HDM-Net: Monocular Non-Rigid 3D Reconstruction with Learned Deformation Model

Vladislav Golyanik[♭,♯], Soshi Shimada[♭,♯], Kiran Varanasi[♭], and Didier Stricker[♭,♯]

[♭]Augmented Vision, DFKI (https://av.dfki.de) [♯]University of Kaiserslautern

Abstract. Monocular dense 3D reconstruction of deformable objects is a hard ill-posed problem in computer vision. Current techniques either require dense correspondences and rely on motion and deformation cues, or assume a highly accurate reconstruction (referred to as a template) of at least a single frame given in advance and operate in the manner of non-rigid tracking. Accurate computation of dense point tracks often requires multiple frames and might be computationally expensive. Availability of a template is a very strong prior which restricts system operation to a pre-defined environment and scenarios. In this work, we propose a new hybrid approach for monocular non-rigid reconstruction which we call *Hybrid Deformation Model Network* (HDM-Net). In our approach, a deformation model is learned by a deep neural network, with a combination of domain-specific loss functions. We train the network with multiple states of a non-rigidly deforming structure with a known shape at rest. HDM-Net learns different reconstruction cues including texture-dependent surface deformations, shading and contours. We show generalisability of HDM-Net to states not presented in the training dataset, with unseen textures and under new illumination conditions. Experiments with noisy data and a comparison with other methods demonstrate the robustness and accuracy of the proposed approach and suggest possible application scenarios of the new technique in interventional diagnostics and augmented reality.

Keywords: Monocular non-rigid reconstruction · Hybrid deformation model · Deep neural network.

1 Introduction

The objective of monocular non-rigid 3D reconstruction (MNR) is the recovery of a time-varying geometry observed by a single moving camera. In the general case, none of the states is observed from multiple views, and at the same time, both the object and the camera move rigidly. This problem is highly ill-posed in the sense of Hadamard since multiple states can cause similar 2D observations. To obtain a reasonable solution, multiple additional priors about the scene, types of motions and deformations as well as camera trajectory are required. Application domains of MNR are numerous and include robotics, medical applications and visual communication systems. MNR also has a long history in augmented reality (AR), and multiple applications have been proposed over the last twenty years ranging from medical systems to communication and entertainment [15, 38].

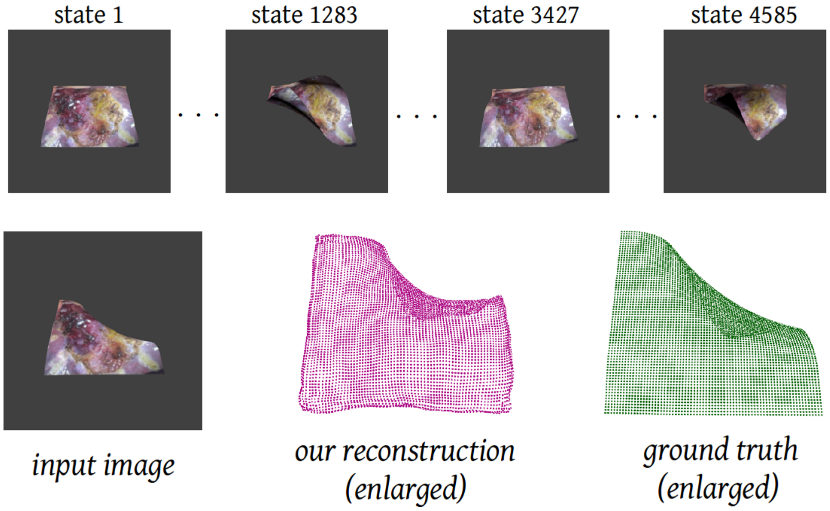


Fig. 1. Reconstruction of an endoscopically textured surface with the proposed HDM-Net. The network is trained on a textured synthetic image sequence with ground truth geometry and accurately reconstructs unseen views in a small fraction of a second (~5ms). Our architecture is potentially suitable for real-time augmented reality applications.

All approaches to MNR can be divided into two main model-based classes — non-rigid structure from motion (NRSfM) and template-based reconstruction (TBR). NRSfM relies on motion and deformation cues and requires dense point correspondences over multiple frames [26, 31]. Most accurate methods for dense correspondences operate on multiple frames and are prohibitively slow for real-time applications [66]. Moreover, their accuracy is volatile and influenced by changing illumination and shading effects in the scene. TBR, per definition, assumes a known template of the scene or an object, *i.e.*, a highly accurate reconstruction for at least one frame of the scene [55, 80]. Sometimes, the template also needs to be accurately positioned, with a minimal initial reprojection error to the reference frame. In this context, TBR can also be comprehended as non-rigid tracking [62]. Obtaining a template is beyond the scope of TBR, though joint solutions were demonstrated in the literature. In some cases, a template is obtained under the rigidity assumption, which might not always be fulfiled in practical applications [80].

Apart from the main classes, methods for monocular scene flow (MSF) and hybrid NRSfM can be named. MSF jointly reconstructs non-rigid geometry and 3D displacement fields [48]. In some cases, it relies on a known camera trajectory or proxy geometry (an initial coarse geometry estimate) [7]. In hybrid NRSfM, a scene-specific shape prior is obtained on-the-fly under non-rigidity, and the input is a sequence of point tracks [31]. Geometry estimation is then conditioned upon the shape prior.

MNR has only recently entered the realm of dense reconstructions [7, 58, 80]. The dense setting brings additional challenges for augmented reality applications such as scalability with the number of points and increased computational and memory complexity.

1.1 Contributions

The scope of this paper is general-purpose MNR, *i.e.*, the reconstruction scenarios are not known in advance. We propose deep neural network (DNN) based deformation model for MNR. We train DNN with a new synthetically generated dataset covering the variety of smooth and isometric deformations occurring in the real world (*e.g.*, clothes deformations, waving flags, bending paper and, to some extent, biological soft tissues). The proposed DNN architecture combines supervised learning with domain-specific loss functions. Our approach with a learned deformation model — Hybrid Deformation Model Network (HDM-Net) — surpasses performances of the evaluated state-of-the-art NRSfM and template-based methods by a considerable margin. We do not require dense point tracks or a well-positioned template. Our initialisation-free solution supports large deformations and copes well with several textures and illuminations. At the same time, it is robust to self-occlusions and noise. In contrast to existing DNN architectures for 3D, we directly regress 3D point clouds (surfaces) and depart from depth maps or volumetric representations.

In the context of MNR methods, our solution can be seen as a TBR with considerably relaxed initial conditions and a broader applicability range per single learned deformation model. Thus, it constitutes a new class of methods — instead of a template, we rather work with a weak shape prior and a shape at rest for a known scenario class.

We generate a new dataset which fills a gap for training DNNs for non-rigid scenes[1] and perform series of extensive tests and comparisons with state-of-the-art MNR methods. Fig. 1 provides an overview of the proposed approach — after training the network, we accurately infer 3D geometry of a deforming surface. Fig. 2 provides a high-level overview of the proposed architecture.

The rest of the paper is partitioned in Related Work (Sec. 2), Architecture of HDM-Net (Sec. 3), Geometry Regression and Comparisons (Sec. 5) and Concluding Remarks (Sec. 6) Sections.

2 Related Work

In this section, we review several algorithm classes and position the proposed HDM-Net among them.

2.1 Non-Rigid Structure from Motion

NRSfM requires coordinates of tracked points throughout an image sequence. The seminal work of Bregler *et al.* [10] marks the origin of batch NRSfM. It constrained surfaces to lie in a linear subspace of several unknown basis shapes. This idea was pursued by several successor methods [9, 51, 73]. Since the basis shapes, as well as their

[1] the dataset is available upon request.

number, are unknown, this subclass is sensitive to noise and parameter choice. Furthermore, an optimal number of basis shapes allowing to express all observed deformation modes does not necessarily always exist [73]. Along with that, multiple further priors were proposed for NRSfM including temporal smoothness [34, 82], basis [79], inextensibility [13, 22, 75] and shape prior [12, 31, 67], among others. The inextensibility constraint penalises deviations from configurations increasing the total surface area. In other words, non-dilatable states are preferred. Several methods investigate a dual trajectory basis and considerably reduce the number of unknowns [5], whereas the other ones explicitly model deformations using physical laws [4]. Multiple general-purpose unsupervised learning techniques were successfully applied to NRSfM including non-linear dimensionality reduction [67] (diffusion maps), [34, 37] (kernel trick) and expectation-maximisation [3, 43]. A milestone in NRSfM was accompanied by a further decrease in the number of unknowns and required prior knowledge for reconstruction. Thus, some of the methods perform a low-rank approximation of a stacked shape matrix [18,26]. A further milestone is associated with the ability to perform dense reconstructions [3, 6, 26, 31, 33].

Several methods allow sequential processing [1, 52, 82]. Starting from an initial estimate obtained on several first frames of a sequence, they perform reconstructions upon arrival of every new frame in an incremental manner. The accuracy of sequential methods is consistently lower than those of the batch counterparts. While still relying on point tracks, they can enable lowest latencies in real-time and interactive applications. Several methods learn and update an elastic model of the observed scene on-the-fly [2] (similarly to the sequential methods, point tracks over the complete sequence are not required). Solving the underlying equations might be slow, and the solution was demonstrated only for sparse settings.

2.2 Template-Based Reconstruction

Approaches of this class assume a known template, *i.e.*, an accurate reconstruction of at least one frame of the sequence. Most methods operate on a short window of frames or single frames. Some TBR methods are known as non-rigid trackers [62]. Early physics-based techniques formalised 3D reconstruction with elastic models and modal analysis [15, 47]. They assumed that some material properties (such as the elastic modulus) of the surface are known and could handle small non-linear deformations.

Multiple priors developed for NRSfM proved their effectiveness for TBR including isometry [11, 49, 55, 60], statistical priors [59], temporal smoothness [61, 80], inextensibility priors [11, 55] and mechanical priors in an improved form [38, 46]. Moreover, modelling image formation process by decomposing observed intensities into lighting, shading and albedo components was also shown to improve tracking accuracy [24, 25, 45, 49, 77].

2.3 Monocular Scene Flow

A somewhat exotic class of approaches developed in parallel to NRSfM and TBR is monocular scene flow (MSF). Birkbeck *et al.*'s approach can handle non-rigid scenes

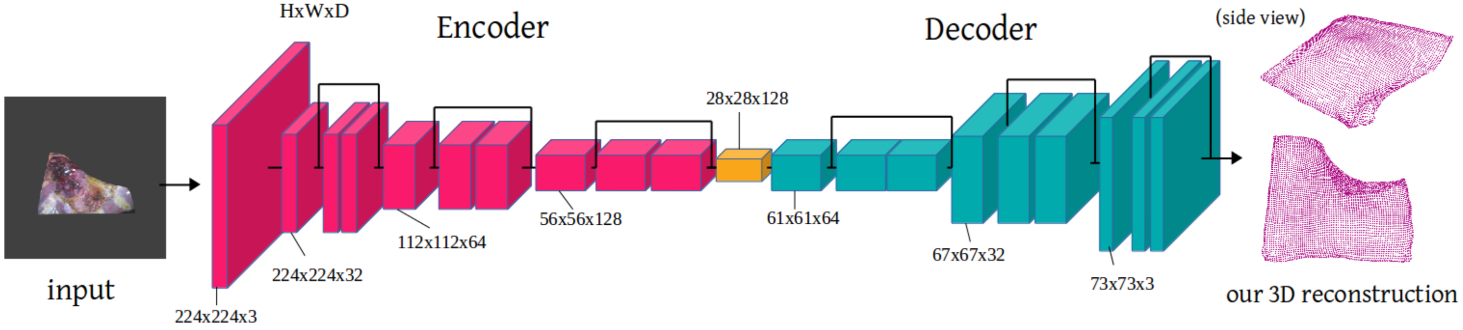


Fig. 2. An overview of the architecture of the proposed HDM-Net with encoder and decoder. The input of HDM-Net is an image of dimensions 224 × 224 with three channels, and the output is a dense reconstructed 3D surface of dimensions 73 × 73 × 3 (a point cloud with 73^2 points).

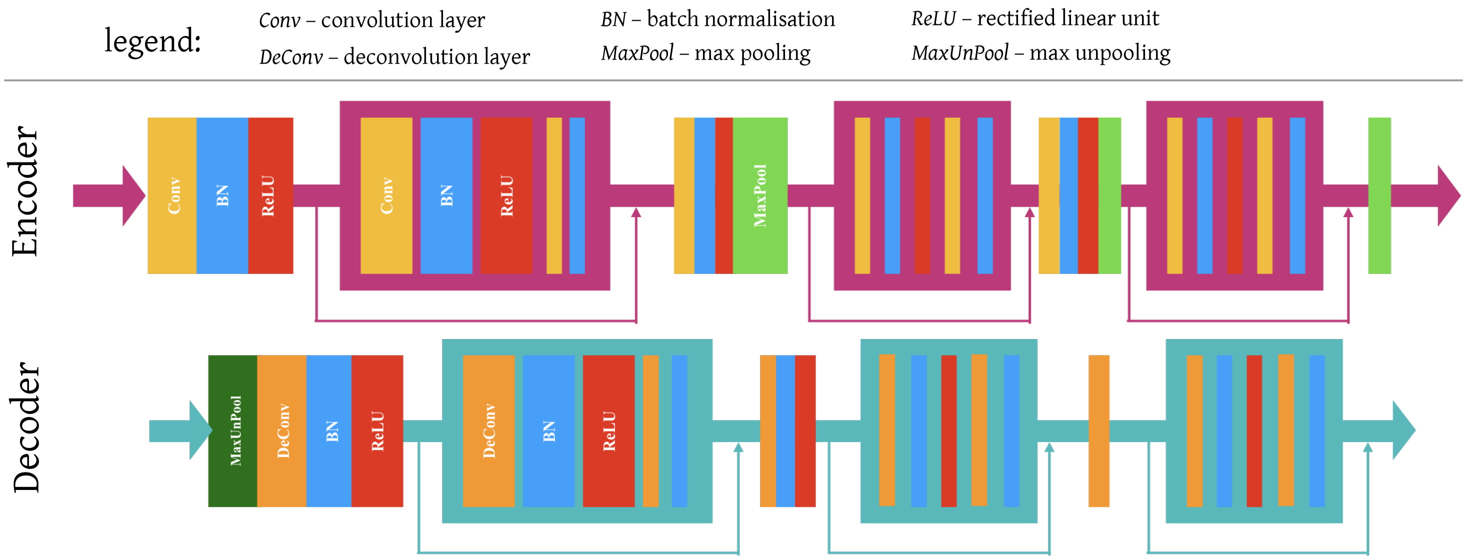

legend:
Conv – convolution layer
DeConv – deconvolution layer
BN – batch normalisation
MaxPool – max pooling
ReLU – rectified linear unit
MaxUnPool – max unpooling
Encoder
Conv
BN
ReLU
Conv
BN
ReLU
MaxPool
Decoder
MaxUnPool
DeConv
BN
ReLU
DeConv
BN
ReLU

Fig. 3. Architecture of the proposed HDM-Net: detailed clarification about the structures of the encoder and decoder.

relying on a known constant camera motion [7]. While camera trajectory can be sometimes available in AR systems, there is no guarantee of its linearity. In [48], a variational solution to rigid multi-body scenes was proposed. Recently, Xiao *et al.* proposed an energy-based method for rigid MSF in the context of automotive scenarios. Their approach is based on a temporal velocity constancy constraint [78].

In general, MSF methods are restricted in the handling of non-rigid surfaces. One exception — NRSfM-Flow of Golyanik *et al.* [32] — takes advantage of known 2D-3D correspondences and relies on batch NRSfM techniques for an accurate scene flow estimation of non-rigid scenes. It inherits the properties of NRSfM and does not assume a known camera trajectory or proxy geometry.

2.4 Specialised Models for Faces and Bodies

For completeness, we provide a concise overview of specialised approaches. Compared to TBR, they are dedicated to the reconstruction of single object classes like human faces [8, 28, 63, 65] or human bodies [35, 76]. They do not use a single prior state (a template), but a whole space of states with feasible deformations and variations. The models et al. are learned from extensive data collections showing a wide variety of forms, expressions (poses) and textures. In almost all cases, reconstruction with these methods means projection into the space of known shapes. To obtain accurate results, post-processing steps are required (*e.g.*, for transferring subtle details to the initial coarse estimates). In many applications, solutions with predefined models might be a right choice, and their accuracy and speed may be sufficient.

2.5 DNN-based 3D Reconstruction

In the recent three years, several promising approaches for inferring 2.5D and 3D geometry have been developed. Most of them regress depth maps [20, 27, 30, 44, 68] or use volumetric representations [14, 57] akin to sign distance fields [17]. Currently, the balance of DNN-based methods for 3D reconstruction is perhaps in favour of face regressors [19, 40, 63, 69]. The alternatives to sparse NRSfM of Tome *et al.* and Zhou *et al.* work exclusively for human poses [72, 81]. The 3D-R2N2 network generates 3D reconstructions from single and multiple views and requires large data collections for training [14]. In contrast to several other methods, it does not require image annotations. Point set generation netwoet al.rk of Fan *et al.* [21] is trained for a single view reconstruction of rigid objects and directly outputs point sets. More and more methods combine supervised learning and model-based losses thus imposing additional problem-specific constraints [21,27,69]. Also, this has often the side effect of decreasing the volume requirements on the datasets [30,69]. The work of Pumarola *et al.* [56] is most closely related to ours. The architecture is separated into three sub-networks which have different roles — creating heat-map of 2D images, depth estimation and 3D geometry inference. Those sub-networks are jointly trained. Our architecture is relatively simple. Encoder and decoder are employed and the output is penalized with three kinds of losses which have different geometrical properties — 3D geometry, smooth surface and contour information after projection onto a 2D plane.

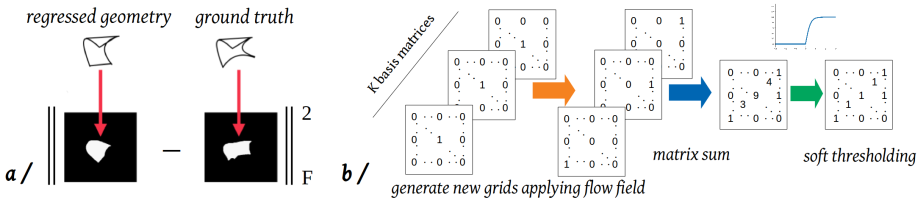


Fig. 4. Our contour loss penalises deviations between reprojections of the regressed geometry and reprojections of the ground truth.

2.6 Attributes of HDM-Net

In this section, we position the proposed approach among the vast body of the literature on MNR. HDM-Net bears a resemblance to DNN-based regressors which use encoder-decoder architecture [69]. In contrast to many DNN-based 3D regressors [14, 21, 69], our network does not include fully connected layers as they impede generalisability (lead to overfitting) as applied to MNR. As most 3D reconstruction approaches, it contains a 3D loss.

In many cases, isometry is an effective and realistic constraint for TBR, as shown in [13,55]. In HDM-Net, isometry is imposed through training data. The network learns the notion of isometry from the opposite, *i.e.*, by not observing other deformation modes. Another strong constraint in TBR is contour information which, however, has not found wide use in MSR, with only a few exceptions [36, 74]. In HDM-Net, we explicitly impose contour constraints by comparing projections of the learned and ground truth surfaces.

Under isometry, the solution space for a given contour is much better constrained compared to the extensible cases. The combined isometry and contour cues enable efficient occlusion handling in HDM-Net. Moreover, contours enable texture invariance up to a certain degree, as a contour remains unchanged irrespective of the texture. Next, through variation of light source positions, we train the network for the notion of shading. Since for every light source configuration, the underlying geometry is the same, HDM-Net acquires awareness of varying illumination. Besides, contours and shading in combination enable reconstruction of texture-less surfaces. To summarise, our framework has unique properties among MSR methods which are rarely found in other MNR techniques, especially when combined.

3 Architecture of HDM-Net

We propose a DNN architecture with encoder and decoder depicted in Fig. 2 (a general overview). The network takes as an input an image of dimensions 224×224 with three channels. Initially, the encoder extracts contour, shading and texture deformation cues and generates a compact latent space representation of dimensions $28 \times 28 \times 128$. Next, the decoder applies a series of deconvolutions and outputs a 3D surface of dimensions $73 \times 73 \times 3$ (a point cloud). It lifts the dimensionality of the latent space until the dimensionality of activation becomes identical to the dimensionality of ground truth. The transition from the implicit representation into 3D occurs on the later stage of decoder through a deconvolution. Fig. 3 provides a detailed clarification about the structures of encoder and decoder.

As can be seen in Fig. 2 and 3, we skip some connections in HDM-Net to avoid vanishing gradients, similar to *resnet* [39]. Due to the nature of convolutions, our deep network might potentially lose some important information in the forward path which might be advantageous in the deeper layers. Thus, connection skipping compensates for this side effect — for each convolution layer — which results in the increased performance. Moreover, in the backward path, shortcut connections help to overcome the vanishing gradient problem, *i.e.*, a series of numerically unstable gradient multiplications leading to vanishing gradients. Thus, the gradients are successfully passed to the shallow layers.

Fully connected (FC) layers are often used in classification tasks [42]. They have more parameters than convolution layers and are known as a frequent cause of overfitting. We have tried FC layers in HDM-Net and observed overfitting on the training dataset. Thus, FC layers reduce generalisation ability of our network. Furthermore, spatial information is destroyed as the data in the decoder is concatenated before being passed to the FC layer. In our task, needless to say, spatial cues are essential for 3D regression. In the end, we omit FC layers and successfully show generalisation ability of 3D reconstruction on the test data.

3.1 Loss Functions

Let $\mathbf{S} = \{\mathbf{S}_f\}$, $f \in \{1,\dots,F\}$ denote predicted 3D states, and $\mathbf{S}^{GT} = \{\mathbf{S}_f^{GT}\}$ is the ground truth geometry; F is the total number of frames and N is the number of points in the 3D surface. In HDM-Net, contour similarity and the isometry constraint are the key innovations and we apply three types of loss functions summarised into the loss energy:

$$\mathbf{E}(\mathbf{S},\mathbf{S}^{GT}) = \mathbf{E}_{3D}(\mathbf{S},\mathbf{S}^{GT}) + \mathbf{E}_{iso}(\mathbf{S}) + \mathbf{E}_{cont.}(\mathbf{S},\mathbf{S}^{GT}). \tag{1}$$

3D error: The 3D loss is the main loss in 3D regression. It penalises the differences between predicted and ground truth 3D states and is common in training for 3D data:

$$\mathbf{E}_{3D}(\mathbf{S},\mathbf{S}^{GT}) = \frac{1}{F}\sum_{f=1}^{F} \|\mathbf{S}_f^{GT} - \mathbf{S}_f\|_{\mathscr{F}}^2, \tag{2}$$

where $\|\cdot\|_{\mathscr{F}}$ denotes the Frobenius norm. Note that we take an average of the squared Frobenius norms of the differences between the learned and ground truth geometries.
Isometry prior: To additionally constrain the regression space, we embed isometry loss which enforces the neighbouring vertices to be located close to each other. Several versions of inextensibility and isometry constraints can be found in MSR — a common one is based on differences between Euclidean and geodesic distances. For our DNN architecture, we choose a differentiable loss which performs Gaussian smoothing of $\mathbf{S}_f$ and penalises the difference between the unembellished and smoothed version $\hat{\mathbf{S}}_i$:

$$\mathbf{E}_{iso}(\mathbf{S}) = \frac{1}{F}\sum_{f=1}^{F} \|\hat{\mathbf{S}}_f - \mathbf{S}_f\|_{\mathscr{F}}, \tag{3}$$

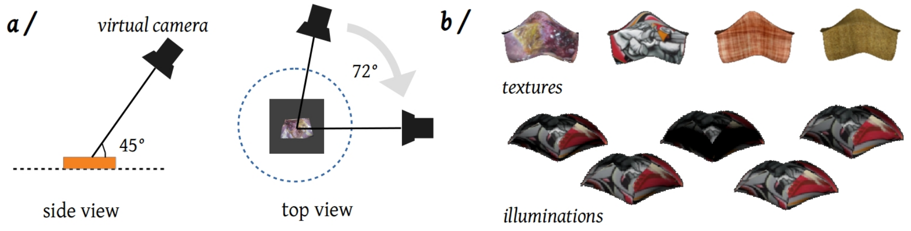


Fig. 5. Camera poses used for the dataset generation (a); different textures applied to the dataset: *endoscopy*, *graffiti*, *clothes* and *carpet* (b-top) and different illuminations (b-bottom).

with

$$\hat{\mathbf{S}}_f = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) * \mathbf{S}_f, \tag{4}$$

where $*$ denotes a convolution operator and σ^2 is the variance of Gaussian.

Contour loss: If the output of the network and the ground truth coordinates are similar, the contour shapes after projection onto a 2D plane have to be similar as well. The main idea of the reprojection loss is visualised in Fig. 4-(a). After the inference of the 3D coordinates by the network, we project them onto the 2D plane and compute the difference between the two projected contours. If focal lengths f_x, f_y as well as the principal point (c_x, c_y) of the camera are known (the $\mathbf{K}$ used for the dataset generation is provided in Sec. 4), observed 3D points $\mathbf{p} = (p_x, p_y, p_z)$ are projected to the image plane by the projection operator $\pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2$:

$$\mathbf{p}'(u,v) = \pi(\mathbf{p}) = \left(f_x \frac{p_x}{p_z} + c_x, f_y \frac{p_y}{p_z} + c_y\right)^{\mathsf{T}}, \tag{5}$$

where $\mathbf{p}'$ is the 2D projection of $\mathbf{p}$ with 2D coordinates u and v. Otherwise, we apply an orthographic camera model.

A naïve shadow casting of a 3D point cloud onto a 2D plane is not differentiable, *i.e.*, the network cannot backpropagate gradients to update the network parameters. The reason is twofold. In particular, the cause for indifferentiability is the transition from point intensities to binary shadow indicators with an ordinary step function (the numerical reason) using point coordinates as indexes on the image grid (the framework-related reason).

Fig. 4-(b) shows how we circumvent this problem. The first step of the procedure is the projection of 3D coordinates onto a 2D plane using either a perspective or an orthographic projection. As a result of this step, we obtain a set of 2D points. We generate $K = 73^2$ translation matrices $\mathbf{T}_j = \begin{pmatrix} 1 & 0 & u \\ 0 & 1 & v \end{pmatrix}$ using 2D points and a flow field tensor of dimension $K \times 99 \times 99 \times 2$ (the size of each binary image is 99×99). Next, we apply bilinear interpolation [41] with generated flow fields on the replicated basis matrix $\mathbf{B}$ K times and obtain K translation indicators. $\mathbf{B}_{99\times 99}$ is a sparse matrix with only a single central non-zero element which equals to 1. Finally, we sum up all translation indicators and softly threshold positive values in the sums to ≈ 1, *i.e.*, our shadow indicator.

Note that to avoid indifferentiability in the last step, the thresholding is performed by a combination of a rectified linear unit (ReLU) and tanh function (see Fig. 4-(b)):

$$\tau(\mathscr{I}(\mathbf{s}_f(n))) = \max(\tanh(2\mathbf{S}_f(n)), 0), \tag{6}$$

where $n \in \{1, \ldots, N\}$ denotes the point index, $\mathbf{s}_f(n)$ denotes a reprojected point $\mathbf{S}_f(n)$ in frame f, and $\mathscr{I}(\cdot)$ fetches intensity of a given point. We denote the differentiable projection operator and differentiable soft thresholding operator by the symbols $\pi^\dagger(\cdot)$ and $\tau(\cdot)$ respectively. Finally, the contour loss reads

$$\mathbf{E}_{cont.}(\mathbf{S}, \mathbf{S}^{GT}) = \frac{1}{F}\sum_{f=1}^{F} \|\tau\big(\pi^\dagger(\mathbf{S}_f)\big) - \tau\big(\pi^\dagger(\mathbf{S}_f^{GT})\big)\|^2_{\mathscr{F}}. \tag{7}$$

Note that object contours correspond to 0-1 transitions.

4 Dataset and Training

For our study, we generated a dataset with a non-rigidly deforming object using *Blender* [23]. In total, there are 4648 different temporally smooth 3D deformation states with structure bendings, smooth foldings and wavings, rendered under Cook-Torrance illumination model [16] (see Fig. 1 for the exemplary frames from our dataset). We have applied five different camera poses, five different light source positions and four different textures corresponding to the scenarios we are interested in — *endoscopy*, *graffiti* (it resembles a waving flag) *clothes* and *carpet* (an example of an arbitrary texture). The endoscopic texture is taken from [29]. Illuminations are generated based on the scheme in Fig. 5-(a), the textures and illuminations are shown in Fig. 5-(b). We project the generated 3D scene by a virtual camera onto a 2D plane upon Eq. (5), with $\mathbf{K} = \begin{pmatrix} 280 & 0 & 128 \\ 0 & 497.7 & 128 \\ 0 & 0 & 1 \end{pmatrix}$. The background in every image is of the same opaque colour. We split the data into training and test subsets in a repetitive manner, see Fig. 6 for the pattern. We train HDM-Net jointly on several textures and illuminations, with the purpose of illumination-invariant and texture-invariant regression. One illumination and one texture are reserved for the test dataset exclusively. Our images are of the dimensions

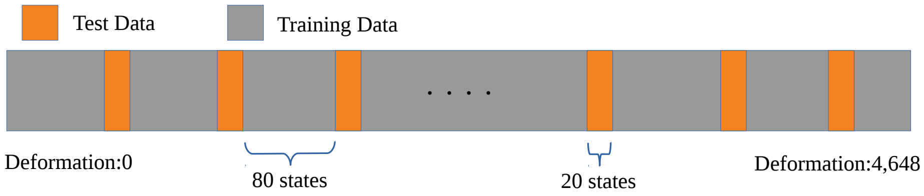


Fig. 6. The pattern of the training and test datasets.

256×256. They reside in 15.2 Gb of memory, and the ground truth geometry requires 1.2 Gb (in total, 16.4 Gb). The hardware configuration consists of two six-core processors Intel(R) Xeon(R) CPU E5-1650 v4 running at 3.60GHz, 16 GB RAM and a GEFORCE GTX 1080Ti GPU with 11GB of global memory. In total, we train for 95 epochs, and the training takes two days in *pytorch* [53, 54]. The evolution of the loss energy is visualised in Fig. 11-(a). The inference of one state takes ca. 5 ms.

5 Geometry Regression and Comparisons

We compare our method with the template-based reconstruction of Yu *et al.* [80], variational NRSfM approach (VA) of Garg *et al.* [26] and NRSfM method of Golyanik *et al.* [33] — Accelerated Metric Projections (AMP). We use an optimised heterogeneous CPU-GPU version of VA written in C++ and CUDA C [50]. AMP is a C++ CPU version which relies on an efficient solution of a semi-definite programming problem and is currently one of the fastest batch NRSfM methods. For VA and AMP, we compute required dense point tracks. Following the standard praxis in NRSfM, we project the ground truth shapes onto a virtual image plane by a slowly moving virtual camera. Camera rotations are parametrised by Euler angles around the x-, y- and z-axes. We rotate for up to 20 degrees around each axis, with five degrees per frame. This variety in motion yields minimal depth changes required for an accurate initialisation in NRSfM. We report runtimes, 3D error

$$e_{3D} = \frac{1}{F} \sum_{f=1}^{F} \frac{\|\mathbf{S}_f^{GT} - \mathbf{S}_f\|_{\mathscr{F}}}{\|\mathbf{S}_f^{GT}\|_{\mathscr{F}}} \tag{8}$$

and standard deviation σ of e_{3D}. Before computing e_{3D}, we align $\mathbf{S}_f$ and the corresponding $\mathbf{S}_f^{GT}$ with Procrustes analysis.

Runtimes, e_{3D} and σ for all three methods are summarised in Table 1. AMP achieves around 30 *fps* and can execute only for 100 frames per batch at a time. However, this estimate does not include often prohibitive computation time of dense correspondences with multi-frame optical flow methods such as [66]. Note that runtime of batch NRSfM depends on the batch size, and the batch size influences the accuracy and ability to reconstruct. VA takes advantage of a GPU and executes with 2.5 *fps*. Yu *et al.* [80] achieves around 0.3 *fps*. In contrast, HDM-Net processes one frame in only 5 ms. This is by far faster than the compared methods. Thus, HDM-Net can compete in runtime with rigid structure from motion [71]. The runtime of the latter method is still considered as the lower runtime bound for NRSfM[2].

At the same time, the accuracy of HDM-Net is the highest among all tested methods. Selected results with complex deformations are shown in Fig. 7. We see that Yu *et al.* [80] copes well with rather small deformations, and our approach accurately resolves even challenging cases not exposed during the training. In the case of Yu *et al.* [80], the

[2] when executed in a batch of 100 frames with 73^2 points each, a C++ version of [71] takes 1.47 ms per frame on our hardware; for 400 frames long batch, it requires 5.27 ms per frame.

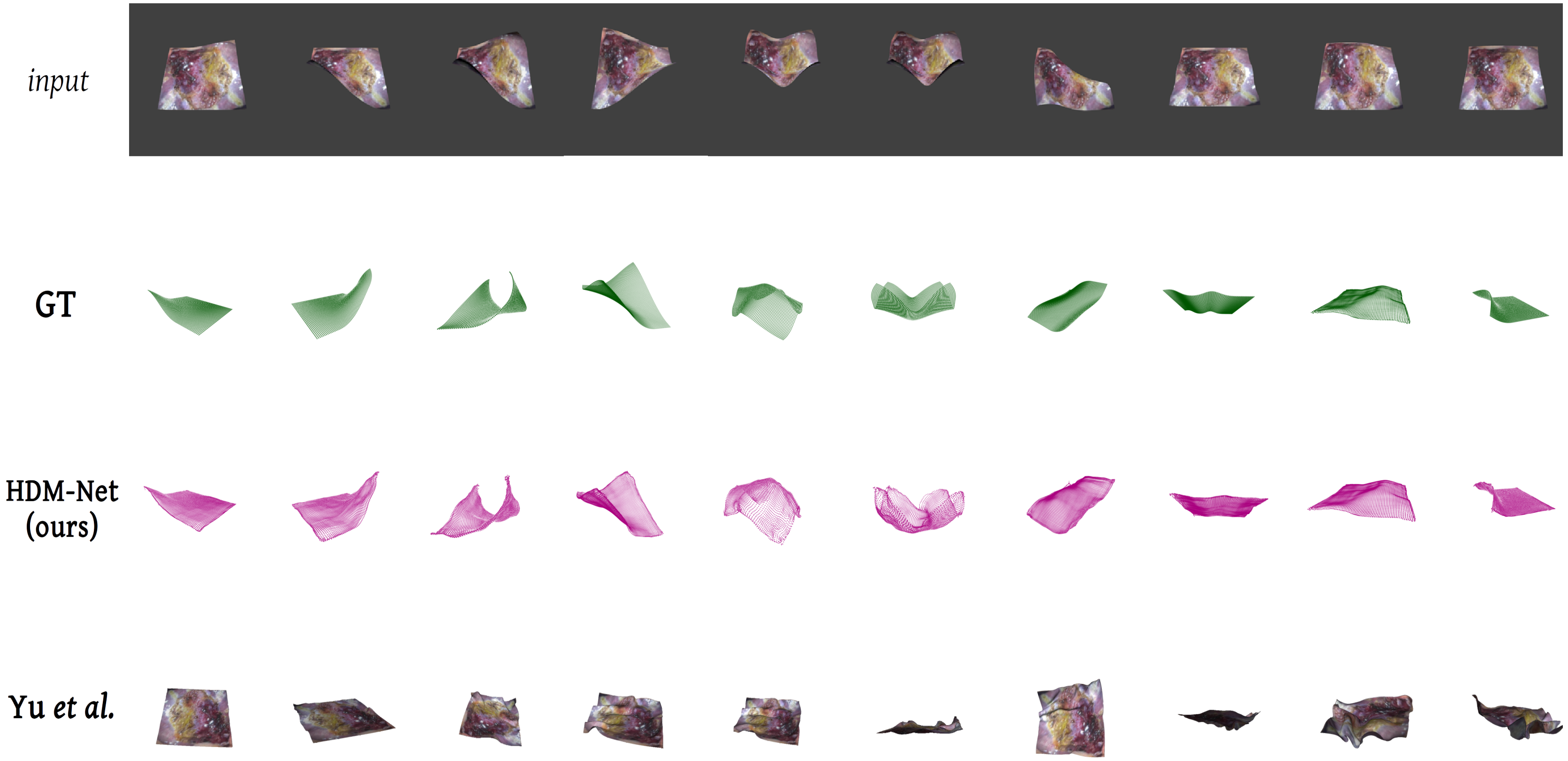


Fig. 7. Selected reconstruction results on endoscopically textured surfaces for HDM-Net (our method) and Yu *et al*. [80].

	Yu *et al.* [80]	AMP [33]	VA [26]	HDM-Net
t, s	3.305	0.035	0.39	**0.005**
e_{3D}	1.3258	1.6189	0.46	**0.0251**
σ	**0.0077**	1.23	0.0334	0.03

Table 1. Per-frame runtime t in *seconds*, e_{3D} and σ comparisons of Yu *et al.* [80], AMP [33] and HDM-Net (proposed method).

	endoscopy	*graffiti*	*clothes*	*carpet*
e_{3D}	**0.0485**	0.0499	0.0489	0.1442
σ	**0.01356**	0.022	0.02648	0.02694

Table 2. Comparison of 3D error for different textures and the same illumination (number 1).

	illum. 1	*illum. 2*	*illum. 3*	*illum. 4*	*illum. 5*
e_{3D}	0.07952	0.0801	0.07942	**0.07845**	**0.07827**
σ	**0.0525**	0.0742	0.0888	0.1009	0.1123

Table 3. Comparison of 3D error for different illuminations.

	3D	3D + Con.	3D + Iso.	3D + Con. + Iso.
e_{3D}	0.0698	**0.0688**	0.0784	0.0773
σ	**0.0761**	0.0736	0.0784	0.0789

Table 4. Comparison of effects of loss functions.

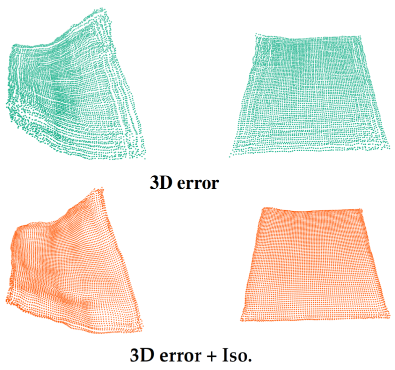


Fig. 8. Comparison of 3D reconstruction with 3D error (top row) and 3D error + isometry prior (bottom row)

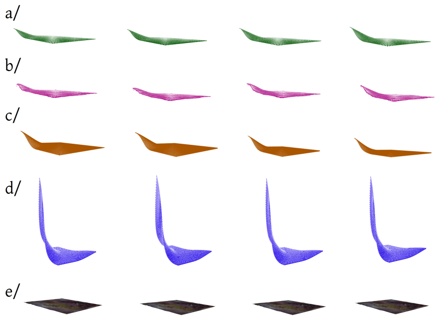


Fig. 9. Qualitative comparisons of ground truth (a), HDM-Net (proposed method) (b), AMP [33] (c), VA [26] (d) and Yu *et al.* [80] (e) on several frames of our test sequence from the first 100 frames (each column corresponds to one frame).

high e_{3D} is explained by a weak handling of self-occlusions and large deformations. In the case of NRSfM methods, the reason for the high e_{3D} is an inaccurate initialisation. Moreover, VA does not handle foldings and large deformations well.

Table 3 summarises e_{3D} for our method under different illumination conditions. We notice that our network copes well with all generated illuminations — the difference in e_{3D} is under 3%. Table 2 shows e_{3D} comparison for different textures. Here, the accuracy of HDM-Net drops on the previously unseen texture by the factor of three, which still corresponds to reasonable reconstructions with the captured main deformation mode. Another quantitative comparison is shown in Fig. 9. In this example, all methods execute on the first 100 frames of the sequence. AMP [33] captures the main deformation mode with $e_{3D} = 0.1564$ but struggles to perform a fine-grained distinction (in Table 1, e_{3D} is reported over the sequence of 400 frames, hence the differing met-

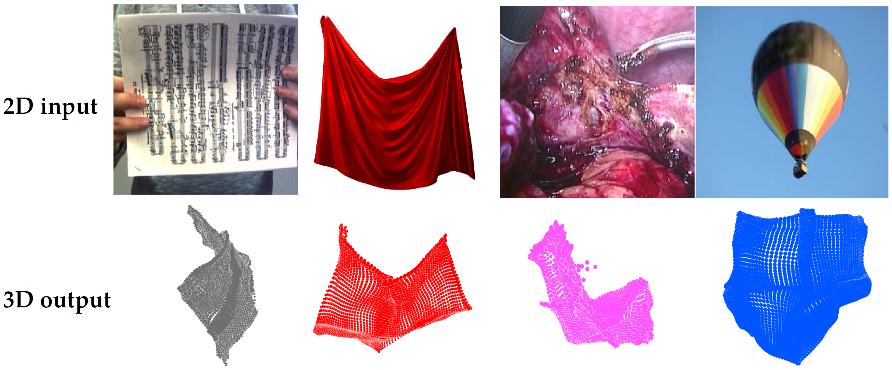


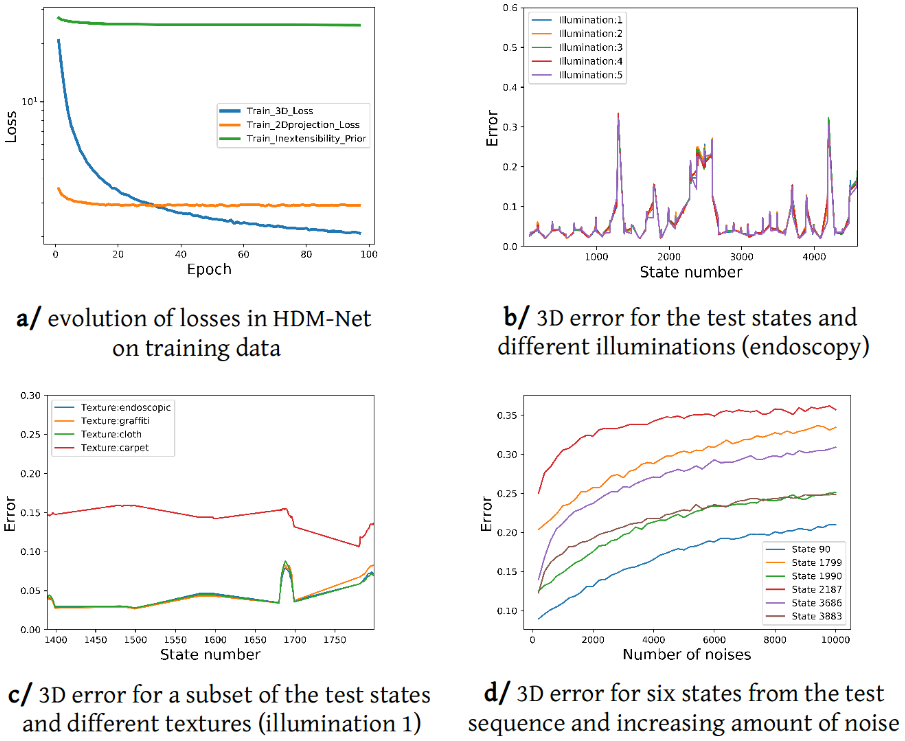
Fig. 10. Exemplary reconstructions from real images obtained by HDM-Net (music notes, a fabric, surgery and an air balloon)

rics). VA suffers under an inaccurate initialisation under rigidity assumption and Yu *et al*. [80], by contrast, does not recognise the variations in the structure. All in all, HDM-Net copes well with self-occlusions. Graphs of e_{3D} as functions of the state index under varying illuminations and textures can be found in Fig. 11-(b,c). Table 4 shows the comparison of e_{3D} using networks trained with various combinations of loss functions. *3D + Con.* shows the lowest e_{3D} and applying *isometry prior* increases e_{3D}. Since *isometry prior* is smoothing loss, the 3D grid becomes smaller in comparison to the outputs without *isometry prior* hence higher e_{3D}. However, as shown in Fig. 8, isometry prior allows the network to generate smoother 3D geometries preserving deformation states.

Next, we evaluate the performance of HDM-Net on noisy input images. Therefore, we augment the dataset with increasing amounts of uniform salt-pepper noise. Fig. 11-(d) shows the evolution of the e_{3D} as a function of the amount of noise, for several exemplary frames corresponding to different input difficulties for the network. We observe that HDM-Net is well-posed w.r.t noise — starting from the respective values obtained for the noiseless images, the e_{3D} increases gradually.

We tested HDM-Net on several challenging real images. Fig. 10 shows the tested images and our reconstructions. We recorded a music note image for an evaluation of our network in real-world scenario. Despite different origin of the inputs (music notes, a fabric [70], an endoscopic view during a surgery [29] and an air balloon [64]), HDM-Net produces realistic and plausible results. Note how different are the regressed geometries which suggests the generalisation ability of the proposed solution.

In many real-world cases, HDM-Net produces acceptable results. However, if the observed states differ a lot from the states in the training data, HDM-Net might fail to recognise and regress the state. This can be addressed by an extension or tailoring of the data set for specific cases. Adding training data originating from motion and geometry capture of real objects might also be an option.



a/ evolution of losses in HDM-Net on training data

b/ 3D error for the test states and different illuminations (endoscopy)

c/ 3D error for a subset of the test states and different textures (illumination 1)

d/ 3D error for six states from the test sequence and increasing amount of noise

Fig. 11. Graphs of e_{3D} for varying illuminations (for *endoscopy* texture), varying textures (for illumination 1) as well as six states under increasing amount of noise. Note that in b/ and c/, only the errors obtained on the test data are plotted. For c/, HDM-Net was trained on a subset of training states (three main textures and one illumination).

6 Concluding Remarks

We have presented a new monocular surface recovery method with a deformation model replaced by a DNN — HDM-Net. The new method reconstructs time-varying geometry from a single image and is robust to self-occlusions, changing illumination and varying texture. Our DNN architecture consists of an encoder, a latent space and a decoder, and is furnished with three domain-specific losses. Apart from the conventional 3D data loss, we propose isometry and reprojection losses. We train HDM-Net with a newly generated dataset with ca. four an a half thousands states, four different illuminations, five different camera poses and three different textures. Experimental results show the validity of our approach and its suitability for reconstruction of small and moderate isometric deformations under self-occlusions. Comparisons with one template-based and two template-free methods have demonstrated a higher accuracy in favour of HDM-Net. Since HDM-Net is one of the first approach of the new kind, there are multiple avenues for investigations and improvements. One apparent direction is the further augmentation of the test dataset with different backgrounds, textures and illuminations. Next, we are going to test more advanced architectures such as generative adversarial networks and recurrent connections for the enhanced temporal smoothness. Currently, we are also investigating the relevance of HDM-Net for medical applications with augmentation of soft biological tissues.

Acknowledgement

Development of HDM-Net was supported by the project DYMANICS (01IW15003) of the German Federal Ministry of Education and Research (BMBF). The authors thank NVIDIA Corporation for the hardware donations.

References

1. Agudo, A., Agapito, L., Calvo, B., Montiel, J.M.M.: Good vibrations: A modal analysis approach for sequential non-rigid structure from motion. In: Computer Vision and Pattern Recognition (CVPR). pp. 1558–1565 (2014)
2. Agudo, A., Moreno-Noguer, F.: Force-based representation for non-rigid shape and elastic model estimation. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2018)
3. Agudo, A., Moreno-Noguer, F.: A scalable, efficient, and accurate solution to non-rigid structure from motion. Computer Vision and Image Understanding (CVIU) (2018)
4. Agudo, A., Moreno-Noguer, F., Calvo, B., Montiel, J.M.M.: Sequential non-rigid structure from motion using physical priors. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2016)
5. 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)
6. Ansari, M., Golyanik, V., Stricker, D.: Scalable dense monocular surface reconstruction. In: International Conference on 3D Vision (3DV) (2017)
7. Birkbeck, N., Cobza, D., Jägersand, M.: Basis constrained 3d scene flow on a dynamic proxy. In: International Conference on Computer Vision (ICCV). pp. 1967–1974 (2011)
8. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: ACM Trans. Graphics (TOG). pp. 187–194 (1999)
9. Brand, M.: A direct method for 3d factorization of nonrigid motion observed in 2d. In: Computer Vision and Pattern Recognition (CVPR). vol. 2, pp. 122–128 (2005)
10. Bregler, C., Hertzmann, A., Biermann, H.: Recovering non-rigid 3d shape from image streams. In: Computer Vision and Pattern Recognition (CVPR). pp. 690–696 (2000)
11. Brunet, F., Hartley, R., Bartoli, A., Navab, N., Malgouyres, R.: Monocular template-based reconstruction of smooth and inextensible surfaces. In: Asian Conference on Computer Vision (ACCV). pp. 52–66 (2010)
12. Bue, A.D.: A factorization approach to structure from motion with shape priors. In: Computer Vision and Pattern Recognition (CVPR) (2008)
13. Chhatkuli, A., Pizarro, D., Collins, T., Bartoli, A.: Inextensible non-rigid structure-from-motion by second-order cone programming. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **PP**(99) (2018)

14. Choy, C.B., 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)
15. Cohen, L.D., Cohen, I.: Deformable models for 3-d medical images using finite elements and balloons. In: Computer Vision and Pattern Recognition (CVPR). pp. 592–598 (1992)
16. Cook, R.L., Torrance, K.E.: A reflectance model for computer graphics. ACM Trans. Graph. (TOG) **1**(1), 7–24 (1982)
17. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: ACM Trans. Graphics (TOG). pp. 303–312 (1996)
18. Dai, Y., Li, H., He, M.: A simple prior-free method for non-rigid structure-from-motion factorization. International Journal of Computer Vision **107**(2), 101–122 (2014)
19. Dou, P., Shah, S.K., Kakadiaris, I.A.: End-to-end 3d face reconstruction with deep neural networks. In: Computer Vision and Pattern Recognition (CVPR) (2017)
20. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in Neural Information Processing Systems (NIPS), pp. 2366–2374 (2014)
21. 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)
22. Fayad, J., Agapito, L., Del Bue, A.: Piecewise quadratic reconstruction of non-rigid surfaces from monocular sequences. In: European Conference on Computer Vision (ECCV) (2010)
23. Foundation., B.: blender, v. 2.79a. open source 3d creation. https://www.blender.org/ (2018)
24. Gallardo, M., Collins, T., Bartoli, A.: Using shading and a 3d template to reconstruct complex surface deformations. In: British Machine Vision Conference (BMVC) (2016)
25. Gallardo, M., Collins, T., Bartoli, A.: Dense non-rigid structure-from-motion and shading with unknown albedos. In: International Conference on Computer Vision (ICCV) (2017)
26. Garg, R., Roussos, A., Agapito, L.: Dense variational reconstruction of non-rigid surfaces from monocular video. In: Computer Vision and Pattern Recognition (CVPR). pp. 1272–1279 (2013)
27. Garg, R., Kumar, V.B.G., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: European Conference on Computer Vision (ECCV). pp. 740–756 (2016)
28. Garrido, P., Zollhöfer, M., Casas, D., Valgaerts, L., Varanasi, K., Perez, P., Theobalt, C.: Reconstruction of personalized 3d face rigs from monocular video **35**(3), 28:1–28:15 (2016)
29. Giannarou, S., Visentini-Scarzanella, M., Yang, G.Z.: Probabilistic tracking of affine-invariant anisotropic regions. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **35**(1), 130–143 (2013)
30. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Computer Vision and Pattern Recognition (CVPR) (2017)
31. 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) (2017)
32. 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)
33. 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)
34. 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)
35. Guan, P., Weiss, A., Blan, A.O., Black, M.J.: Estimating human shape and pose from a single image. In: International Conference on Computer Vision (ICCV). pp. 1381–1388 (2009)

36. Gumerov, N., Zandifar, A., Duraiswami, R., Davis, L.S.: Structure of applicable surfaces from single views. In: European Conference on Computer Vision (ECCV). pp. 482–496 (2004)
37. Hamsici, O.C., Gotardo, P.F.U., Martinez, A.M.: Learning spatially-smooth mappings in non-rigid structure from motion. In: European Conference on Computer Vision (ECCV) (2012)
38. Haouchine, N., Dequidt, J., Berger, M.O., Cotin, S.: Single view augmentation of 3d elastic objects. In: International Symposium on Mixed and Augmented Reality (ISMAR). pp. 229–236 (2014)
39. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition (CVPR) (2016)
40. Jackson, A.S., Bulat, A., Argyriou, V., Tzimiropoulos, G.: Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In: International Conference on Computer Vision (ICCV) (2017)
41. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Advances in Neural Information Processing Systems (NIPS). pp. 2017–2025 (2015)
42. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105 (2012)
43. Lee, M., Cho, J., Oh, S.: Procrustean normal distribution for non-rigid structure from motion. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **39**(7), 1388–1400 (2017)
44. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: Computer Vision and Pattern Recognition (CVPR) (2015)
45. Liu-Yin, Q., Yu, R., Agapito, L., Fitzgibbon, A., Russell, C.: Better together: Joint reasoning for non-rigid 3d reconstruction with specularities and shading. In: British Machine Vision Conference (BMVC) (2016)
46. Malti, A., Hartley, R., Bartoli, A., Kim, J.H.: Monocular template-based 3d reconstruction of extensible surfaces with local linear elasticity. In: Computer Vision and Pattern Recognition (CVPR). pp. 1522–1529 (2013)
47. McInerney, T., Terzopoulos, D.: A finite element model for 3d shape reconstruction and nonrigid motion tracking. In: International Conference on Computer Vision (ICCV). pp. 518–523 (1993)
48. Mitiche, A., Mathlouthi, Y., Ben Ayed, I.: Monocular concurrent recovery of structure and motion scene flow. Frontiers in ICT **2**, 16 (2015)
49. Moreno-Noguer, F., Porta, J.M., Fua, P.: Exploring ambiguities for monocular non-rigid shape estimation. In: European Conference on Computer Vision (ECCV). pp. 370–383 (2010)
50. NVIDIA Corporation: NVIDIA CUDA C programming guide (2018), version 9.0
51. 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)
52. Paladini, M., Bartoli, A., Agapito, L.: Sequential non-rigid structure-from-motion with the 3d-implicit low-rank shape model. In: European Conference on Computer Vision (ECCV). pp. 15–28 (2010)
53. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: Advances in Neural Information Processing Systems Workshops (NIPS-W) (2017)
54. Paszke, A., Gross, S., Massa, F., Chintala, S.: pytorch. https://github.com/pytorch (2018)

55. Perriollat, M., Hartley, R., Bartoli, A.: Monocular template-based reconstruction of inextensible surfaces. International Journal of Computer Vision (IJCV) **95**(2), 124–137 (2011)
56. 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). pp. 4681–4690 (2018)
57. Riegler, G., Ulusoy, A.O., Bischof, H., Geiger, A.: Octnetfusion: Learning depth fusion from data. In: International Conference on 3D Vision (3DV) (2017)
58. Russell, C., Fayad, J., Agapito, L.: Dense non-rigid structure from motion. In: International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT). pp. 509–516 (2012)
59. Salzmann, M., Fua, P.: Reconstructing sharply folding surfaces: A convex formulation. In: Computer Vision and Pattern Recognition (CVPR). pp. 1054–1061 (2009)
60. Salzmann, M., Fua, P.: Linear local models for monocular reconstruction of deformable surfaces. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **33**(5), 931–944 (2011)
61. Salzmann, M., Hartley, R., Fua, P.: Convex optimization for deformable surface 3-d tracking. In: International Conference on Computer Vision (ICCV) (2007)
62. Salzmann, M., Lepetit, V., Fua, P.: Deformable surface tracking ambiguities. In: Computer Vision and Pattern Recognition (CVPR) (2007)
63. Sela, M., Richardson, E., Kimmel, R.: Unrestricted facial geometry reconstruction using image-to-image translation. In: International Conference on Computer Vision (ICCV) (2017)
64. Stay & Play Rotorua Ltd: *A hot balloon*. http://stayandplaynz.com/rotorua/the-real-new-zealand-experience/, [Online; accessed June 29, 2018]
65. Suwajanakorn, S., Kemelmacher-Shlizerman, I., Seitz, S.M.: Total moving face reconstruction. In: European Conference on Computer Vision (ECCV) (2014)
66. 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)
67. Tao, L., Matuszewski, B.J.: Non-rigid structure from motion with diffusion maps prior. In: Computer Vision and Pattern Recognition (CVPR). pp. 1530–1537 (2013)
68. Tateno, K., Tombari, F., Laina, I., Navab, N.: Cnn-slam: Real-time dense monocular slam with learned depth prediction. In: Computer Vision and Pattern Recognition (CVPR) (2017)
69. Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Theobalt, C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: International Conference on Computer Vision (ICCV) (2017)
70. Textures.com: *WrincklesHanging0037*. https://www.textures.com/browse/hanging/112398, [Online; accessed June 29, 2018]
71. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: a factorization method. International Journal of Computer Vision (IJCV) **9**, 137–154 (1992)
72. Tome, D., Russell, C., Agapito, L.: Lifting from the deep: Convolutional 3d pose estimation from a single image. In: Computer Vision and Pattern Recognition (CVPR) (2017)
73. 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)
74. Varol, A., Shaji, A., Salzmann, M., Fua, P.: Monocular 3d reconstruction of locally textured surfaces. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **34**(6), 1118–1130 (2012)
75. Vicente, S., Agapito, L.: Soft inextensibility constraints for template-free non-rigid reconstruction. In: European Conference on Computer Vision (ECCV). pp. 426–440 (2012)

76. Wandt, B., Ackermann, H., Rosenhahn, B.: 3d reconstruction of human motion from monocular image sequences. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **38**(8), 1505–1516 (2016)
77. White, R., Forsyth, D.A.: Combining cues: Shape from shading and texture. In: Computer Vision and Pattern Recognition (CVPR). pp. 1809–1816 (2006)
78. Xiao, D., Yang, Q., Yang, B., Wei, W.: Monocular scene flow estimation via variational method. Multimedia Tools and Applications **76**(8), 10575–10597 (2017)
79. Xiao, J., Chai, J., Kanade, T.: A closed-form solution to non-rigid shape and motion recovery. International Journal of Computer Vision (IJCV) **67**(2), 233–246 (2006)
80. 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)
81. Zhou, X., Zhu, M., Pavlakos, G., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2018)
82. Zhu, S., Zhang, L., Smith, B.M.: Model evolution: An incremental approach to non-rigid structure from motion. In: Computer Vision and Pattern Recognition (CVPR). pp. 1165–1172 (2010)