IsMo-GAN: Adversarial Learning for Monocular Non-Rigid 3D Reconstruction

Soshi Shimada[1,2] Vladislav Golyanik[1,3] Christian Theobalt[3] Didier Stricker[1,2]

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

Abstract

The majority of the existing methods for non-rigid 3D surface regression from a single 2D image require an object template or point tracks over multiple frames as an input, and are still far from real-time processing rates. In this work, we present the Isometry-Aware Monocular Generative Adversarial Network (IsMo-GAN) — an approach for direct 3D reconstruction from a single image, trained for the deformation model in an adversarial manner on a light-weight synthetic dataset. IsMo-GAN reconstructs surfaces from real images under varying illumination, camera poses, textures and shading at over 250 *Hz. In multiple experiments, it consistently outperforms multiple approaches in the reconstruction accuracy, runtime, generalisation to unknown surfaces and robustness to occlusions. In comparison to the state-of-the-art, we reduce the reconstruction error by* 10-30% *including the textureless case and our surfaces evince fewer artefacts qualitatively.*

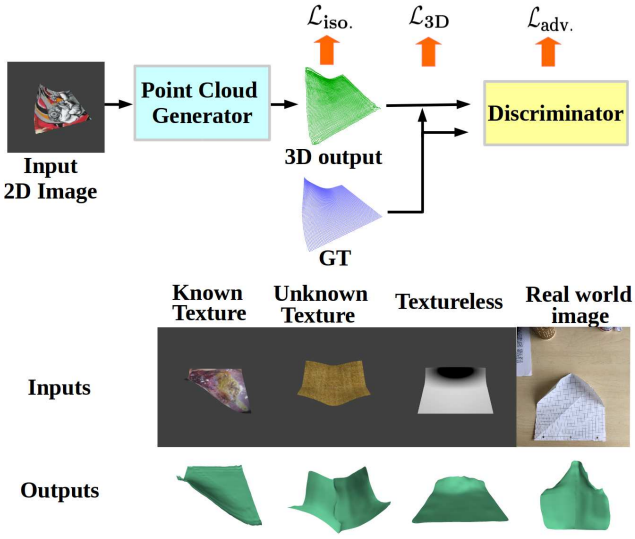

Figure 1: Overview of our IsMo-GAN approach. (top) The *generator* network accepts a 2D RGB image segmented by the *object detection* network (OD-Net) and returns a 3D point cloud. The output and ground truth (GT) are fed to the *discriminator* network which serves as a surface regulariser. (bottom) Example reconstructions by IsMo-GAN in different scenarios: a known texture, an unknown texture, a textureless surface and a reconstruction of a real image.

1. Introduction

Monocular non-rigid 3D reconstruction from single 2D images is a challenging ill-posed problem in computer vision with many useful applications. Such factors as varying illumination, external and self occlusions in the scene and lack of texture further complicate the setting. In recent times, dense monocular non-rigid reconstruction was mostly tackled by shape-from-template (SfT) techniques and non-rigid structure from motion (NRSfM). SfT requires a template — an accurate geometry estimate corresponding to one of the 2D views known in advance [56, 50, 4, 44, 22, 71] —, whereas NRSfM relies on motion and deformation cues in the input point tracks over multiple views [6, 61, 20, 15, 46, 18, 32]. Currently, there is a lack of approaches supporting real-time processing rates which is a desired property for interactive applications.

At the same time, convolutional neural networks (CNN) [34] have been successfully applied in various domains of computer vision including, among other architectures, fully convolutional encoder-decoders to convert data modalities, as in object segmentation and contour detection [3, 7, 8, 24]. Many applications benefit from the properties of different modifications of generative adversarial networks (GAN) [25, 28, 43, 52, 57, 73]. GAN include two competing neural networks which are trained simultaneously during the training phase — the *generator* and *discriminator* networks. Starting from arbitrary signals, the generator mimics data distributions of the training dataset and learns to pass the discriminator's test on sample authenticity. The discriminator estimates the probabilities that given outputs originate from the training dataset or from the generator. This adversarial manner allows the generator to pursue a high-level objective, *i.e, "generate outputs that look authentic and have the properties of the representative samples"*.

In this paper, we propose *Isometry-Aware Monocular Generative Adversarial Network* (IsMo-GAN) — a framework with several CNNs for the recovery of a deformable 3D structure from 2D images, see Fig. 1 for an overview. Our approach learns a deformation model, and the individual CNNs are trained in an adversarial manner to enable generalisation to unknown data and robustness to noise. In the 3D reconstruction task, the adversarial training is targeted at the objective *"generate realistic 3D geometry"*. This high-level objective improves the reconstruction qualitatively because lower Euclidean distances between the predicted and ground truth geometry do not necessarily imply higher visual quality.

1.1. Contributions

By combining a CNN with skipping connections for 3D reconstruction, an adversarial learning (a discriminator and geometry regulariser) and a confidence map indicator for object segmentation, we develop an approach that directly regresses 3D point clouds while consistently outperforming competing methods [15, 71, 60, 38, 18, 17, 5] quantitatively by 10-30% across various experiments and scenarios (see Fig. 2 and Sec. 4). IsMo-GAN enhances the **reconstruction accuracy of real images** compared to the competing methods, including the regression of **textureless surfaces**. The demonstrated improvement is due to the key technical **contributions** of the method — first, **the adversarial regulariser loss** and, second, the integrated **object detection network** (OD-Net) for the foreground-background segmentation, as we show in the comparison with the most closely related previous method [17] (refer to Sec. 4).

IsMo-GAN does not require a template, camera calibration parameters or point tracks over multiple frames. Our pipeline is robust to varying illumination and camera poses, internal and external occlusions and unknown textures, and all that with a training on light-weight datasets of non-rigid surfaces [17, 5]. Concerning the runtime, IsMo-GAN exceeds conventional methods by a large margin and reconstructs **up to 250 states per second**. Compared to computationally expensive 3D [41, 9, 53] and graph convolutions [10, 62], IsMo-GAN applies 2D convolutions [31] for 3D surface regression from 2D images. To the best of our knowledge, our study is the first one for deformation model-aware non-rigid 3D surface regression from single monocular images with point set representation trained in an adversarial manner and a masking network in a single pipeline.

1.2. Paper Structure

The rest of the paper is organised as follows. In Sec. 2, we discuss related works. Technical details and the network architectures are elaborated in Sec. 3. Sec. 4 describes the experiments. Finally, we discuss the method including its limitations in Sec. 5 and summarise the study in Sec. 6.

2. Related Work

In this section, we review the most related model-based (Sec. 2.1) and deep neural network (DNN)-based techniques (Secs. 2.2–2.3).

2.1. Unsupervised Learning Methods

NRSfM factorises point tracks over multiple views into camera poses and non-rigid shapes relying on motion and deformation cues as well as weak prior assumptions (*e.g.,* temporal state smoothness or expected deformation complexity) [6, 61, 20, 15, 32]. Only recently NRSfM has entered the realm of dense reconstructions [55, 15, 2, 18]. Dense NRSfM requires distinctive textures on the target object during the tracking phase [14, 59, 37]. Even though the reconstruction can be performed at interactive rates [2], obtaining dense correspondences from real images can significantly decrease the overall throughput of the pipeline. The recent work of Gallardo *et al.* [13] can cope with textureless objects by considering shading and still, their solution is computationally expensive. IsMo-GAN reconstructs textureless objects upon the learned deformation model while fulfilling the real-time requirement.

SfT, also known as non-rigid 3D tracking, requires a 3D template known in advance, *i.e.,* an accurate reconstruction with given 2D-3D correspondences [56, 4, 71]. Several approaches enhance robustness of SfT to illumination changes with the shape-from-shading component [40, 38]. Our method does not require a template — all we need as an input is a single monocular 2D image during the surface inference phase. At the same time, IsMo-GAN is trained in the supervised manner. The training dataset contains a sequence of 3D states along with the corresponding 2D images [17]. Thus, our framework bears a remote analogy with SfT, as IsMo-GAN is trained for a deformation model with a pre-defined surface at rest (or multiple surfaces at rest, in the extended version).

2.2. DNN-Based 3D Reconstruction Techniques

Methods for 3D reconstruction with DNNs primarily focus on rigid scenes [69, 26, 21, 11, 9, 16, 53, 33] while only a few approaches were proposed for the non-rigid case so far [17, 51]. Volumetric representation is often used in DNN based approaches [41, 9, 53]. In most cases, it relies on computationally costly 3D convolutions limiting the techniques in the supported resolution and throughput. Qualitatively, volumetric representations lead to discretisation artefacts. Our approach directly regresses 3D point coordinates by applying computationally less expensive 2D convolutions [34, 31], and surfaces recovered by IsMo-GAN are smoother and more realistic qualitatively.

Golyanik *et al.* [17] recently proposed Hybrid Deformation Model Network (HDM-Net) for monocular non-rigid

3D reconstruction targeting virtual reality applications. In their method, an encoder-decoder network is trained for a deformation model with a light-weight synthetic dataset of thin plate states in the point cloud representation. Rather than treating every image as a different rigid instance of a pre-defined object class [27], HDM-Net associates every input image with a non-rigid surface state imposing the isometry and feasibility constraint upon the learned deformation model. In addition, its objective function includes a contour loss. We do not use the contour loss as it increases the training time and does not make a significant difference in the reconstruction accuracy. We regress 50 states per second more on average with a higher accuracy compared to HDM-Net [17]. Moreover, IsMo-GAN shows more accurate results for occluded and textureless surfaces as well as when reconstructing from real images.

Pumarola *et al.* [51] combine three sub-networks for 2D heat-map generation with object detection, depth estimation and 3D surface regression. For the real-world scenario, they have to finetune the pipeline. In contrast, IsMo-GAN automatically segments and reconstructs real images, with no need for further parameter tuning. Bednařík *et al.* [5] employ a trident network with a single encoder and three decoders for the depth-map, normal map and 3D mesh estimation. For mesh decoding, they use a fully-connected layer. Similar to [17, 51], our generator consists of 2D convolutional layers and includes multiple sub-networks. In contrast, IsMo-GAN uses an adversarial loss which leads to the consistently improved accuracy across different scenarios.

2.3. Adversarial Learning in Computer Vision

GAN were initially introduced as a generative model for the sampling of new instances from a predefined class [19]. In GAN, learning to sample from a training distribution is performed through a two-player game and formalised as the adversarial loss. GAN were applied for various tasks including image inpainting [49, 70], video generation [68, 63], 2D image resolution enhancement [35, 64], image texture transfer [36] and a transfer from texts to images [72], among others. Several improvements for training convergence and performance of GAN were subsequently proposed over the last years [25, 43, 52, 73]. The adversarial loss is also applicable as a fidelity regulariser in rigid 3D reconstruction [26]. In [26], the conditional adversarial loss demands the inference result to be close to the shape probability distribution of the training set. Adversarial loss in IsMo-GAN targets the deformation model of a thin structure instead of the space of multiple shapes, *i.e.,* the recovered surfaces are constrained to be reasonable with respect to the probability distribution of the learned space of non-rigid states. To the best of our knowledge, it is the first time an adversarial loss is applied in monocular non-rigid surface reconstruction with DNNs.

3. The Proposed Method

In this section, we first describe the proposed architecture (Sec. 3.1) followed by the loss functions (Sec. 3.2). Next, we provide details about the dataset (Sec. 3.3) and IsMo-GAN training (Sec. 3.4).

3.1. Network Architecture

We propose a DNN architecture that consists of a generator and discriminator networks, see Fig. 2 for the schematic visualisation. The generator is, in turn, composed of OD-Net and Reconstruction Network (Rec-Net), both based on an encoder-decoder architecture with skipping connections [23]. The input images are of the resolution 224×224. OD-Net has a U-net structure [54, 42], and it is responsible for the generation of a grayscale confidence map indicating the position of the target object. The generated confidence map is subsequently binarised [45] and the target object is extracted with the algorithm of Suzuki *et al.* [58]. Compared to the customised U-Net [42], the number of downsampling and upsampling convolutional blocks is reduced by one in our OD-Net due to the relatively small size of the training dataset (see Sec. 3.3). Rec-Net is a residual encoder-decoder network. The encoder extracts relevant features for 3D reconstruction from the given 2D inputs and converts them into the latent space representation. The decoder increases the dimensionality of the latent space in height and width and adjusts the depth of the latent space until its activation reaches the dimensionality of $73 \times 73 \times 3$, *i.e.,* the dimensionality of the ground truth training states.

Our discriminator consists of four blocks — a convolutional layer, leaky rectified linear unit (ReLU) [39], batch normalisation and a fully-connected layer. To enhance training stability, the first layer set of the discriminator does not contain batch normalisation [52]. The output from Rec-Net is evaluated by several loss functions. First, we penalise Euclidean distances between the ground truth 3D geometry and output of the generator with the sum of absolute differences (SAD). Next, similar to [17], we assume the observed surfaces to be isometric and introduce a soft isometry constraint, *i.e.,* a loss function penalising the roughness and non-isometric effects (*e.g.,* shrinking and dilatation) of the predicted 3D geometry in an unsupervised manner. For more plausible and realistic outputs, we introduce an adversarial loss [19] which targets the deformation model of a surface. In the following section, all three losses of IsMo-GAN are described in detail.

3.2. Loss Functions

Suppose $\mathbf{I} = \{\mathbf{I}_m^n\}$, $m \in \{1, \dots, M\}$, $n \in \{1, \dots, N\}$ denote 2D input images, with the total number of states M and the total number of images for each state N. Let $\mathbf{S}^{\text{GT}} = \{\mathbf{S}_m^{\text{GT}}\}$ be the ground truth geometry. $\mathbf{G}$ and $\mathbf{D}$

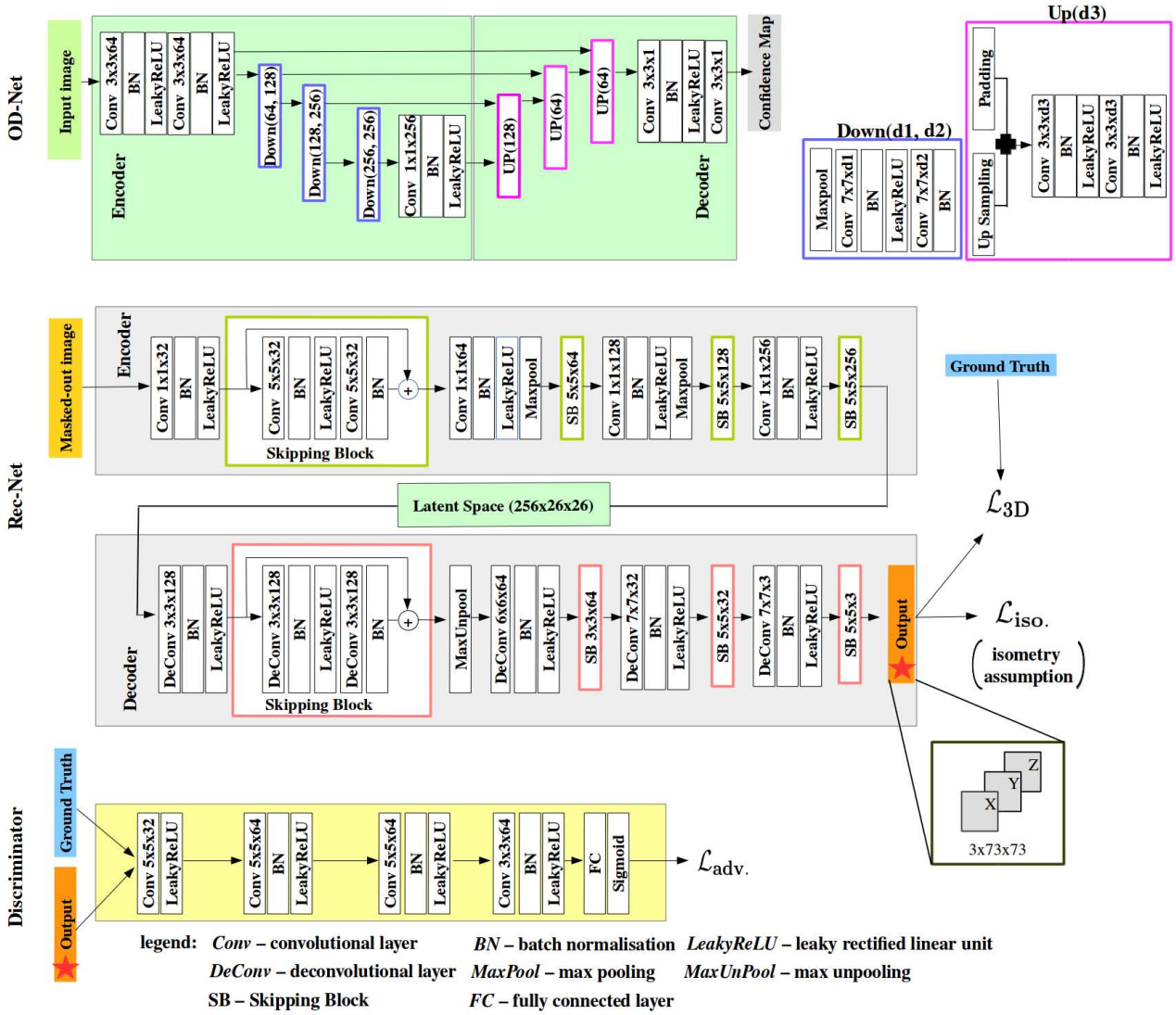

Figure 2: Architecture of the proposed IsMo-GAN framework. *Up Sampling* in OD-Net doubles the width and height of the input using binary interpolation. OD-Net applies *padding* on the inputs to equalise the input dimensionalities if necessary. Rec-Net accepts images of the size $224 \times 224 \times 3$ (with three colour channels). The output is a $73 \times 73 \times 3$ dense reconstruction, with 73^2 points per frame. The fully-connected layer in the discriminator converts the dimensionality from 3136 to 1 in order to generate the probabilistic decision about the input authenticity (the activation from the fourth convolutional layer is of the dimension $7 \times 7 \times 64$ leading to the dimensionality 3136 when concatenated).

denote the generator (Rec-Net) and discriminator components. The total loss of IsMo-GAN reads:

$$\mathcal{L}(\mathbf{I}, \mathbf{S}^{\text{GT}}) = \mathcal{L}_{\mathbf{adv.}}(\mathbf{I}, \mathbf{S}^{\text{GT}}) + \mathcal{L}_{\mathbf{iso.}}(\mathbf{G}(\mathbf{I})) + \mathcal{L}_{\mathbf{3D}}(\mathbf{G}(\mathbf{I}), \mathbf{S}^{\text{GT}}), \tag{1}$$

where $\mathbf{G}(\mathbf{I})$ stands for the reconstructed 3D surfaces.

3D Loss. The 3D loss is based on SAD function which penalises the Euclidean distance between ground truth geometry and the predicted 3D geometry per point:

$$\mathcal{L}_{\mathbf{3D}}(\mathbf{G}(\mathbf{I}), \mathbf{S}^{\text{GT}}) = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} |\mathbf{S}_m^{\text{GT}} - \mathbf{G}(\mathbf{I}_m^n)|. \tag{2}$$

Isometry Prior. The isometry prior penalises surface roughness. We assume the target object to be isometric which implies that every 3D point has to be located close to the neighbouring points. This loss was already effectively applied in HDM-Net [17]. The corresponding loss function is expressed in terms of the difference between the predicted geometry and its smoothed version:

$$\mathcal{L}_{\mathbf{iso.}}(\mathbf{G}(\mathbf{I})) = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} |\hat{\mathbf{S}}_m^n - \mathbf{G}(\mathbf{I}_m^n)|. \tag{3}$$

In Eq. (3), $\hat{\mathbf{S}}_m^n$ denotes the surface smoothed by a Gaussian kernel:

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

where $*$ is the convolution operator, σ is the standard deviation of the Gaussian kernel, and x and y stand for the point coordinates.

Adversarial Loss. As an objective function of the adversarial training, we employ binary cross entropy (BCE) [19] defined as

$$\mathcal{L}_{\mathbf{G}}(\mathbf{I}) = -\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\log(\mathbf{D}(\mathbf{G}(\mathbf{I}_m^n)) \tag{5}$$

for the generator, and

$$\mathcal{L}_{\mathbf{D}}(\mathbf{I}, \mathbf{S}^{\mathrm{GT}}) = -\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\left[\log(\mathbf{D}(\mathbf{S}_m^{\mathrm{GT}}))+\log(1-\mathbf{D}(\mathbf{G}(\mathbf{I}_m^n))\right] \tag{6}$$

for the discriminator. The adversarial loss is then comprised of the sum of both components:

$$\mathcal{L}_{\mathbf{adv.}}(\mathbf{I}, \mathbf{S}^{\mathrm{GT}}) = \mathcal{L}_{\mathbf{G}}(\mathbf{I}) + \mathcal{L}_{\mathbf{D}}(\mathbf{I}, \mathbf{S}^{\mathrm{GT}}). \tag{7}$$

The adversarial loss in Eq. (7) defines the high-level goal that encourages IsMo-GAN to generate visually more realistic surfaces. It is the core component which enables IsMo-GAN to outperform HDM-Net [17] by $10-15\%$ quantitatively as well as qualitatively on real images (see Sec. 4.1).

We observed that using SAD as 3D loss tends to propagate the surface roughness from the input to the output. The isometry prior reduces the roughness, slightly shrinks the output and smoothes the corners. The adversarial loss compensates for these undesired effects of the 3D loss and the isometry prior, and serves as a novel regulariser for surface deformations.

3.3. Training Datasets

In this section, we elaborate on the main datasets [17, 30] used to train the OD-Net, Rec-Net and the discriminator. In Sec. 4.2, we extra use the *textureless cloth* dataset [5] to train a variation of our pipeline and compare its performance on textureless surfaces.

3.3.1 Deformation Model Dataset

We use the synthetic 2D-3D thin plate dataset from [17] for the training and tests. In total, the dataset contains 4648 states representing different isometric non-linear deformations of a thin plate structure (*e.g.,* waving deformations and bending). Due to the original 4:1 training-test split, $M = 3728$, and $N = 60$ (three textures illuminated by a light source at four different locations, and each combination of the texture and illumination is rendered with five virtual cameras). Every 3D state contains 73^2 3D points sampled on a regular grid at rest, with a consistent topology across all states. For each 3D state, there are corresponding rendered 2D images of the resolution 256×256[1] for the combinations with five different positions of the light source, four different textures (*endoscopy*, *graffiti*, *clothes* and *carpet*) and five different camera poses. To train IsMo-GAN and competing methods for the shape-from-shading, we extend the thin plate dataset [17] with a subsequence of deforming textureless surfaces (the states are left the same while the texture is removed). In our dataset extension, $M = 3728$ and $N = 5$ (no texture, five virtual cameras).

3.3.2 OD-Net Dataset

To train OD-Net, we generate a mixed image dataset with varying backgrounds (*sky*, *office* and *forest*) and the corresponding binary masks. First, we randomly translate the target object in the images from the deformation model dataset (Sec. 3.3.1). Next, we combine the first part with a dataset of real-world RGB images and the corresponding binary masks from [30]. In total, our mixed dataset contains $\approx 14k$ images and corresponding binary masks.

3.4. Training Details

We use Adam [29] for optimisation of network parameters, with the learning rate of 10^{-3} and the batch size of 8. OD-Net and Rec-Net are separately trained using the mixed binary mask dataset (Sec. 3.3.2) and 2D-3D dataset (Sec. 3.3.1) respectively. In total, we train Rec-Net and OD-Net for 130 and 30 epochs respectively. The architecture is implemented using *PyTorch* [47, 48]. In the 2D-3D dataset, we extract 20 sequential states out of every 100 consecutive states for testing and use the remaining data for Rec-Net training. Likewise, we divide the binary mask dataset in the ratios $8:2$ for the training and testing of OD-Net. We use mean squared error (MSE) to penalise the discrepancy between the output and the ground truth binary images.

4. Experimental Evaluation

We evaluate the reconstruction accuracy of IsMo-GAN with different illuminations, textures and occlusions in the input images. Our system for training and experiments includes 256 GB RAM, Intel Xeon CPU E5-2687W v3 running at 3.10 GHz and GeForce GTX 1080Ti GPU with 11 GB RAM running under Ubuntu 16.04. We compare our framework with three template-based reconstruction methods of Yu *et al.* [71], Liu-Yin *et al.* [38] and Tien Ngo *et al.* [60], two NRSfM approaches based on different principles, *i.e.,* variational NRSfM approach (VA) [15] and

[1] the input images are resized to 224×224 in our pipeline

	Yu *et al.* [71]	Liu-Yin *et al.* [38]	AMP [18]	VA [15]	HDM-Net [17]	IsMo-GAN
t, *sec.*	3.305	5.42	0.035	0.39	0.005	**0.004**
e_{3D}	1.3258	1.0049	1.6189	0.46	0.0251	**0.0175**
σ	**0.007**	0.0176	1.23	0.0334	0.03	0.01

Table 1: Reconstruction times per frame t in *seconds*, e_{3D} and standard deviation σ for Yu *et al.* [71], Liu-Yin *et al.* [38], AMP [18], VA [15], HDM-Net [17] and our IsMo-GAN method, for the test interval of 400 frames.

		illum. 1	*illum. 2*	*illum. 3*	*illum. 4*	*illum. 5*
HDM-Net [17]	e_{3D}	0.07952	0.0801	0.07942	0.07845	0.07827
	σ	0.0525	0.0742	0.0888	0.1009	0.1123
IsMo-GAN	e_{3D}	**0.06803**	**0.06908**	**0.06737**	**0.06754**	**0.06685**
	σ	**0.0499**	**0.0696**	**0.0824**	**0.093**	**0.102**

Table 2: Comparison of 3D error for different illuminations. The *illuminations* 1-4 are known, and the *illumination* 5 is unknown.

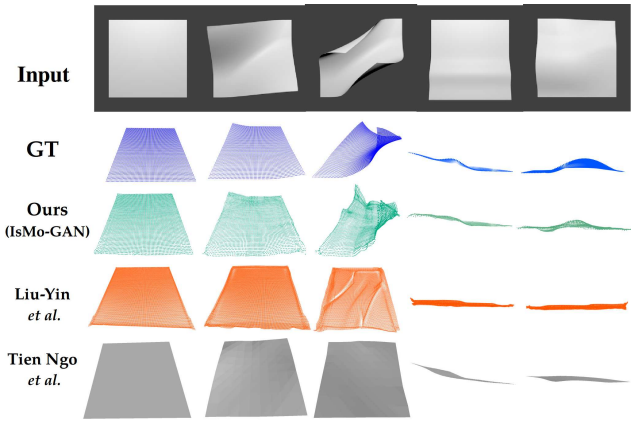

Figure 3: Selected reconstruction results of Liu-Yin *et al.* [38], Tien Ngo *et al.* [60] and IsMo-GAN on the textureless surfaces from the training set.

Accelerated Metric Projections (AMP) [18], HDM-Net of Golyanik *et al.* [17] and monocular surface reconstruction approach for textureless surfaces of Bednařík *et al.* [5]. [38] is an extension of [71] with a shape-from-shading component. For consistency, we adopt the evaluation setting as proposed in [17] and report the 3D reconstruction error e_{3D} along with the standard deviation of e_{3D} over a set of frames denoted by σ. e_{3D} is defined as

$$e_{3D} = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \frac{\|\mathbf{S}_m^{\mathrm{GT}} - \mathbf{G}(\mathbf{I}_m^n)\|_{\mathcal{F}}}{\|\mathbf{S}_m^{\mathrm{GT}}\|_{\mathcal{F}}}, \qquad (8)$$

where $\|\cdot\|_{\mathcal{F}}$ denotes the Frobenius norm.

4.1. Synthetic Thin Plate Dataset [17]

Table 1 summarises the accuracy and the runtimes on a test sub-sequence with 400 frames chosen such that it can be processed by all tested methods. AMP [18] has the highest throughput, and [15] shows the highest accuracy among non deep learning methods. IsMo-GAN outperforms all other methods in the reconstruction accuracy. Compared to HDM-Net [17], the runtime improves by 0.001 seconds per frame on average which means that IsMo-GAN supports processing rates of up to 250 Hz compared to 200 Hz of HDM-Net. As shown in Table 2, our framework also excels HDM-Net [17] in the test with varying illuminations. We do not observe a large difference in e_{3D} for different positions of the light source, which suggests the enhanced property of illumination invariance. We report e_{3D} for known (*endoscopy*, *graffiti* and *clothes*) and unknown (*carpet*) textures in Table 3. In all cases, our approach outperforms HDM-Net [17] reducing the error by $> 20\%$ on average. As expected, e_{3D} is higher for the unknown texture compared to the known ones. Still, we do not find severe qualitative faults in the reconstructions. In the textureless case, our approach shows much lower e_{3D} than Liu-Yin *et al.* [38] and $\approx 30\%$ lower e_{3D} than HDM-Net, see Table 4 and Fig. 3 with visualisations. Liu-Yin *et al.* [38] assume the contour of the target object to be consistent since it uses masking to distinguish the region of interest from the background. Therefore, for a fair comparison, we choose predominantly small deformations from our dataset (see Fig. 3). Tien Ngo *et al.* [60] support poorly textured surfaces when the observed deformations are rather small. All in all, this is a significant improvement compared to the baseline HDM-Net approach [17], as IsMo-GAN uses the same training dataset for the geometry regression as HDM-Net, while relying on other regression criteria (*e.g.,* adversarial loss).

		endoscopy	*graffiti*	*clothes*	*carpet*
HDM-Net [17]	e_{3D}	0.0485	0.0499	0.0489	0.1442
	σ	**0.0135**	0.022	0.0264	0.0269
IsMo-GAN	e_{3D}	**0.0336**	**0.0333**	**0.0353**	**0.1105**
	σ	0.0148	**0.0208**	**0.0242**	**0.0268**

Table 3: e_{3D} comparison for differently textured surfaces under the same illumination (*illumination* 1).

	Liu-Yin *et al.* [38]	Tien Ngo *et al.* [60]	HDM-Net [17]	IsMo-GAN
e_{3D}	0.9109	0.0945	0.0994	**0.0677**
σ	**0.0677**	0.1170	0.0809	**0.0697**

Table 4: e_{3D} comparison of the template-based approaches [38, 60], HDM-Net [17] and IsMo-GAN on the textureless surfaces from the dataset of Golyanik *et al.* [17].

External Occlusions. Next, we evaluate IsMO-GAN in the scenario with external occlusions. We select an arbitrary 3D state from the test dataset with a comparably large deformation and introduce random circular noise (grey circles) into the corresponding 2D images. The size and the number of occluders vary as shown in Fig. 4-(a). We show the reconstruction results with five introduced occluders in Fig. 4-(b). For each combination of the occluder's size and

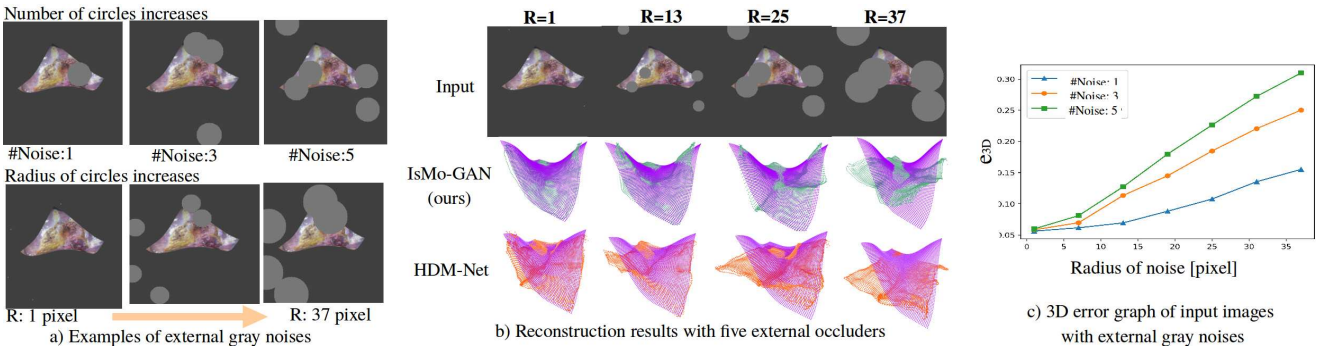

Figure 4: a) Exemplary occluded images with the increasing number of occluders (the top row) and the increasing size of the occluders (the bottom row). **b)** Outputs of our network and HDM-Net [17] with five external occluders — ground truth shapes (purple), reconstructions by IsMo-GAN (green) and HDM-Net (orange). **c)** 3D error graph for images with external occlusions. In **a)** and **b)**, R denotes radii of occluders. Best viewed in colour.

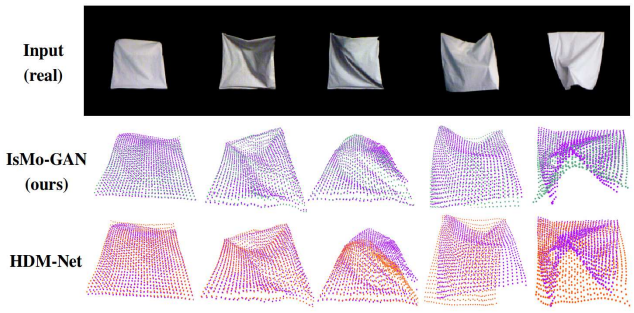

Figure 5: Selected reconstructions of the textureless *cloth* dataset [5].

the number of occluders, we generate ten images and report the average e_{3D} of the IsMo-GAN reconstructions for these images, see Fig. 4-(c). Unless the input image contains large occlusions, our network keeps the high reconstruction accuracy. When the occluder's size reaches 7 pixels, the slope of the graph increases which marks the robustness threshold, with up to 40% of the object being occluded.

4.2. Real Textureless Cloth Dataset [5]

We also evaluate IsMo-GAN on the real *cloth* dataset [5] with textureless deforming surfaces with varying shading. For every frame, the dataset includes ground truth meshes of the observed surfaces (with 31^2 points per state) obtained by fitting a mesh template to the captured depth maps [5]. Similarly to the evaluation with the thin plate dataset [17], we split all frames in the propotion 80-20% for the training and test subsets respectively. Since the *cloth* dataset contains 6237 samples and is smaller than the thin plate dataset, we omit two layer blocks in the generator's encoder (sets of convolutions, batch normalisation, leaky ReLU and max pooling) as well as two layer blocks in the generator's decoder (sets with deconvolutions, batch normalisation and leaky ReLU) and adjust the kernel sizes. The dimensionality of the latent space is reduced to $11 \times 11 \times 256$.

We compare the proposed IsMo-GAN with HDM-Net [17] and the monocular 3D reconstruction approach for non-rigid textureless surfaces of Bednařík *et al.* [5]. While Bednařík *et al.* report the SAD of $21.48\ mm$ [5], HDM-Net [17] achieves $17.65\ mm$. SAD of our IsMo-GAN amounts to $15.79\ mm$ which is a 26.5% improvement in comparison to Bednařík *et al.* [5][2], and a 10.5% improvement versus HDM-Net [17]. Compared to Bednařík *et al.* [5], we use deconvolutional layers in the decoder instead of the fully-connected layers. We believe that point adjacencies provide a strong cue for surface reconstruction. Fig. 5 shows selected reconstructions of challenging states. Even though SAD of HDM-Net is just $1.86\ mm$ larger as compared to IsMo-GAN on average, HDM-Net often fails to reconstruct states with large folds and deformations. Our architecture is not restricted to globally smooth surfaces and captures fine geometric details revealed by the shading cue.

4.3. Real Images (Qualitative Results)

Next, we evaluate IsMo-GAN on a collection of real images. In comparison to HDM-Net [17], the strength of IsMo-GAN is the enhanced generalisability to real data, even though the deformation model is trained on the synthetic dataset. Fig. 6 shows several reconstructions from real images by HDM-Net [17] and IsMo-GAN. We choose images with a different textures, deformations, illuminations and scene context, *i.e.,* waving flags, a hot air balloon, a bent paper, and a carpet with wrinkles. None of the textures were present in the training dataset, and IsMo-GAN captures well the main deformation mode and shape. The scene with the hot air balloon (Fig. 6-(c)) has an inhomogeneous background. Thanks to the OD-Net, IsMo-GAN generates qualitatively a more realistic reconstruction than HDM-Net. Fig. 6-(e) is an example of a deformation state which is the most dissimilar to the states in the training dataset. Remarkably, our approach recovers the rough

[2]note that details on the dataset split are not provided in [5]

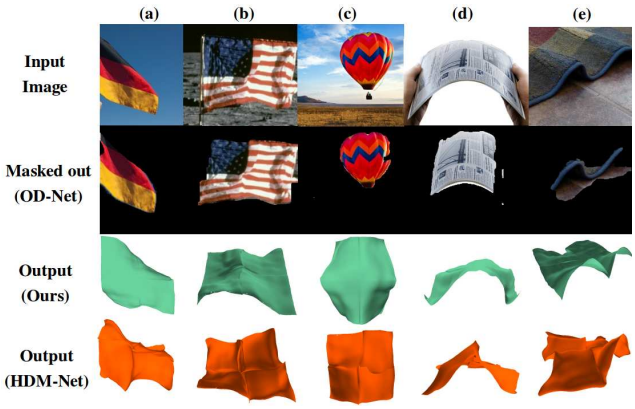

Figure 6: 3D reconstruction results from real images: a German flag [12], an American flag [1], a hot air balloon [66], a bent surface [65] and a carpet with a double wrinkle [67]. All input images are unknown to our pipeline. Note the qualitative improvement in the results of IsMo-GAN compared to the previous HDM-Net method [17]. Best viewed enlarged.

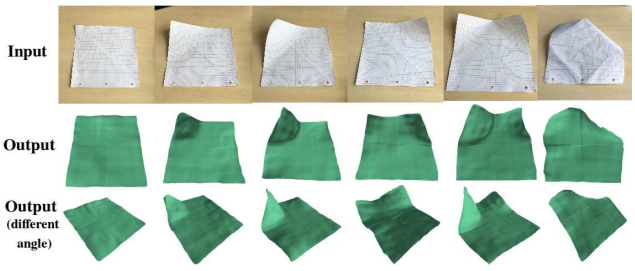

Figure 7: 3D reconstruction results by IsMo-GAN on the new real *origami* video sequence. Best viewed enlarged.

geometry of the object in the scene whereas HDM-Net fails to capture it.

Fig. 7 shows the reconstruction results by IsMo-GAN on the new *origami* video sequence. For *origami*, the main reconstruction cue is shading. Our approach captures well the global deformation of the target object with a weak texture in the real-world scene. Even though IsMo-GAN operates on individual images, the resulting dynamic reconstruction is temporally smooth.

5. Discussion

The experiments demonstrate the significant qualitative improvement of IsMo-GAN when reconstructing from real images compared to the previous most related method HDM-Net [17]. We can reconstruct surfaces more accurately in the challenging cases with external occlusions and lack of texture. The experiment with textureless *cloth* dataset [5] in Sec. 4.2 shows that our pipeline generalises well, can be easily adjusted for other scenarios (*e.g.,* different primary reconstruction cues, surface properties, types of deformations, *etc.*) and even outperform competing specialised methods. Even though we do not explicitly assume gradual frame-to-frame surface deformations, IsMo-GAN recovers temporally smooth surfaces from a video sequence as shown in Sec. 4.3. Especially the enhanced accuracy for textureless surfaces is a valuable property in passive 3D capture devices operating in real human-made environments. The inference in IsMo-GAN is light-weight (running at 250 Hz) and would require low energy, making it appealing for mobile augmented reality devices.

IsMo-GAN shows plausible results especially when similar non-rigid states appear in the training dataset or when the target state can be represented as a blend of known deformation states. Otherwise, IsMo-GAN can be retrained with a dataset encoding another deformation model or covering more deformation modes, as has been demonstrated in Sec. 4.2. Moreover, the accuracy of our approach depends on the accuracy of the binary mask generation in the real-world scenario, and this aspect can also be improved for pre-defined scenarios.

6. Conclusion

In this study, we introduce IsMo-GAN — the first DNN-based framework for deformation model-aware non-rigid 3D surface regression from single monocular images with point set representation trained in an adversarial manner. The proposed approach regresses realistic general non-rigid surfaces from real images while being trained on a synthetic dataset of non-rigid states with varying light sources, textures and camera poses. Compared to the previously proposed DNN based methods [17, 51], our pipeline localises the target object with an OD-Net. Thanks to the point cloud representation, we take advantage of computationally efficient 2D convolutions.

In the extensive experiments, IsMo-GAN outperforms competing methods, both model-based and DNN-based, in the reconstruction accuracy, throughput, robustness to occlusions as well as the ability to handle textureless surfaces. In future work, we plan to collect more real data and test IsMo-GAN in the context of medical applications. For video sequences such as *origami* reconstructed in Sec. 4.3, a temporal smoothness term could further improve the results. Another future direction is network pruning for deployment of IsMo-GAN on an embedded device. Besides, a superordinate system can include IsMo-GAN as a component for shape recognition or surface augmentation.

Acknowledgement

This work was supported by project VIDETE (01IW18002) of the the German Federal Ministry of Education and Research (BMBF).

References

[1] Ricardo Salam Pez. *Flag and Earth*. https://www.quora.com/What-are-the-flaws-in-this-conspiracy-theory-about-the-pictures-of-flags-on-the-Moon-from-the-Apollo-missions. [Online; acc. March 3, 2019].

[2] A. Agudo, J. M. M. Montiel, L. Agapito, and B. Calvo. Online dense non-rigid 3d shape and camera motion recovery. In *BMVC*, 2014.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *TPAMI*, 39(12):2481–2495, 2017.

[4] A. Bartoli, Y. Grard, F. Chadebecq, and T. Collins. On template-based reconstruction from a single view: Analytical solutions and proofs of well-posedness for developable, isometric and conformal surfaces. In *CVPR*, pages 2026–2033, 2012.

[5] J. Bednařík, P. Fua, and M. Salzmann. Learning to reconstruct texture-less deformable surfaces from a single view. In *3DV*, 2018.

[6] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *CVPR*, pages 690–696, 2000.

[7] J. Chen, L. Yang, Y. Zhang, M. Alber, and D. Z. Chen. Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. In *NIPS*, pages 3036–3044, 2016.

[8] K. Chen, M. Seuret, J. Hennebert, and R. Ingold. Convolutional neural networks for page segmentation of historical document images. In *ICDAR*, volume 1, pages 965–970, 2017.

[9] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3D-R2N2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016.

[10] M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NIPS*, pages 3844–3852, 2016.

[11] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, pages 2366–2374, 2014.

[12] empresassa.com. *A German flag*. http://www.empresassa.com.br/2018/05/alemanha-oferece-bolsa-de-estudos-com.html. [Online; acc. March 3, 2019].

[13] M. Gallardo, T. Collins, A. Bartoli, and F. Mathias. Dense non-rigid structure-from-motion and shading with unknown albedos. In *ICCV*, 2017.

[14] R. Garg, L. Pizarro, D. Rueckert, and L. Agapito. Dense multi-frame optic flow for non-rigid objects using subspace constraints. In *ACCV*, pages 460–473, 2011.

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

[16] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[17] V. Golyanik, S. Shimada, K. Varanasi, and D. Stricker. Hdm-net: Monocular non-rigid 3d reconstruction with learned deformation model. In *EuroVR*, 2018.

[18] V. Golyanik and D. Stricker. Dense batch non-rigid structure from motion in a second. In *WACV*, pages 254–263, 2017.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[20] P. F. U. Gotardo and A. M. Martinez. Non-rigid structure from motion with complementary rank-3 spaces. In *CVPR*, pages 3065–3072, 2011.

[21] J. Gwak, C. B. Choy, M. Chandraker, A. Garg, and S. Savarese. Weakly supervised 3d reconstruction with adversarial constraint. In *3DV*, pages 263–272, 2017.

[22] N. Haouchine, J. Dequidt, M.-O. Berger, and S. Cotin. Single View Augmentation of 3D Elastic Objects. In *ISMAR*, 2014.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[24] L. Hongsheng, Z. Rui, and W. Xiaogang. Highly efficient forward and backward propagation of convolutional neural networks for pixelwise classification. *arXiv preprint 1412.4526*, 2014.

[25] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, pages 5967–5976, 2017.

[26] L. Jiang, S. Shi, X. Qi, and J. Jia. Gal: Geometric adversarial loss for single-view 3d-object reconstruction. In *ECCV*, pages 820–834, 2018.

[27] A. Kanazawa, S. Tulsiani, A. A. Efros, and J. Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018.

[28] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ICLR*, 2018.

[29] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint 1412.6980*, 2014.

[30] A. Kolesnikov, M. Guillaumin, V. Ferrari, and C. H. Lampert. Closed-form approximate crf training for scalable image segmentation. In *ECCV*, pages 550–565, 2014.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[32] S. Kumar, A. Cherian, Y. Dai, and H. Li. Scalable dense non-rigid structure-from-motion: A grassmannian perspective. In *CVPR*, 2018.

[33] A. Kurenkov, J. Ji, A. Garg, V. Mehta, J. Gwak, C. B. Choy, and S. Savarese. Deformnet: Free-form deformation network for 3d shape reconstruction from a single image. In *WACV*, pages 858–866, 2018.

[34] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[35] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

[36] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *ECCV*, pages 702–716, 2016.

[37] W. Li, D. Cosker, and M. Brown. Drift robust non-rigid optical flow enhancement for long sequences. *Journal of Intelligent and Fuzzy Systems*, 31(5):2583–2595, 2016.

[38] Q. Liu-Yin, R. Yu, L. Agapito, A. Fitzgibbon, and C. Russell. Better together: Joint reasoning for non-rigid 3d reconstruction with specularities and shading. In *BMVC*, 2016.

[39] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013.

[40] A. Malti, A. Bartoli, and T. Collins. Template-based conformal shape-from-motion-and-shading for laparoscopy. In *IPCAI*, 2012.

[41] D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IROS*, pages 922–928, 2015.

[42] Milesial. *Pytorch implementation of the U-Net for image semantic segmentation, with dense CRF post-processing*. https://github.com/milesial/Pytorch-UNet, 2016. [Online; acc. March 3, 2019].

[43] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint 1411.1784*, 2014.

[44] J. Östlund, A. Varol, D. T. Ngo, and P. Fua. Laplacian meshes for monocular 3d shape recovery. In *ECCV*, 2012.

[45] N. Otsu. A threshold selection method from gray-level histograms. *IEEE T-SMC*, 9(1):62–66, Jan 1979.

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

[47] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS Workshops*, 2017.

[48] A. Paszke, S. Gross, F. Massa, and S. Chintala. pytorch. https://github.com/pytorch, 2018.

[49] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

[50] M. Perriollat, R. I. Hartley, and A. Bartoli. Monocular template-based reconstruction of inextensible surfaces. *IJCV*, 95:124–137, 2010.

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

[52] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint 1511.06434*, 2015.

[53] G. Riegler, A. O. Ulusoy, H. Bischof, and A. Geiger. Octnetfusion: Learning depth fusion from data. In *3DV*, 2017.

[54] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[55] C. Russell, J. Fayad, and L. Agapito. Dense non-rigid structure from motion. In *3DIMPVT*, pages 509–516, 2012.

[56] M. Salzmann, R. Urtasun, and P. Fua. Local deformation models for monocular 3d shape recovery. In *CVPR*, 2008.

[57] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, 2017.

[58] S. Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer vision, graphics, and image processing*, 30(1):32–46, 1985.

[59] B. Taetz, G. Bleser, V. Golyanik, and D. Stricke. Occlusion-aware video registration for highly non-rigid objects. In *WACV*, 2016.

[60] D. Tien Ngo, S. Park, A. Jorstad, A. Crivellaro, C. D. Yoo, and P. Fua. Dense image registration and deformable surface reconstruction in presence of occlusions and minimal texture. In *ICCV*, 2015.

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

[62] N. Verma, E. Boyer, and J. Verbeek. FeaStNet: Feature-Steered Graph Convolutions for 3D Shape Analysis. In *CVPR*, pages 2598–2606, 2018.

[63] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *NIPS*, pages 613–621, 2016.

[64] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang. Esrgan: Enhanced super-resolution generative adversarial networks. *ECCV Workshops*, 2018.

[65] webtekno.com. *A bent surface*. https://www.webtekno.com/apple-iphone-larda-esnek-lcd-ekranlar-kullanabilir-h24541.html. [Online; acc. March 3, 2019].

[66] William E Hollon. *A hot balloon*. https://www.pinterest.ca/pin/446137906809803925. [Online; acc. March 3, 2019].

[67] William E Hollon. *A wrinkled carpet*. https://www.answerplane.com/how-to-straighten-a-wrinkled-rug. [Online; acc. March 3, 2019].

[68] H. Wu, S. Zheng, J. Zhang, and K. Huang. Gp-gan: Towards realistic high-resolution image blending. *arXiv preprint 1703.07195*, 2017.

[69] J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NIPS*, pages 82–90, 2016.

[70] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018.

[71] R. Yu, C. Russell, N. D. F. Campbell, and L. Agapito. Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video. In *ICCV*, 2015.

[72] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *ICCV*, 2017.

[73] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.