correlate with the pattern of RMSE much and keeps at *ca.* 0.055. The latter phenomenon stems from the decoupled nature of the geometry and camera poses.

Varying β. Next, we vary β under four different types of noise — 2 and 8 pixels of uniform perturbances and 5% and 11% of missing entries (Fig. 3-(b)). With small disturbances (2 pixels perturbance and 5% of missing data), RMSE and QE vary slightly. The lowest errors are reached with small β. By and large, the errors are smaller for the case of smaller disturbances. For 8 pixels perturbance and 11% of missing data, the optimal metrics are achieved with a larger β ($\beta \approx 0.05$ for 8 pixels perturbance and $\beta \approx 0.5$ for 11% of missing data) suggesting that the shape smoothness term is more effective for more noisy point tracks.

4.5. Joint Evaluation of Flow and DSPR

Joint evaluation of input flow fields and NRSfM considers the influence of the dense correspondence tracking on the reconstructions. Even though still not being widespread in the NRSfM literature, it is highly relevant for practical scenarios. We perform a joint evaluation of DSP generation, the influence of optical flow and DSPR on the adapted *actor mocap* sequence [54] of 100 frames with $3.5 \cdot 10^4$ points in each shape. It contains ground truth geometry, camera poses, corresponding rendered images, a reference image with the face segmentation mask and ground truth multi-frame optical flow (MFOF), *i.e.,* a series of optical flows between the reference frame and every other frame in the sequence. In our modification, we rotate the ground truth surfaces and project them onto an image plane by ray tracing to render the images and the mask. The ground truth MFOF is obtained as the distances between the projections of the corresponding points in the image plane.

In addition to the ground truth MFOF, we compute dense correspondences by the method of Sun *et al.* [48] in the pairwise manner, as well as global MFOF with point trajectory regularisation over the whole batch [16]. The *average endpoint error* (AEPE) of two-frame optical flow (TFOF) and MFOF amount to 1.218 and 1.123, respectively. Next, we evaluate DSPR with the ground truth flow, TFOF and MFOF while using as DSP either ground truth geometry or



	GT shapes as DSP			CDF shapes as DSP		
	GT flow	TFOF	MFOF	GT flow	TFOF	MFOF
RMSE	0.00027	0.00042	0.00041	0.00525	0.00527	0.00527
QE	0.09136	0.09382	0.09271	0.12977	0.15017	0.13508

Figure 4: Exemplary frames from the adapted *actor mocap* sequence (first row), corresponding ground truth dense flow (second row), flow obtained by the method of Sun *et al.* [48] (third row) and MFOF [15] (fourth row). The table underneath lists errors for all evaluated combinations.

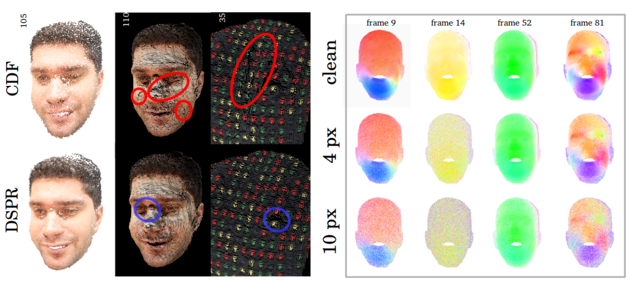


Figure 5: Exemplary reconstructions of CDF [21] and DSPR on point tracks with 10 pixels of perturbation magnitude (left column) and the comparison of the compressed states (the second and third columns). Compression artefacts highlighted in red are more pronounced for CDF, even though it achieves 2.34-2.65 times smaller compression ratio. The blue circles emphasise artefacts due to the tracking. The examples of clean and noisy point tracks with perturbances of 4 and 10 pixels are on the right.

shapes obtained by CDF [21] on the MFOF point tracks. In both cases, DSP contains 65 states after the compression. Fig. 4 shows exemplary images and different types of flow fields, while the associated table summarises the results. We see that the errors achieved with MFOF are slightly and consistently more accurate than those obtained with TFOF. Still, the TFOF errors do not worsen much attesting that DSPR tolerates less reliable and noisy point tracks.

4.6. Experiments with Real Data and Applications

We perform tests with real *face* [15], *back* [43], *liver* [34] and two *heart* [47, 46] sequences. Apart from the monocular non-rigid reconstruction, several other modes of operation are conceivable for DSPR. First, if we rerun DSPR on the point tracks which are used to compute DSP, we obtain a compressed version of the reconstructions. With the increasing density and the number of views, the space required for storage of a dynamic reconstruction grows fast. Especially in embedded and mobile devices, limits on the data bandwidth can become a bottleneck. Hence, compression of dynamic reconstructions is of high practical relevance. In the compression mode, we need to save a DSP, a shape prior identifier and a camera pose for every frame. This adds up to 12 bytes for camera pose in the axis-angle representation and one-two bytes for the shape prior identifier. Second, we are free to mix the sources of the DSP and incoming measurements. By computing correspondences between a reference frame of one sequence and frames observing a similar scene from another sequence, we can reenact 3D deformation states as if they were another scene.

μ	MSGD seeds	*face* [15]		*back* [43]	
		|DSP|	C	|DSP|	C
1.5	20	73	1.64	67	2.23
2.5	20	59	2.03	57	2.63
5	20	42	2.85	40	3.75
10.0	12	25	4.8	25	6
20.0	8	15	8	16	9.375
30.0	5	10	12	11	13.63
40.0	4	8	15	9	16.6

Table 3: The summary of the experiment for the compression of dynamic reconstructions with the achieved compression ratios.

DSPR for Shape Compression. We compare DSPR and CDF [21] — which is explicitly designed for compressible representations — for shape compression. We use MFOF [16] point tracks of *face* and *back* sequences, and extra prepare perturbed measurements of *face*. For the latter, CDF achieves compression ratio $C = 7.0$ on clean tracks. On the noisy tracks with the perturbation magnitudes of 4 and 10 pixels, its C decays to 3.0 and 1.582, respectively ($\epsilon = 1.6 \cdot 10^{-3}$). DSPR reaches $C = 8.0$ for $\mu = 20.0$ under 10 pixels of perturbations. If DSP is computed on clean reconstructions, the compression ratio is only weakly affected by the noise in point tracks, and only slight qualitative differences can be noticed (see the supplementary video). On the *back*, CDF achieves $C = 4.0$ ($\epsilon = 8 \cdot 10^{-4}$) and DSPR converges at $C = 9.375$ with $\mu = 20.0$. Recall that for DSPR, the longer an image sequence is, the higher are the compression ratios. The compression quality depends on how accurate the representative sequence for DSP generation reflects the shape space in the interactive mode. Fig. 5 compares the reconstructions obtained by CDF and DSRP. The left column shows the resulting states on noisy tracks (10 pixels of perturbation magnitude). As we see, especially with high compression ratios, CDF causes noticeable compression artefacts. For DSPR, Table 3 reports all combinations of the tested thresholds μ, corresponding DSP cardinalities, the number of MSGD seeds and the attained compression ratios for the *face* and *back* sequences.

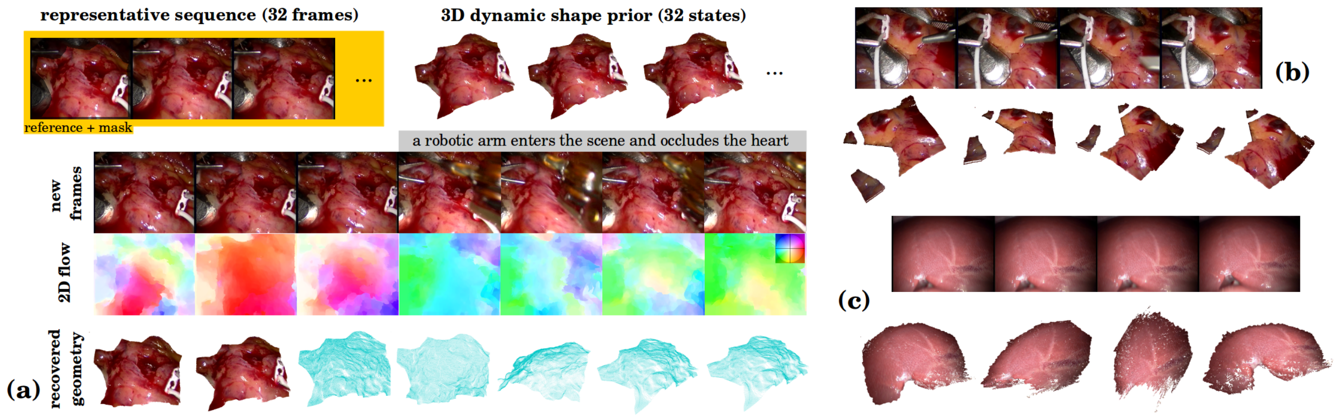


Figure 6: Application of DSPR in endoscopic scenarios with pronounced reoccurring deformations. (a): The first *heart* sequence [47]. The representative sequence and exemplary DSP states are shown in the top row. The new incoming frames and calculated flow fields visualised with Middlebury colour scheme [6] are given underneath. Our reconstructions from different perspectives are displayed in the bottom row. (b), (c): Exemplary frames and our reconstructions of another *heart* [47] and the *liver* sequences [34], respectively. Best viewed in colour. Furthermore, see our supplementary video.

DSPR in Endoscopy. Scenarios with temporally-disjoint rigidity often occur in the endoscopy. Examples are a human heart undergoing a series of repetitive contractions or a liver with periodic respiratory deformations. We test the proposed DSPR on two *heart* sequences from Stoyanov *et al.* [47, 46] and *liver* sequence from [34].

The first *heart* sequence contains 1573 frames in total[3]. For the DSP reconstruction, we choose 32 unoccluded frames — this duration corresponds to one complete cardiac cycle — and compute MFOF [49]. Next, we reconstruct a DSP with 32 states and $68k$ points per state, see Fig. 6-(a). The geometries during the diastole (refilling) and the systole (contraction) are all different, and, hence, we *do not* perform state compression. Next, we compute TFOF [48] between the reference and every remaining frame, and execute DSPR achieving five frames per seconds on our system ($\sim$4.1Mbps for the incoming optical flow). The reconstruction follows the cardiac cycle, *even if the robotic arm partially occludes the heart*. Following similar steps, we reconstruct the second *heart* sequence (899 frames, $55k$ points per surface), and the *liver* (250 frames, $54.5k$ points per surface). The *heart* sequence in Fig. 6-(b) also has occlusions, and the reconstruction distinctly reflects the cardiac cycle. The *liver* in Fig. 6-(c) contains large displacements in the point tracks. See our video with dynamic visualisations.

4.7. Discussion

We see that the substitution of deformation weights in classical low-rank NRSfM by a selection mechanism for each DSP element results in a fast and robust dense sequential NRSfM. We also witness that the proposed optimisation procedure is very quick, possibly faster than most of the NRSfM algorithms in the literature. In the online mode, DSPR achieves multiple frames per second, and the throughput can be further increased by parallelisation.

All in all, we match the state-of-the-art performance while drastically improving the robustness on deteriorated point tracks — in reality, data is often far from perfect. Last but not least, the compression evaluation is also rarely seen in NRSfM but is of practical relevance.

[3] its shorter parts are often used in the NRSfM evaluation [2, 5, 12, 29]

5. Conclusions and Future Directions

We introduce a new hybrid NRSfM method relying on temporally-disjoint rigidity effects. In the first step of our DSPR approach, we reconstruct a representative set of views and generate DSP. Next, for new incoming dense point tracks, we solve a light-weight optimisation problem with a zero-norm which selects the closest shape from DSP while positioning it as observed in the measurements.

The robustness to inaccurate point tracks, the possibility to use faster and less accurate dense flow fields, the highest compression ratios and the suitability of the proposed technique for medical applications with repetitive deformations significantly broaden the scope of modern NRSfM, especially when handling real data. We show experimentally that DSPR successfully bridges the gap between the accuracy of dense correspondences and reconstructions, and we believe that it can have a high practical impact. Furthermore, our light-weight dense incremental NRSfM can enable various new applications in augmented reality.

There are multiple future work directions. Though not explicitly tested, the sparse setting can also be investigated, because runtime of dense NRSfM has always been more of an issue. DSPR can be deployed on a low-power consumption device such as augmented reality glasses for applications involving deformable objects. Moreover, DSP signatures are worth trying for object class recognition.

References

[1] A. Agudo, J. M. M. Montiel, L. Agapito, and B. n. Calvo. Online dense non-rigid 3D shape and camera motion recovery. In *British Machine Vision Conference (BMVC)*, 2014.

[2] A. Agudo and F. Moreno-Noguer. Learning shape, motion and elastic models in force space. In *International Conference on Computer Vision (ICCV)*, pages 756–764, 2015.

[3] A. Agudo and F. Moreno-Noguer. A scalable, efficient, and accurate solution to non-rigid structure from motion. *Computer Vision and Image Understanding (CVIU)*, 167:121–133, 2018.

[4] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Trajectory space: A dual representation for nonrigid structure from motion. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(7):1442–1456, 2011.

[5] M. D. Ansari, V. Golyanik, and D. Stricker. Scalable dense monocular surface reconstruction. In *International Conference on 3D Vision (3DV)*, 2017.

[6] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision (IJCV)*, 92(1):1–31, 2011.

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

[8] M. Brand. A direct method for 3d factorization of nonrigid motion observed in 2d. In *Computer Vision and Pattern Recognition (CVPR)*, pages 122–128, 2005.

[9] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *Computer Vision and Pattern Recognition (CVPR)*, pages 690–696, 2000.

[10] A. D. Bue. A factorization approach to structure from motion with shape priors. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[11] A. Chhatkuli, D. Pizarro, T. Collins, and A. Bartoli. Inextensible non-rigid shape-from-motion by second-order cone programming. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] Y. Dai, H. Deng, and M. He. Dense non-rigid structure-from-motion made easy – a spatial-temporal smoothness based solution. In *International Conference on Image Processing (ICIP)*, pages 4532–4536, 2017.

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

[14] A. Del Bue, J. Xavier, L. Agapito, and M. Paladini. Bilinear factorization via augmented lagrange multipliers. In *European Conference on Computer Vision (ECCV)*, pages 283–296, 2010.

[15] R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1272–1279, 2013.

[16] R. Garg, A. Roussos, and L. Agapito. A variational approach to video registration with subspace constraints. *International Journal of Computer Vision (IJCV)*, 104(3):286–314, 2013.

[17] V. Golyanik. Shape priors in dense non-rigid structure from motion (chapter 5). In *Robust Methods for Dense Monocular Non-Rigid 3D Reconstruction and Alignment of Point Clouds*, pages 89–133. Springer Nature, 2020.

[18] V. Golyanik, T. Fetzer, and D. Stricker. Accurate 3d reconstruction of dynamic scenes from monocular image sequences with severe occlusions. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 282–291, 2017.

[19] V. Golyanik, A. Jonas, and D. Stricker. Consolidating segmentwise non-rigid structure from motion. In *Machine Vision Applications (MVA)*, 2019.

[20] V. Golyanik and D. Stricker. Dense batch non-rigid structure from motion in a second. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 254–263, 2017.

[21] V. Golyanik and D. Stricker. High dimensional space model for dense monocular surface recovery. In *International Conference on 3D Vision (3DV)*, 2017.

[22] P. F. U. Gotardo and A. M. Martinez. Computing smooth time trajectories for camera and deformable shape in structure from motion with occlusion. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(10):2051–2065, 2011.

[23] P. F. U. Gotardo and A. M. Martinez. Non-rigid structure from motion with complementary rank-3 spaces. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3065–3072, 2011.

[24] S. Hoppe Nesgaard Jensen, A. Del Bue, M. E. Brix Doest, and H. Aanæs. A Benchmark and Evaluation of Non-Rigid Structure from Motion. *arXiv e-prints*, 2018.

[25] L. Kausch, A. Hilsmann, and P. Eisert. Template-based 3d non-rigid shape estimation from monocular image sequences. In *Vision, Modeling & Visualization (VMV)*, 2017.

[26] C. Kong and S. Lucey. Prior-less compressible structure from motion. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[27] O. Kovalenko, V. Golyanik, J. Malik, A. Elhayek, and D. Stricker. Structure from Articulated Motion: An Accurate and Stable Monocular 3D Reconstruction Approach without Training Data. *arXiv e-prints*, 2019.

[28] S. Kumar. Jumping manifolds: Geometry aware dense non-rigid structure from motion. In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[29] S. Kumar, A. Cherian, Y. Dai, and H. Li. Scalable dense non-rigid structure-from-motion: A grassmannian perspective. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[30] S. Kumar, Y. Dai, and H. Li. Spatio-temporal union of subspaces for multi-body non-rigid structure-from-motion. *Pattern Recognition*, 71:428 – 443, 2017.

[31] M. Lee, J. Cho, and S. Oh. Consensus of non-rigid reconstructions. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[32] M. Lee, J. Cho, and S. Oh. Procrustean normal distribution for non-rigid structure from motion. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39:1388–1400, 2017.

[33] X. Li, H. Li, H. Joo, Y. Liu, and Y. Sheikh. Structure from recurrent motion: From rigidity to recurrency. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[34] P. Mountney, D. Stoyanov, and G.-Z. Yang. Three-dimensional tissue deformation recovery and tracking. *IEEE Signal Processing Magazine*, 27(4):14–24, 2010.

[35] S. I. Olsen and A. Bartoli. Implicit non-rigid structure-from-motion with priors. *Journal of Mathematical Imaging and Vision (JMIV)*, 31(2-3):233–244, 2008.

[36] M. Paladini, A. Bartoli, and L. Agapito. Sequential non-rigid structure-from-motion with the 3d-implicit low-rank shape model. In *European Conference on Computer Vision (ECCV)*, pages 15–28, 2010.

[37] M. Paladini, A. Del Bue, J. a. Xavier, L. Agapito, M. Stosić, and M. Dodig. Optimal metric projections for deformable and articulated structure-from-motion. *International Journal of Computer Vision (IJCV)*, 96(2):252–276, 2012.

[38] M. Perriollat, R. Hartley, and A. Bartoli. Monocular template-based reconstruction of inextensible surfaces. *International Journal of Computer Vision (IJCV)*, 95(2):124–137, 2011.

[39] A. Pumarola, A. Agudo, L. Porzi, A. Sanfeliu, V. Lepetit, and F. Moreno-Noguer. Geometry-aware network for non-rigid shape prediction from a single view. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[40] N. Qian, J. Wang, F. Mueller, F. Bernard, V. Golyanik, and C. Theobalt. HTML: A Parametric Hand Texture Model for 3D Hand Reconstruction and Personalization. In *European Conference on Computer Vision (ECCV)*, 2020.

[41] R. Rabaud and S. Belongie. Re-thinking non-rigid structure from motion. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[42] A. Rehan, A. Zaheer, I. Akhter, A. Saeed, M. H. Usmani, B. Mahmood, and S. Khan. Nrsfm using local rigidity. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 69–74, 2014.

[43] C. Russell, J. Fayad, and L. Agapito. Energy based multiple model fitting for non-rigid structure from motion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3009–3016, 2011.

[44] S. Shimada, V. Golyanik, C. Theobalt, and D. Stricker. IsMo-GAN: Adversarial learning for monocular non-rigid 3d reconstruction. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

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

[46] D. Stoyanov. Stereoscopic scene flow for robotic assisted minimally invasive surgery. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 479–486, 2012.

[47] D. Stoyanov, G. Mylonas, F. Deligianni, A. Darzi, and G. Z. Yang. Soft-tissue motion tracking and structure estimation for robotic assisted mis procedures. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 139–146, 2005.

[48] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2439, 2010.

[49] B. Taetz, G. Bleser, V. Golyanik, and D. Stricker. Occlusion-aware video registration for highly non-rigid objects. In *Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[50] J. Taylor, A. D. Jepson, and K. N. Kutulakos. Non-rigid structure from locally-rigid motion. In *Computer Vision and Pattern Recognition (CVPR)*, 2010.

[51] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and T. Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *International Conference on Computer Vision (ICCV)*, 2017.

[52] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision (IJCV)*, 9(2):137–154, 1992.

[53] L. Torresani, A. Hertzmann, and C. Bregler. Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(5):878–892, 2008.

[54] L. Valgaerts, C. Wu, A. Bruhn, H.-P. Seidel, and C. Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. *ACM Transactions on Graphics (TOG)*, 31(6):187:1–187:11, 2012.

[55] R. Yu, C. Russell, N. D. F. Campbell, and L. Agapito. Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video. In *International Conference on Computer Vision (ICCV)*, 2015.

[56] Y. Zhu, D. Huang, F. D. L. Torre, and S. Lucey. Complex non-rigid motion 3d reconstruction by union of subspaces. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1542–1549, 2014.