Neural Dense Non-Rigid Structure from Motion with Latent Space Constraints*

Vikramjit Sidhu[1,2] Edgar Tretschk[1]
Vladislav Golyanik[1] Antonio Agudo[3] Christian Theobalt[1]

[1]Max Planck Institute for Informatics, SIC [2]Saarland University, SIC
[2]Institut de Robòtica i Informàtica Industrial, CSIC-UPC

Abstract. We introduce the first dense neural non-rigid structure from motion (N-NRSfM) approach, which can be trained end-to-end in an unsupervised manner from 2D point tracks. Compared to the competing methods, our combination of loss functions is fully-differentiable and can be readily integrated into deep-learning systems. We formulate the deformation model by an auto-decoder and impose subspace constraints on the recovered latent space function in a frequency domain. Thanks to the state recurrence cue, we classify the reconstructed non-rigid surfaces based on their similarity and recover the period of the input sequence. Our N-NRSfM approach achieves competitive accuracy on widely-used benchmark sequences and high visual quality on various real videos. Apart from being a standalone technique, our method enables multiple applications including shape compression, completion and interpolation, among others. Combined with an encoder trained directly on 2D images, we perform scenario-specific monocular 3D shape reconstruction at interactive frame rates. To facilitate the reproducibility of the results and boost the new research direction, we open-source our code and provide trained models for research purposes[1].

Keywords: Neural non-rigid structure from motion, sequence period detection, latent space constraints, deformation auto-decoder.

1 Introduction

Non-Rigid Structure from Motion (NRSfM) reconstructs non-rigid surfaces and camera poses from monocular image sequences using multi-frame 2D correspondences calculated across the input views. It relies on motion and deformation cues as well as weak prior assumptions, and is object-class-independent in contrast to monocular 3D reconstruction methods which make use of parametric

* Supported by the ERC Consolidator Grant 4DReply (770784) and the Spanish Ministry of Science and Innovation under project HuMoUR TIN2017-90086-R. The authors thank Mallikarjun B R for help with running the FML method [58] on our data. Please contact the corresponding author via golyanik@mpi-inf.mpg.de.

[1] `http://gvv.mpi-inf.mpg.de/projects/Neural_NRSfM/`

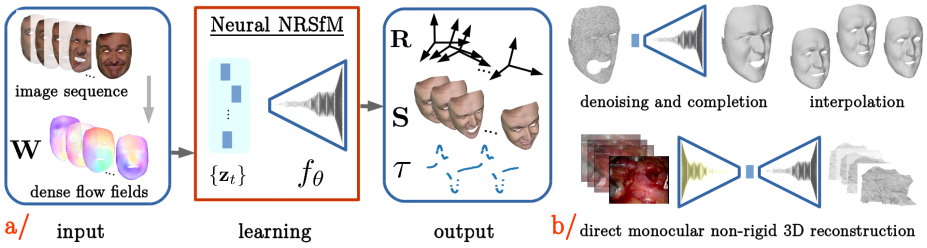

Fig. 1: Neural non-rigid structure from motion (N-NRSfM). Our approach reconstructs monocular image sequences in 3D from dense flow fields (shown using the Middlebury optical flow scheme [9]). In contrast to all other methods, we represent the deformation model with a neural auto-decoder $f_{\boldsymbol{\theta}}$ which decodes latent variables $\mathbf{z}_t$ into 3D shapes (a/). This brings a higher expressivity and flexibility which results in state-of-the-art results and new applications such as shape completion, denoising and interpolation, as well as direct monocular non-rigid 3D reconstruction (b/).

models [59]. Dense NRSfM has achieved remarkable progress during the last several years [1,8,19,37,51]. While the accuracy of dense NRSfM has been recently only marginally improved, learning-based direct methods for monocular rigid and non-rigid 3D reconstruction have become an active research area in computer vision [13,33,47,54,66].

Motivated by these advances, we make the first step towards learning-based dense NRSfM, as it can be seen in Fig. 1. At the same time, we remain in the classical NRSfM setting without strong priors (which restrict to object-specific scenarios) or assuming the availability of training data with 3D geometry. We find that among several algorithmic design choices, replacing an explicit deformation model by an implicit one, *i.e.,* a neural network with latent variables for each shape, brings multiple advantages and enables new applications compared to the previous work such as temporal state segmentation, shape completion, interpolation and direct monocular non-rigid 3D reconstruction (see Fig. 1-b/ for some examples).

By varying the number of parameters in our neural component, we can express our assumption on the complexity of the observed deformations. We observe that most real-world deformations evince state recurrence which can serve as an additional reconstruction constraint. By imposing constraints on the latent space, we can thus detect a period of the sequence, denoted by τ, *i.e.,* the duration in frames after which the underlying non-rigid 3D states repeat, and classify the recovered 3D states based on their similarity. Next, by attaching an image encoder to the learnt neural deformation model (deformation auto-decoder), we can perform in testing direct monocular non-rigid 3D reconstruction at interactive frame rates. Moreover, an auto-decoder represents non-rigid states in a compressed form due to its compactness.

Note that the vast majority of the energy functions proposed in the literature so far is not fully differentiable or cannot be easily used in learning-based systems due to computational or memory requirements [1,8,19,37]. We combine a data

loss, along with constraints in the metric and trajectory spaces, a temporal smoothness loss as well as latent space constraints into single energy — with the non-rigid shape parametrised by an auto-decoder — and optimise it with the back-propagation algorithm [49]. The experimental evaluation indicates that the proposed N-NRSfM approach obtains competitive solutions in terms of 3D reconstruction, and outperforms competing methods on several sequences, but also represents a useful tool for non-rigid shape analysis and processing.

Contributions. In summary, the primary contributions of this work are:

- ⋆ The first, to the best of our belief, fully differentiable dense neural NRSfM approach with a novel auto-decoder-based deformation model (Secs. 3, 4);
- ⋆ Subspace constraints on the latent space imposed in the Fourier domain. They enhance the reconstruction accuracy and enable temporal classification of the recovered non-rigid 3D states with period detection (Sec. 4.2);
- ⋆ Several applications of the deformation model including shape compression, interpolation and completion, as well as fast direct non-rigid 3D reconstruction from monocular image sequences (Sec. 4.4);
- ⋆ An extensive experimental evaluation of the core N-NRSfM technique and its applications with state-of-the-art results (Sec. 5).

2 Related Work

Recovering a non-rigid 3D shape from a single monocular camera has been an active research area in the past two decades. In the literature, two main classes of approaches have proved most effective so far: template-based formulations and NRSfM. On the one hand, template-based approaches relied on establishing correspondences with a reference image in which the 3D shape is already known in advance [42,53]. To avoid ambiguities, additional constraints were included in the optimisation, such as the inextensibility [42,65], as rigid as possible priors [68], providing very robust solutions but limiting its applicability to almost inelastic surfaces. While the results provided by template-based approaches are promising, knowing a 3D template in advance can become a hard requirement. In order to avoid that, NRSfM approaches have reduced these requirements, making their applicability easier. In this context, NRSfM has been addressed in the literature by means of model-based approaches, and more recently, by the use of deep-learning-based methods. We next review the most related work to solve this problem by considering both perspectives.

Non-Rigid Structure from Motion. NRSfM has been proposed to solve the problem from 2D tracking data in a monocular video (in the literature, 2D trajectories are collected in a measurement matrix). The most standard approach to address the inherent ambiguity of the NRSfM problem is by assuming the underlying 3D shape is low-rank. In order to estimate such low-rank model, both factorisation- [11] and optimisation-based approaches [43,61] have been proposed, considering single low-dimensional shape spaces [16,19], or a union of temporal [69] or spatio-temporal subspaces [3]. Low-rank models were also

extended to the other domains, by exploiting pre-defined trajectory basis [7], the combination of shape-trajectory vectors [28,29], and the force space that induces the deformations [5]. On top of these models, additional spatial [38] or temporal [2,10,39] smoothness constraints, as well as shape priors [12,21,35] have also been considered. However, in contrast to their rigid counterparts, NRSfM methods are typically sparse, limiting their application to a small set of salient points. Whereas several methods are adaptations of sparse techniques to dense data [22,51], other techniques were explicitly designed for the dense setting [1,19,37] relying on sophisticated optimisation strategies.

Neural Monocular Non-Rigid 3D Reconstruction. Another possibility to perform monocular non-rigid 3D reconstruction is to use learning-based approaches. Recently, many works have been presented for rigid [13,18,30,40,66] and non-rigid [27,47,54,62] shape reconstruction. These methods exploited a large and annotated dataset to learn the solution space, limiting their applicability to the type of shapes that are observed in the dataset. Unfortunately, this supervision is a hard task to be handled in real applications, where the acquisition of 3D data to train a neural network is not trivial.

While there has been work at the intersection of NRSfM and deep learning, the methods require large training datasets [34,41,52] and address only the sparse case [34,41]. *C3DPO* [41] learns basis shapes from 2D observations and does not require 3D supervision, similar to our approach. Neural methods for monocular non-rigid reconstruction have to be trained for every new object class or shape configuration within the class. In contrast to the latter methods — and similar to the classical NRSfM — we solely rely on motion and deformation cues. Our approach is unsupervised and requires only dense 2D point tracks for the recovery of non-rigid shapes. Thus, we combine the best of both worlds, *i.e.,* the expressivity of neural representations for deformation models and improvements upon weak prior assumptions elaborated in previous works on dense NRSfM. We leverage the latter in the way so that we find an energy function which is fully differentiable and can be optimised with modern machine-learning tools.

3 Revisiting NRSfM

We next review the NRSfM formulation that will be used later to describe our neural approach. Let us consider a set of P points densely tracked across T frames. Let $\mathbf{s}_t^p = [x_t^p, y_t^p, z_t^p]^\top$ be the 3D coordinates of the p-th point in image t, and $\hat{\mathbf{w}}_t^p = [u_t^p, v_t^p]^\top$ its 2D position according to an orthographic projection. In order to simplify subsequent formulation, the camera translation $\mathbf{t}_t = \sum_p \hat{\mathbf{w}}_t^p / P$ can be subtracted from the 2D projections, considering centred measurements as $\mathbf{w}_t^p = \hat{\mathbf{w}}_t^p - \mathbf{t}_t$. We can then build a linear system to map the 3D-to-2D point coordinates as:

$$\underbrace{\begin{bmatrix} \mathbf{w}_1^1 & \dots & \mathbf{w}_1^P \\ \vdots & \ddots & \vdots \\ \mathbf{w}_T^1 & \dots & \mathbf{w}_T^P \end{bmatrix}}_{\mathbf{W}} = \underbrace{\begin{bmatrix} \mathbf{R}_1 & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{R}_T \end{bmatrix}}_{\mathbf{R}} \underbrace{\begin{bmatrix} \mathbf{s}_1^1 & \dots & \mathbf{s}_1^P \\ \vdots & \ddots & \vdots \\ \mathbf{s}_T^1 & \dots & \mathbf{s}_T^P \end{bmatrix}}_{\mathbf{S}}, \tag{1}$$

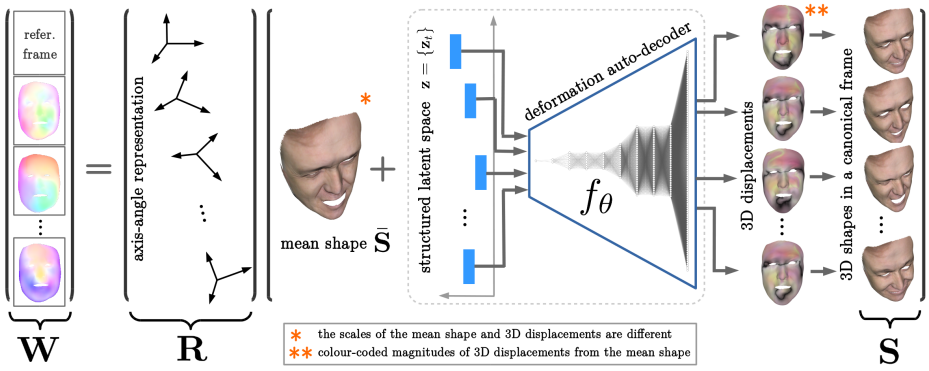

Fig. 2: Overview of our N-NRSfM approach to factorise a measurement input matrix W into motion R and shape S factors. To enable an end-to-end learning, we formulate a fully-differentiable neural energy function, where each $\mathbf{S}_t$ is mapped by means of a deformation auto-decoder $f_{\boldsymbol{\theta}}$ from a latent space $\mathbf{z}_t$, plus a mean shape $\bar{\mathbf{S}}$. After obtaining optimal network parameters $\boldsymbol{\theta}$, the latent space becomes structured allowing the scene deformation pattern analysis.

where $\mathbf{W}$ is a $2T \times P$ measurement matrix with the 2D measurements arranged in columns, $\mathbf{R}$ is a $2T \times 3T$ block diagonal matrix made of T truncated 2×3 camera rotations $\mathbf{R}_t \equiv \mathbf{\Pi}\mathbf{G}_t$ with the full rotation matrix $\mathbf{G}_t$ and $\mathbf{\Pi} = \begin{bmatrix}1&0&0\\0&1&0\end{bmatrix}$; and $\mathbf{S}$ is a $3T \times P$ matrix with the non-rigid 3D shapes. Every $\mathbf{G}_t$ lies in the $SO(3)$ group, that we enforce using an axis-angle representation encoding the rotation by a vector $\boldsymbol{\alpha}_t = (\alpha_t^x, \alpha_t^y, \alpha_t^z)$, that can be related to $\mathbf{G}_t$ by the Rodrigues' rotation formula. On balance, the problem consists in estimating the time-varying 3D shape $\mathbf{S}^t$ as well as the camera motion $\mathbf{G}^t$ with $t = \{1, \ldots, T\}$, from 2D trajectories $\mathbf{W}$.

4 Deformation Model with Shape Auto-Decoder

In the case of dynamic objects, the 3D shape changes as a function of time. Usually, this function is unknown, and many efforts have been made to model it. The type of deformation model largely determines which observed non-rigid states can be accurately reconstructed, *i.e.,* the goal is to find a simple model with large expressibility. In this context, perhaps the most used model in the literature consists in enforcing the deformation shape to lie in a linear subspace [11]. While this model has been proved to be effective, the form in which the shape bases are estimated can be decisive. For example, it is well known that some constraints cannot be effectively imposed in factorisation methods [11,67], forcing the proposal of more sophisticated optimisation approaches [3,16,69]. In this paper, we propose to depart from the traditional formulations based on linear subspace models and embrace a different formulation that can regress the deformation modes in a unsupervised manner during a neural network training,

see Fig. 2 for a method overview. By controlling the architecture and composition of the layers, we can express our assumptions about the complexity and type of the observed deformations. We will use the name of Neural Non-Rigid Structure from Motion (N-NRSfM) to denote our approach.

4.1 Modelling Deformation with Neural Networks

We propose to implement our non-rigid model network as a deformation auto-decoder $f_{\boldsymbol{\theta}}$, as it was done for rigid shape categories [44], where $\boldsymbol{\theta}$ denotes the learned network parameters. Specifically, we construct $f_{\boldsymbol{\theta}}$ as a series of nine fully-connected layers with small hidden dimensions $(2, 8, 8, 8, 16, 32, 32, B, |\mathbf{S}_t|)$, and exponential linear unit (ELU) activations [14] (except after the penultimate and final layers). B — set to 32 by default — can be interpreted as an analogue to the number of basis shapes in linear subspace models. $f_{\boldsymbol{\theta}}$ is a function of the latent space $\mathbf{z}_t$, that is related to the shape space $\mathbf{S}_t$ by means of:

$$\mathbf{S}_t = \bar{\mathbf{S}} + f_{\boldsymbol{\theta}}(\mathbf{z}_t), \tag{2}$$

where $\bar{\mathbf{S}}$ is a $3 \times P$ mean shape matrix. We can also obtain the time-varying shape $\mathbf{S}$ in Eq. (1) by $\mathbf{S} = (\mathbf{1}_T \otimes \bar{\mathbf{S}}) + f_{\boldsymbol{\theta}}(\mathbf{z})$, with $\mathbf{1}_T$ a T-dimensional vector of ones and $\otimes$ a Kronecker product. The fully-connected layers of $f_{\boldsymbol{\theta}}$ are initialised using He initialisation [31], and the bias value of the last layer is set to a rigid shape estimate $\bar{\mathbf{S}}$, which is kept fixed during optimisation. Both $\bar{\mathbf{S}}$ and $\mathbf{R}_t$ with $t = \{1, \dots, T\}$ are initialised by rigid factorisation [60] from $\mathbf{W}$. Note that we estimate displacements (coded by $f_{\boldsymbol{\theta}}(\mathbf{z}_t)$) from $\bar{\mathbf{S}}$ instead of absolute point positions. Considering that, the weight matrix of the final fully-connected layer of $f_{\boldsymbol{\theta}}$ can be interpreted as a low-rank linear subspace where every vector denotes a 3D displacement from the mean shape. This contributes to the compactness of the recovered space and serves as an additional constraint, similar to the common practice of the principal component analysis [46].

To learn $\boldsymbol{\theta}$, and update it during training, we require gradients with respect to a full energy $\mathbf{E}$ that we will propose later, such that:

$$\frac{\partial \mathbf{E}}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{T} \frac{\partial \mathbf{E}}{\partial \mathbf{S}_t} \, \frac{\partial \mathbf{S}_t}{\partial \boldsymbol{\theta}} \,, \tag{3}$$

connecting $f_{\boldsymbol{\theta}}$ into a fully-differentiable loss function, in which $\mathbf{S}_t$, $t = \{1, \dots, T\}$ are optimised as free variables via gradients. We next describe our novel energy function $\mathbf{E}$, which is compatible with $f_{\boldsymbol{\theta}}$ and supports gradient back-propagation.

4.2 Differentiable Energy Function

To solve the NRSfM problem as it was defined in Section 3, we propose to minimise a differentiable energy function with respect to motion parameters $\mathbf{R}$ and shape ones (coded by $\boldsymbol{\theta}$ and $\mathbf{z}$) as:

$$\mathbf{E} = \mathbf{E}_{\text{data}}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{R}) + \beta \, \mathbf{E}_{\text{temp}}(\boldsymbol{\theta}, \mathbf{z}) + \gamma \, \mathbf{E}_{\text{spat}}(\boldsymbol{\theta}, \mathbf{z}) + \eta \, \mathbf{E}_{\text{traj}}(\boldsymbol{\theta}, \mathbf{z}) + \omega \, \mathbf{E}_{\text{latent}}(\mathbf{z}), \tag{4}$$

where $\mathbf{E}_{\text{data}}$ is a data term, and $\{\mathbf{E}_{\text{temp}}, \mathbf{E}_{\text{spat}}, \mathbf{E}_{\text{traj}}, \mathbf{E}_{\text{latent}}\}$ encode the priors that we consider. β, γ, η and ω are weight coefficients to balance the influence of every term. We now describe each of these terms in detail.

The data term $\mathbf{E}_{\text{data}}$ is derived from the projection equation (1), and it is to penalise the image re-projection errors as:

$$\mathbf{E}_{\text{data}}(\boldsymbol{\theta}, \mathbf{z}, \mathbf{R}) = \left\| \mathbf{W} - \mathbf{R}\left((\mathbf{1}_T \otimes \bar{\mathbf{S}}) + f_{\boldsymbol{\theta}}(\mathbf{z})\right) \right\|_{\epsilon}, \tag{5}$$

where $\|\cdot\|_{\epsilon}$ denotes the Huber loss of a matrix.

The temporal smoothness term $\mathbf{E}_{\text{temp}}$ enforces temporal-preserving regularisation of the 3D shape via its latent space as:

$$\mathbf{E}_{\text{temp}}(\boldsymbol{\theta}, \mathbf{z}) = \sum_{t=1}^{T-1} \left\| f_{\boldsymbol{\theta}}(\mathbf{z}_{t+1}) - f_{\boldsymbol{\theta}}(\mathbf{z}_t) \right\|_{\epsilon}. \tag{6}$$

Thanks to this soft-constraint prior, our algorithm can generate clean surfaces that also stabilise the camera motion estimation.

The spatial smoothness term $\mathbf{E}_{\text{spat}}$ imposes spatial-preserving regularisation for a neighbourhood. This is especially relevant for dense observations, where most of the points in a local neighbourhood can follow a similar motion pattern. To define this constraint, let $\mathcal{N}(\mathbf{p})$ be a 1-ring neighbourhood of $\mathbf{p} \in \mathbf{S}_t$, that will be used to define a Laplacian term (widely used in computer graphics [55]). For robustness, we complete the spatial smoothness with a depth penalty term. Combining both ideas, we define this term as:

$$\mathbf{E}_{\text{spat}}(\boldsymbol{\theta}, \mathbf{z}) = \underbrace{\sum_{t=0}^{T-1} \sum_{\mathbf{p} \in \mathbf{S}_t} \left\| \mathbf{p} - \frac{1}{|\mathcal{N}(\mathbf{p})|} \sum_{\mathbf{q} \in \mathcal{N}(\mathbf{p})} \mathbf{q} \right\|_1}_{\text{Laplacian smoothing}} - \lambda \underbrace{\sum_{t=1}^{T} \left\| \mathcal{P}_z(\mathbf{G}_t \mathbf{S}_t) \right\|_2}_{\text{depth control}}, \tag{7}$$

where $\mathcal{P}_z$ denotes an operator to extract z-coordinates, $\|\cdot\|_1$ and $\|\cdot\|_2$ are the l_1- and l_2-norm, respectively, and $\lambda > 0$ is a weight coefficient. Thanks to the depth term, our N-NRSfM approach automatically achieves more supervision over the z-coordinate of the 3D shapes, since it can lead to an increase in the shape extent along the z-axis.

The point trajectory term $\mathbf{E}_{\text{traj}}$ imposes a subspace constraint on point trajectories throughout the whole sequence, as it was exploited by [6,7]. To this end, the 3D point trajectories are coded by a linear combination of K fixed trajectory vectors by a $T \times K$ matrix $\boldsymbol{\Phi}$ together with a $3K \times P$ matrix $\mathbf{A}$ of unknown coefficients. The penalty term can be then written as:

$$\mathbf{E}_{\text{traj}}(\boldsymbol{\theta}, \mathbf{z}) = \left\| (\mathbf{1}_T \otimes \bar{\mathbf{S}}) + f_{\boldsymbol{\theta}}(\mathbf{z}) - (\boldsymbol{\Phi} \otimes \mathbf{I}_3)\mathbf{A} \right\|_{\epsilon}, \quad \boldsymbol{\Phi} = \begin{pmatrix} \phi_{1,1} & \dots & \phi_{1,K} \\ \vdots & \ddots & \vdots \\ \phi_{T,1} & \dots & \phi_{T,K} \end{pmatrix}, \tag{8}$$

where $\phi_{t,k} = \frac{\sigma_k}{\sqrt{2}} \cos\left(\frac{\pi}{2T}(2t-1)(k-1)\right)$, with $\sigma_k = 1$ for $k = 1$, and $\sigma_k = \sqrt{2}$, otherwise. $\mathbf{I}_3$ is a 3×3 identity matrix. We experimentally find that this term is not redundant with the rest of terms, and it provides a soft regularisation of $f_{\boldsymbol{\theta}}$.

Finally, the latent term $\mathbf{E}_{\text{latent}}$ imposes sparsity constraints over the latent vector $\mathbf{z}$. This type of regularisation is enabled by the new form to express the deformation model with an auto-decoder $f_{\boldsymbol{\theta}}$, and it can be expressed as:

$$\mathbf{E}_{\text{latent}}(\mathbf{z}) = \|\mathcal{F}(\mathbf{z})\|_1 \,, \tag{9}$$

where $\mathcal{F}(\cdot)$ denotes the Fourier transform (FT) operator. Thanks to this penalty term, we can impose several effects which were previously not possible. First, $\mathbf{E}_{\text{latent}}$ imposes structure on the latent space by encouraging the sparsity of the Fourier series and removing less relevant frequency components. In other words, this can be interpreted as subspace constraints on the trajectory of the latent space variable, where the basis trajectories are periodic functions. Second, by analysing the structured latent space, we can extract the period of a periodic sequence and temporally segment the shapes according to their similarity. Our motivation for $\mathbf{E}_{\text{latent}}$ is manifold and partially comes from the observation that many real-world scenes evince recurrence, *i.e.,* they repeat their non-rigid states either in periodic or non-periodic manner.

Period Detection and Sequence Segmentation. The period of the sequence can be recovered from the estimated $\mathcal{F}(\mathbf{z})$, by extracting the dominant frequency in terms of energy within the frequency spectrum. If a dominant frequency ω_d is identified, its period can be directly computed as $\tau = \frac{T}{\omega_d}$. Unfortunately, in some real scenarios, the frequency spectrum that we obtain may not be unimodal (two or more relevant peaks can be observed in the spectrum), and therefore we obtain $\tau = T$. Irrespective whether a sequence is periodic or not, the latent space is temporally segmented so that similar values are decoded into similar shapes. This enables applications such as shape interpolation, completing and denoising.

4.3 Implementation Details

The proposed energy in Eq. (4) and the deformation auto-decoder $f_{\boldsymbol{\theta}}$ are fully-differentiable by construction, and therefore the gradients that flow into $\mathbf{S}_t$ can be further back-propagated into $\boldsymbol{\theta}$. Our deformation model is trained to simultaneously recover the motion parameters $\mathbf{R}$, the latent space $\mathbf{z}$ to encode shape deformations, and the model parameters $\boldsymbol{\theta}$. Additionally, the trajectory coefficients in $\mathbf{A}$ are also learned in this manner (see Eq. (8)). For initialisation, we use rigid factorisation to obtain $\mathbf{R}$ and $\bar{\mathbf{S}}$, random values in the interval $[-1, 1]$ for $\mathbf{z}$, and a null matrix for $\mathbf{A}$. The weights $\beta, \gamma, \eta, \omega, \lambda$ are determined empirically and selected from the determined ranges in most experiments we describe in Sec. 5, unless mentioned otherwise. The values we set are 10^2 for $\mathbf{E}_{\text{data}}$, $\beta = 1$, $\gamma \in [10^{-6}; 10^{-4}]$, $\eta \in [1; 10]$, $\omega = 1$, $\lambda \in [0; 10^{-3}]$ and $B = 32$ in $f_{\boldsymbol{\theta}}$. In addition, we use $K = 7$ as default value to define our low-rank trajectory model in Eq. (8).

Our N-NRSfM approach is implemented in pytorch [45]. As all the training data are available at the same time, we use the RProp optimiser [48] with a learning rate of 0.0001, and train for $60,000$ epochs. All experiments are performed on NVIDIA Tesla V100 and K80 GPUs with a Debian 9 Operating System. Depending on the size of the dataset, training takes between three (*e.g.*, the *back* sequence [50]) and twelve (the *barn-owl* sequence [26]) hours on our hardware.

4.4 Applications of the Deformation Auto-Decoder $f_{\boldsymbol{\theta}}$

Our deformation auto-decoder $f_{\boldsymbol{\theta}}$ can be used for several applications which were not easily possible in the context of NRSfM before, including shape denoising, shape completion and interpolation as well as correspondence-free monocular 3D reconstruction of non-rigid surfaces with reoccurring deformations.

Shape Compression, Interpolation, Denoising and Completion. The trained $f_{\boldsymbol{\theta}}$ combined with $\bar{\mathbf{S}}$ represents a compressed version of a 4D reconstruction and requires much less memory compared to the uncompressed shapes in the explicit representation $\mathbf{S}_t$ with $t = \{1, \dots, T\}$. The number of parameters required to capture all 3D deformations accurately depends on the complexity of the observed deformations, and not on the length of a sequence. Thus, the longer a sequence with repetitive states is, the higher is the compression ratio c. Next, let us suppose we are given a partial and noisy shape $\tilde{\mathbf{S}}$, and we would like to obtain a complete and smooth version of it $\mathbf{S}_{\boldsymbol{\theta}}$ upon the learned deformation model prior. We use our pre-trained auto-decoder and optimise for the latent code $\mathbf{z}$, using the per-vertex error as the loss. In the case of a partial shape, the unknown vertices are assumed to have some dummy values. Moreover, since the learned latent space is smooth and statistically assigns similar variables to similar shapes (displacements), we can interpolate the latent variables which will result in the smooth interpolation of the shapes (displacements).

Direct Monocular Non-Rigid 3D Reconstruction with Occlusion Handling. Pre-trained $f_{\boldsymbol{\theta}}$ can also be combined with other machine-learning components. We are interested in direct monocular non-rigid 3D reconstruction for endoscopic scenarios (though N-NRSfM is not restricted to those). Therefore, we train an image encoder which relates images to the resulting latent space of shapes (after the N-NRSfM training). Such image-to-mesh encoder-decoder is also robust against moderate partial scene occlusions — which frequently occur is endoscopic scenarios — as the deformations model $f_{\boldsymbol{\theta}}$ can also rely on partial observations. We build the image encoder based on ResNet-50 [32] pre-trained on the ImageNet [17] dataset.

At test time, we can reconstruct a surface from a single image, assuming state recurrence. Since the latent space is structured, we are modelling in-between states obtained by interpolation of the observed surfaces. This contrasts to the DSPR method [25], which *de facto* allows only state re-identification. Next, with the gradual degradation of the views, the accuracy of our image-to-surface reconstructor degrades gracefully. We can feed images with occlusions or a constant camera pose bias — such as those observed by changing from the left to the right camera in stereo recordings — and still expect accurate reconstructions.

5 Experiments

In this section, we describe the experimental results. We first compare our N-NRSfM approach to competing approaches on several widely-used benchmarks and real datasets following the established evaluation methodology for NRSfM

(Sec. 5.1). We next evaluate how accurately our method detects the periods and how well it segments sequences with non-periodic state recurrence (Sec. 5.2). For the sequences with 3D ground truth geometry $\mathbf{S}^{\mathrm{GT}}$, we report the 3D error e_{3D} — after shape-wise orthogonal Procrustes alignment — defined as $e_{3D} = \frac{1}{T}\sum_t \frac{\|\mathbf{S}_t^{\mathrm{GT}} - \mathbf{S}_t\|_{\mathcal{F}}}{\|\mathbf{S}_t^{\mathrm{GT}}\|_{\mathcal{F}}}$, where $\|\cdot\|_{\mathcal{F}}$ denoted Frobenius norm. Note that e_{3D} also implicitly evaluates the accuracy of $\mathbf{R}_t$ because of the mutual dependence between $\mathbf{R}_t$ and $\mathbf{S}_t$. Finally, for periodic sequences, we compare the estimated pulse τ with the known one τ^{GT}.

5.1 Quantitative Comparisons

We use three benchmark datasets in the quantitative comparison: *synthetic faces* (two sequences with 99 frames and two different camera trajectories denoted by *traj. A* and *traj. B*, with 28,000 points per frame) [19], *expressions* (384 frames with 997 points per frame) [4], and Kinect *t-shirt* (313 frames with 77,000 points) and *paper* (193 frames with 58,000 points) sequences taken from [64]. In the case if 3D ground truth shapes are available, ground truth dense point tracks are obtained by a virtual orthographic camera. Otherwise, dense correspondences are calculated by multi-frame optical flow [20,57].

Synthetic Faces. e_{3D} for the *synthetic faces* are reported in Table 1. We compare our N-NRSfM to Metric Projections (MP) [43], Trajectory Basis (TB) approach [7], Variational Approach (VA) [19], Dense Spatio-Temporal Approach (DSTA) [15], Coherent Depth Fields (CDF) [23], Consolidating Monocular Dynamic Reconstruction (CMDR) [24,25], Grassmannian Manifold (GM) [37], Jumping Manifolds (JM) [36], Scalable Monocular Surface Reconstruction (SMSR) [8], Expectation-Maximisation Finite Element Method (EM-FEM) [1] and Probabilistic Point Trajectory Approach (PPTA) [6]. Our N-NRSfM comes close to the most accurate methods on *traj. A* and comes in the middle on *traj. B* among all methods. Note that GM and JM use Procrustes alignment with scaling, which results in the comparison having slightly differing metrics. Still, we include these methods for completeness. *Traj. B* is reportedly more challenging compared to *traj. A* for all tested methods which we also confirm in our runs. We observed that without the depth control term in Eq. (7), the e_{3D} on *traj. B* was higher by $\sim$30%. Fig. 4-(a) displays the effect of Eq. (7) on the 3D reconstructions from real images, when the dense point tracks and initialisations can be noisy.

Expressions. The usage of *expressions* allows us to compare N-NRSfM to even more methods from the literature including Expectation-Maximisation Linear Dynamical System (EM-LDS) [61], Column Space Fitting, version 2 (CSF2) [29], Kernel Shape Trajectory Approach (KSTA) [28] and Global Model with Local Interpretation (GMLI) [4]. The results are summarised in Table 2. We achieve $e_{3D} = 0.026$ on par with GMLI, *i.e.,* currently the best method on this sequence. The complexity of facial deformations in the *expressions* is similar to those of the *synthetic faces* [19]. This experiment shows that our novel neural model for NRSfM with constraints in metric and trajectory space is superior to multiple older NRSfM methods.

Table 1: e_{3D} for the 99 frames long synthetic face sequence [19] (*traj. A* and *traj. B*). * denotes methods which use Procrustes analysis for shape alignment, whereas most methods use orthogonal Procrustes. † indicates sequential method. Compared to the default settings, the lowest e_{3D} of N-NRSfM is obtained with $B = 10$, $K = 30$, $\lambda = 0$ and $\eta = 10$ for *traj. A* (denoted by "♭") and $K = 40$ for *traj. B* (denoted by "♮").

	TB [7]	**MP** [43]	**VA** [19]	**DSTA** [15]	**CDF** [23]	**CMDR** [24]
traj. *A*	0.1252	0.0611	0.0346	0.0374	0.0886	0.0324
traj. *B*	0.1348	0.0762	0.0379	0.0428	0.0905	0.0369
	GM* [37]	**JM*** [36]	**SMSR** [8]	**PPTA** [6]	**EM-FEM** [1]†	**N-NRSfM (ours)**
traj. *A*	0.0294	0.0280	0.0304	0.0309	0.0389	0.045 / 0.032$^\flat$
traj. *B*	0.0309	0.0327	0.0319	0.0572	0.0304	0.049 / 0.0389$^\natural$

Table 2: Qualitative comparison on the *expressions* dataset [4].

	EM-LDS [61]	**PTA** [7]	**CSF2** [29]	**KSTA** [28]	**GMLI** [4]	**N-NRSfM (ours)**
Expr.	0.044	0.048	0.03	0.035	**0.026**	**0.026**

Table 3: Quantitative comparison on the Kinect *paper* and *t-shirt* sequences [64].

	TB [7]	**MP** [43]	**DSTA** [15]	**GM** [37]	**JM** [36]	**N-NRSfM (ours)**
paper	0.0918	0.0827	0.0612	0.0394	0.0338	**0.0332**
t-shirt	0.0712	0.0741	0.0636	0.0362	0.0386	**0.0309**

Kinect Sequences. For a fair evaluation, we pre-process the Kinect *t-shirt* and *paper* sequences along with their respective reference depth measurements as described in Kumar *et al.* [37]. As it is suggested there, we run multi-frame optical flow [20] with default parameters to obtain dense correspondences. e_{3D} for the Kinect sequences are listed in Table 3. Visualisations of selected reconstructions of Kinect sequences can be found in Fig. 6-(top row). On Kinect *paper* and *t-shirt* sequences, we outperform all competing methods, including the current state of the art by significant margins of 1% and 20%, respectively. These sequences evince larger deformations compared to the face sequence, and, on the other hand, a simpler camera trajectory.

5.2 Period Detection and Sequence Segmentation

We evaluate the capability of our N-NRSfM method in period detection and sequence segmentation on the *actor mocap* sequence (100 frames with $3.5 \cdot 10^4$ points per frame) [25,63]. It has highly deformed facial expressions with ground truth shapes, ground truth dense flow fields and rendered images under orthographic projection. We duplicate the sequence and run N-NRSfM on the obtained point tracks. Our approach reconstructs the entire sequence and returns the fre-

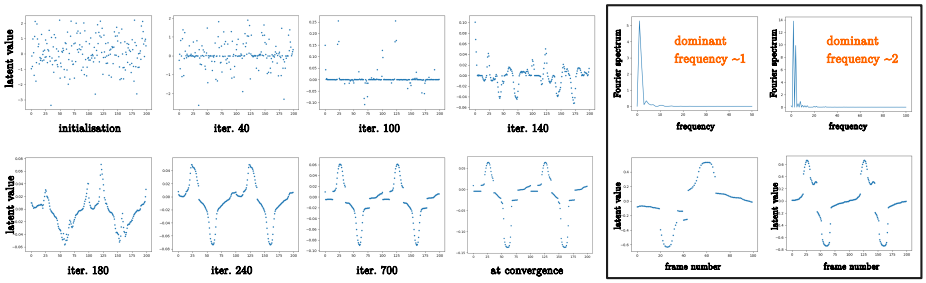

Fig. 3: Visualisations of the latent space during the training of N-NRSfM on *actor mocap* [25]. We show which effect our latent space constraints have on the latent space function. Left: The evolution of the latent space function from initialisation until convergence. Right: Frequency spectrum for the case with 100 and 200 frames.

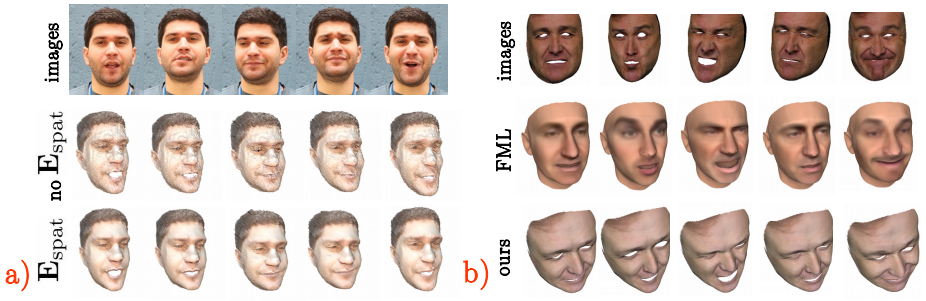

Fig. 4: a): 3D reconstructions of the *real face* with and without $\mathbf{E}_{\text{spat}}$. b): Input images of the *actor mocap* sequence; and 3D reconstructions by FML [58] and our approach.

quency equal to 2, as can be seen in the Fourier spectrum. Given 200 input frames, it implies a period of 100. The latent space function for this experiment and the evolution of the latent space function are shown in Fig. 3. Note that for the same shapes, the resulting latent variables are also similar. This confirms that our N-NRSfM segments the sequence based on the shape similarity.

Next, we confirm that the period detection works well on real heart bypass surgery sequence [56] with 201 frames and 68,000 point per frame (see Fig. 6-(bottom right) for the exemplary frames and our reconstructions). This sequence evinces natural periodicity, and the flow fields are computed individually for every frame without duplication. We emphasise that images do not repeat as — even though the states are recurrent — they are observed under varying illumination and different occlusions. We recover the dominant frequency of 7.035, whereas the observed number of heartbeats amounts to ∼7.2. Knowing that the video was recorded at 24 frames per second, we obtain the pulse τ of $\tau = 7.035 \text{ beats} \cdot \frac{24 \text{ fps}}{201 \text{ frames}} = 0.84$ beats per second or ∼50 beats per minute — which is in the expected pulse range of a human during bypass surgery.

5.3 Qualitative Results and Applications

The *actor mocap* sequence allows us to qualitatively compare N-NRSfM to a state-of-the-art method for monocular 3D face reconstruction. Thus, we run the Face Model Learning (FML) approach of Tewari *et al.* [58] on it and show qualitative results in Fig. 4-(b). We observe that it is difficult to recognise the person in the FML 3D estimates ($e_{3D} = 0.092$ after Procrustes alignment of the ground truth shapes and FML reconstructions with rescaling of the latter). Since FML runs per-frame, its 3D shapes evince variation going beyond changing facial expressions, *i.e.,* it changes the identity. In contrast, N-NRSfM produces recognizable and consistent shapes at the cost of accurate dense correspondences across an image batch ($e_{3D} = 0.0181$, $\sim$5 times lower compared to $e_{3D} = 0.092$ of FML).

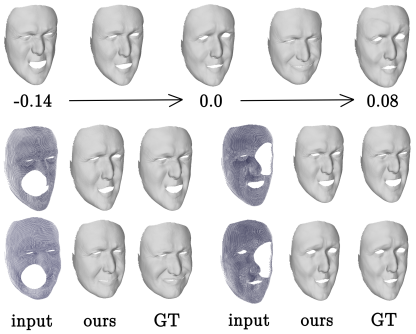

Fig. 5: Shape interpolation and completion. Top. A shape interpolation over the *actor* sequence is performed. **Bottom.** A series of three shapes are displayed, *i.e.,* input data, our estimation after completion, and the corresponding ground truth (GT).

Our auto-decoder $f_{\boldsymbol{\theta}}$ is a flexible building block which can be used in multiple applications which were not easily possible with classical NRSfM methods. Those include shape completion, denoising, compression and interpolation, fast direct monocular non-rigid 3D reconstruction as well as sequence segmentation.

Shape Interpolation and Completion. To obtain shape interpolations, we can linearly interpolate the latent variables, see Fig. 5-(top row) for an example with the *actor mocap* reconstructions. Note that the interpolation result depends on the shape order in the latent space. For shape with significantly differing latent variables, it is possible that the resulting interpolations will not be equivalent to linear interpolations between the shapes and include non-linear point trajectories. Results of shape denoising and completion are shown in Fig. 5-(bottom rows). We feed point clouds with missing areas (mouth and the upper head area) and obtain surfaces completed upon our learned $f_{\boldsymbol{\theta}}$ prior.

Direct Monocular Non-Rigid 3D Reconstruction. We attach an image encoder to $f_{\boldsymbol{\theta}}$ — as described in Sec. 4.4 — and test it in the endoscopic scenario with the *heart* sequence. Our reconstructions follow the cardiac cycle outside of the image sub-sequence, which has been used for the training. Please, see our supplemental material for extra visualisations.

Real Image Sequences. Finally, we reconstruct several real image sequence, *i.e., barn owl* [26], *back* [50] (see Fig. 6) and *real face* [19] (see Fig. 4-(a) which also highlights the influence of the spatial smoothness term). All our reconstructions are of high visual quality and match state of the art. Please, see our supplementary video for time-varying visualisations.

Fig. 6: Qualitative results on real sequences. In all cases, from left to right. **Top:** *T-shirt* and *paper* sequences [64]. **Bottom:** *Barn owl* [26], *back* [50] and *heart* [56] sequences. On both Kinect sequences, we achieve the lowest e_{3D} among all tested methods. The *heart* sequence is also used in the experiment with direct monocular non-rigid 3D reconstruction. For the visualisation of the *real face* sequence, see Fig. 4-(a).

6 Concluding Remarks

This paper introduces the first end-to-end trainable neural dense NRSfM method with a deformation model auto-decoder and learnable latent space function. Our approach operates on dense 2D point tracks without 3D supervision. Structuring the latent space to detect and exploit periodicity is a promising first step towards new regularisation techniques for NRSfM. Period detection and temporal segmentation of the reconstructed sequences, automatically learned deformation model, shape compression, completion and interpolation — all that is obtained with a single neural component in our formulation. Experiments have shown that the new model results in smooth and accurate surfaces while achieving low 3D reconstruction errors in a variety of scenarios. One of the limitations of N-NRSfM is the sensitivity to inaccurate points tracks and the dependence on the mean shape obtained by rigid initialisation. We also found that our method does not cope well with large and sudden changes, even though the mean shape is plausible. Another limitation is the handling of articulated motions.

We believe that our work opens a new perspective on dense NRSfM. In future research, more sophisticated neural components for deformation models can be tested to support stronger non-linear deformations and composite scenes. Moreover, we plan to generalise our model to sequential NRSfM scenarios.

Supplementary Material. Our supplementary material contains more experimental results (*e.g.,* with noisy point tracks), details of the experimental evaluations (*e.g.,* alignment of the FML reconstructions to the ground truth shapes) as well as two application examples of pre-trained shape auto-decoders.

References

1. Agudo, A., Montiel, J.M.M., Agapito, L., Calvo, B.: Online dense non-rigid 3D shape and camera motion recovery. In: British Machine Vision Conference (BMVC) (2014)
2. Agudo, A., Montiel, J.M.M., Calvo, B., Moreno-Noguer, F.: Mode-shape interpretation: Re-thinking modal space for recovering deformable shapes. In: Winter Conference on Applications of Computer Vision (WACV) (2016)
3. Agudo, A., Moreno-Noguer, F.: DUST: Dual union of spatio-temporal subspaces for monocular multiple object 3D reconstruction. In: Computer Vision and Pattern Recognition (CVPR) (2017)
4. Agudo, A., Moreno-Noguer, F.: Global model with local interpretation for dynamic shape reconstruction. In: Winter Conference on Applications of Computer Vision (WACV) (2017)
5. Agudo, A., Moreno-Noguer, F.: Force-based representation for non-rigid shape and elastic model estimation. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **40**(9), 2137–2150 (2018)
6. Agudo, A., Moreno-Noguer, F.: A scalable, efficient, and accurate solution to non-rigid structure from motion. Computer Vision and Image Understanding (CVIU) **167**, 121–133 (2018)
7. Akhter, I., Sheikh, Y., Khan, S., Kanade, T.: Trajectory space: A dual representation for nonrigid structure from motion. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **33**(7), 1442–1456 (2011)
8. Ansari, M., Golyanik, V., Stricker, D.: Scalable dense monocular surface reconstruction. In: International Conference on 3D Vision (3DV) (2017)
9. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. International Journal of Computer Vision (IJCV) **92**(1), 1–31 (2011)
10. Bartoli, A., Gay-Bellile, V., Castellani, U., Peyras, J., Olsen, S., Sayd, P.: Coarse-to-fine low-rank structure-from-motion. In: Computer Vision and Pattern Recognition (CVPR) (2008)
11. Bregler, C., Hertzmann, A., Biermann, H.: Recovering non-rigid 3D shape from image streams. In: Computer Vision and Pattern Recognition (CVPR) (2000)
12. Bue, A.D.: A factorization approach to structure from motion with shape priors. In: Computer Vision and Pattern Recognition (CVPR) (2008)
13. Choy, C., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In: European Conference on Computer Vision (ECCV) (2016)
14. Clevert, D., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). In: International Conference on Learning Representations (ICLR) (2016)
15. Dai, Y., Deng, H., He, M.: Dense non-rigid structure-from-motion made easy – a spatial-temporal smoothness based solution. In: International Conference on Image Processing (ICIP). pp. 4532–4536 (2017)
16. Dai, Y., Li, H., He, M.: Simple prior-free method for non-rigid structure-from-motion factorization. International Journal of Computer Vision (IJCV) **107**, 101–122 (2014)
17. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: Computer Vision and Pattern Recognition (CVPR) (2009)

18. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3D object reconstruction from a single image. In: Computer Vision and Pattern Recognition (CVPR) (2017)
19. Garg, R., Roussos, A., Agapito, L.: Dense variational reconstruction of non-rigid surfaces from monocular video. In: Computer Vision and Pattern Recognition (CVPR) (2013)
20. Garg, R., Roussos, A., Agapito, L.: A variational approach to video registration with subspace constraints. International Journal of Computer Vision (IJCV) **104**(3), 286–314 (2013)
21. Golyanik, V., Fetzer, T., Stricker, D.: Accurate 3D reconstruction of dynamic scenes from monocular image sequences with severe occlusions. In: Winter Conference on Applications of Computer Vision (WACV). pp. 282–291 (2017)
22. Golyanik, V., Stricker, D.: Dense batch non-rigid structure from motion in a second. In: Winter Conference on Applications of Computer Vision (WACV). pp. 254–263 (2017)
23. Golyanik, V., Fetzer, T., Stricker, D.: Introduction to coherent depth fields for dense monocular surface recovery. In: British Machine Vision Conference (BMVC) (2017)
24. Golyanik, V., Jonas, A., Stricker, D.: Consolidating segmentwise non-rigid structure from motion. In: Machine Vision Applications (MVA) (2019)
25. Golyanik, V., Jonas, A., Stricker, D., Theobalt, C.: Intrinsic Dynamic Shape Prior for Fast, Sequential and Dense Non-Rigid Structure from Motion with Detection of Temporally-Disjoint Rigidity. arXiv e-prints (2019)
26. Golyanik, V., Mathur, A.S., Stricker, D.: Nrsfm-flow: Recovering non-rigid scene flow from monocular image sequences. In: British Machine Vision Conference (BMVC) (2016)
27. Golyanik, V., Shimada, S., Varanasi, K., Stricker, D.: HDM-Net: Monocular non-rigid 3d reconstruction with learned deformation model. In: EuroVR (2018)
28. Gotardo, P.F.U., Martinez, A.M.: Kernel non-rigid structure from motion. In: International Conference on Computer Vision (ICCV). pp. 802–809 (2011)
29. Gotardo, P.F.U., Martinez, A.M.: Non-rigid structure from motion with complementary rank-3 spaces. In: Computer Vision and Pattern Recognition (CVPR). pp. 3065–3072 (2011)
30. Groueix, T., Fisher, M., Kim, V.G., Russell, B., Aubry, M.: AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In: Computer Vision and Pattern Recognition (CVPR) (2018)
31. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: International Conference on Computer Vision (ICCV). pp. 1026–1034 (2015)
32. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
33. Kanazawa, A., Tulsiani, S., Efros, A.A., Malik, J.: Learning category-specific mesh reconstruction from image collections. In: European Conference on Computer Vision (ECCV) (2018)
34. Kong, C., Lucey, S.: Deep non-rigid structure from motion. In: International Conference on Computer Vision (ICCV) (2019)
35. Kovalenko, O., Golyanik, V., Malik, J., Elhayek, A., Stricker, D.: Structure from Articulated Motion: Accurate and Stable Monocular 3D Reconstruction without Training Data. Sensors **19**(20) (2019)
36. Kumar, S.: Jumping manifolds: Geometry aware dense non-rigid structure from motion. In: Computer Vision and Pattern Recognition (CVPR) (2019)

37. Kumar, S., Cherian, A., Dai, Y., Li, H.: Scalable dense non-rigid structure-from-motion: A grassmannian perspective. In: Computer Vision and Pattern Recognition (CVPR) (2018)
38. Lee, M., Cho, J., Choi, C.H., Oh, S.: Procrustean normal distribution for non-rigid structure from motion. In: Computer Vision and Pattern Recognition (CVPR) (2013)
39. Lee, M., Choi, C.H., Oh, S.: A procrustean markov process for non-rigid structure recovery. In: Computer Vision and Pattern Recognition (CVPR) (2014)
40. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Computer Vision and Pattern Recognition (CVPR) (2019)
41. Novotny, D., Ravi, N., Graham, B., Neverova, N., Vedaldi, A.: C3DPO: Canonical 3D pose networks for non-rigid structure from motion. In: International Conference on Computer Vision (ICCV) (2019)
42. Ostlund, J., Varol, A., Fua, P.: Laplacian meshes for monocular 3D shape recovery. In: European Conference on Computer Vision (ECCV). pp. 412–425 (2012)
43. Paladini, M., Del Bue, A., Xavier, J., Agapito, L., Stosić, M., Dodig, M.: Optimal metric projections for deformable and articulated structure-from-motion. International Journal of Computer Vision (IJCV) **96**(2), 252–276 (2012)
44. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Computer Vision and Pattern Recognition (CVPR) (2019)
45. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
46. Pearson, K.: On lines and planes of closest fit to systems of points in space. Philosophical Magazine **2**, 559–572 (1901)
47. Pumarola, A., Agudo, A., Porzi, L., Sanfeliu, A., Lepetit, V., Moreno-Noguer, F.: Geometry-aware network for non-rigid shape prediction from a single view. In: Computer Vision and Pattern Recognition (CVPR) (2018)
48. Riedmiller, M., Braun, H.: A direct adaptive method for faster backpropagation learning: The rprop algorithm. In: International Conference on Neural Networks (ICNN). pp. 586–591 (1993)
49. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**, 533–536 (1986)
50. Russell, C., Fayad, J., Agapito, L.: Energy based multiple model fitting for non-rigid structure from motion. In: Computer Vision and Pattern Recognition (CVPR). pp. 3009–3016 (2011)
51. Russell, C., Fayad, J., Agapito, L.: Dense non-rigid structure from motion. In: 2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization Transmission (3DIMPVT) (2012)
52. Sahasrabudhe, M., Shu, Z., Bartrum, E., Alp Güler, R., Samaras, D., Kokkinos, I.: Lifting autoencoders: Unsupervised learning of a fully-disentangled 3D morphable model using deep non-rigid structure from motion. In: International Conference on Computer Vision Workshops (ICCVW) (2019)
53. Salzmann, M., Fua, P.: Reconstructing sharply folding surfaces: A convex formulation. In: Computer Vision and Pattern Recognition (CVPR). pp. 1054–1061 (2009)

54. Shimada, S., Golyanik, V., Theobalt, C., Stricker, D.: IsMo-GAN: Adversarial learning for monocular non-rigid 3D reconstruction. In: Computer Vision and Pattern Recognition Workshops (CVPRW) (2019)
55. Sorkine, O.: Laplacian mesh processing. In: Annual Conference of the European Association for Computer Graphics (Eurographics) (2005)
56. Stoyanov, D.: Stereoscopic scene flow for robotic assisted minimally invasive surgery. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 479–486 (2012)
57. Taetz, B., Bleser, G., Golyanik, V., Stricker, D.: Occlusion-aware video registration for highly non-rigid objects. In: Winter Conference on Applications of Computer Vision (WACV) (2016)
58. Tewari, A., Bernard, F., Garrido, P., Bharaj, G., Elgharib, M., Seidel, H., Pérez, P., Zollhöfer, M., Theobalt, C.: Fml: Face model learning from videos. In: Computer Vision and Pattern Recognition (CVPR) (2019)
59. Tewari, A., Zollöfer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Christian, T.: MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In: International Conference on Computer Vision (ICCV) (2017)
60. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: a factorization method. International Journal of Computer Vision (IJCV) **9**(2), 137–154 (1992)
61. Torresani, L., Hertzmann, A., Bregler, C.: Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **30**(5), 878–892 (2008)
62. Tsoli, A., Argyros, A.A.: Patch-based reconstruction of a textureless deformable 3D surface from a single rgb image. In: International Conference on Computer Vision Workshops (ICCVW) (2019)
63. Valgaerts, L., Wu, C., Bruhn, A., Seidel, H.P., Theobalt, C.: Lightweight binocular facial performance capture under uncontrolled lighting. ACM Transactions on Graphics (TOG) **31**(6), 187:1–187:11 (2012)
64. Varol, A., Salzmann, M., Fua, P., Urtasun, R.: A constrained latent variable model. In: Computer Vision and Pattern Recognition (CVPR) (2012)
65. Vicente, S., Agapito, L.: Soft inextensibility constraints for template-free non-rigid reconstruction. In: European Conference on Computer Vision. pp. 426–440 (2012)
66. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2mesh: Generating 3D mesh models from single RGB images. In: European Conference on Computer Vision (ECCV) (2018)
67. Xiao, J., Chai, J., Kanade, T.: A closed-form solution to non-rigid shape and motion recovery. In: European Conference on Computer Vision (ECCV) (2004)
68. Yu, R., Russell, C., Campbell, N.D.F., Agapito, L.: Direct, dense, and deformable: Template-based non-rigid 3D reconstruction from RGB video. In: International Conference on Computer Vision (ICCV) (2015)
69. Zhu, Y., Huang, D., Torre, F.D.L., Lucey, S.: Complex non-rigid motion 3D reconstruction by union of subspaces. In: Computer Vision and Pattern Recognition (CVPR). pp. 1542–1549 (2014)