Learning non-rigid surface reconstruction from spatio-temporal image patches

Matteo Pedone, Abdelrahman Mostafa and Janne Heikkilä

Center for Machine Vision Research and Signal Analysis
University of Oulu, Finland

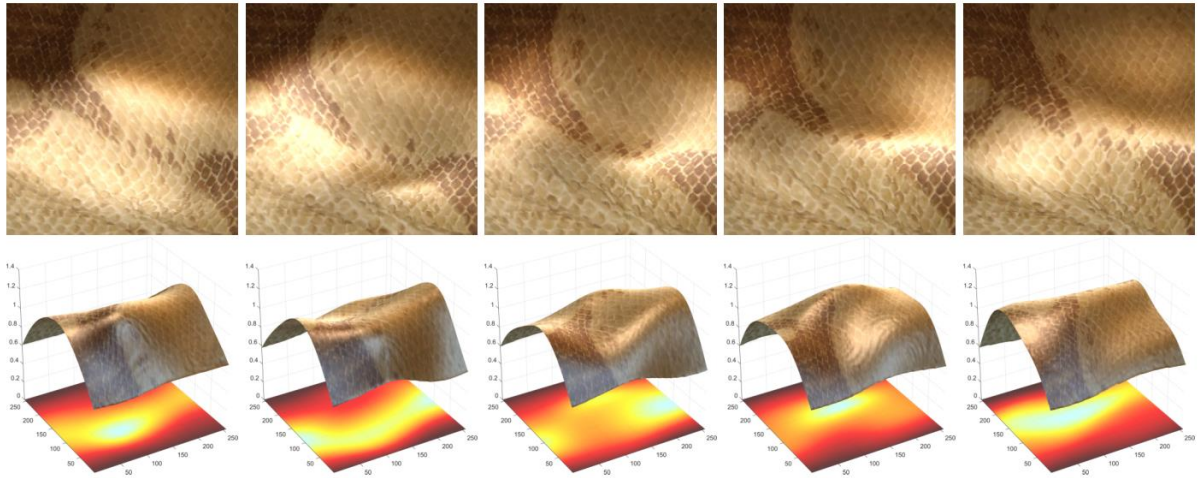
Learning non-rigid surface reconstruction from spatio-temporal image patches

- *Input:* **a video sequence of a moving object**

- *Output*: ***XYZ*-coordinates of the points**

- *Typical solutions involve*:

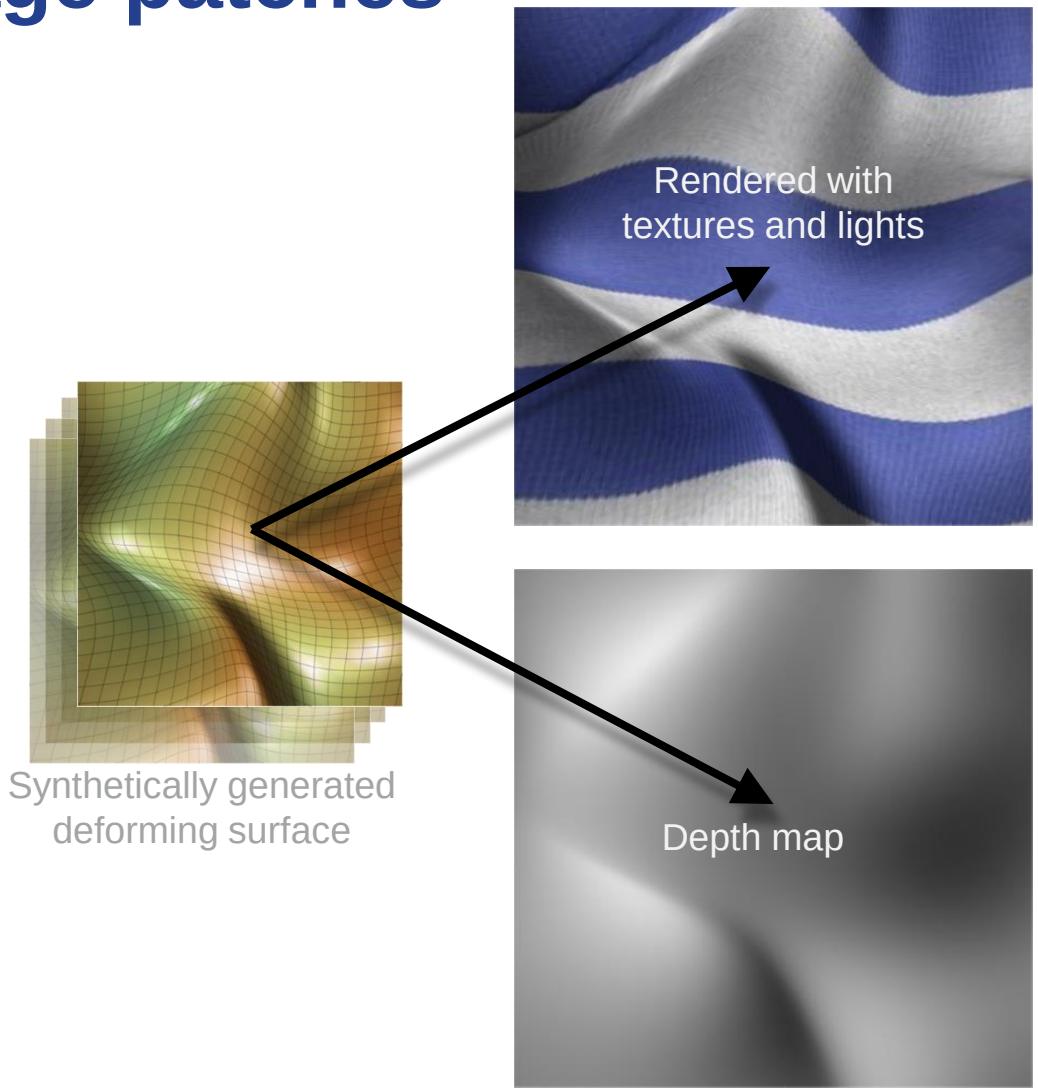
1. **Tracking feature points across frames + NRSfM**
2. **Exploiting assumptions on camera, shape of the object, trajectories etc.**



…ill-posed, mathematically challenging

Learning non-rigid surface reconstruction from spatio-temporal image patches

- *Motivation:* **Circumvent difficult mathematical challenges and avoid point tracking**
- *Idea:* **Train a network to infer shape directly from the video sequence…but how?**
- **Synthetically generate database of short movie clips of realistically deforming surfaces, and their corresponding depth maps.**
- **Divide the video into patches, estimate depth, combine together**



Learning non-rigid surface reconstruction from spatio-temporal image patches

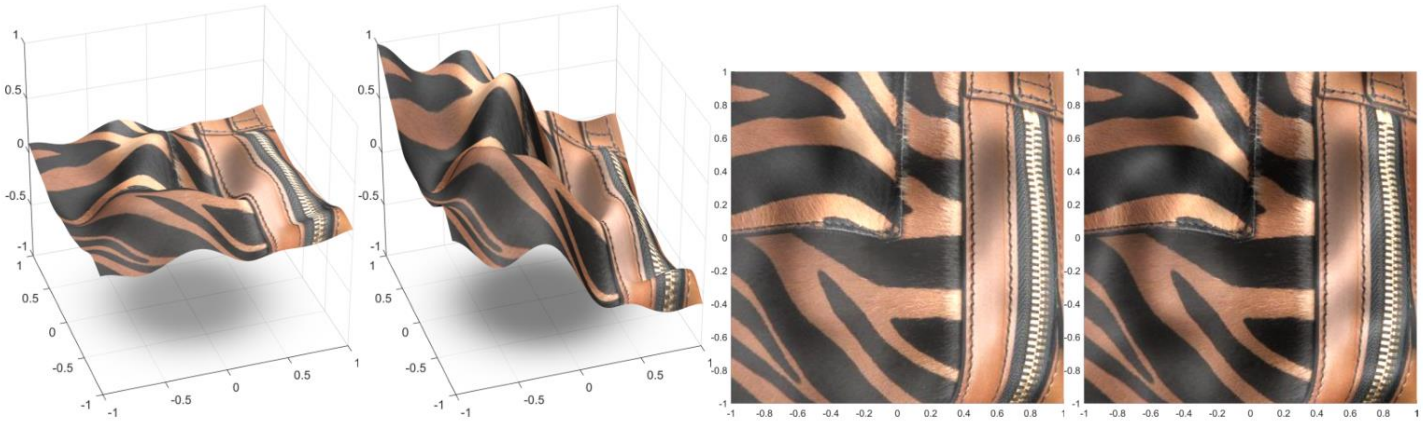
Assumptions:

1. **Static and orthographic camera (=> video depth estimation)**
2. **Non-negligible deformation of the object across time**
3. **Locally, the 4D structure of the object can be approximated with a parametric model**

Learning non-rigid surface reconstruction from spatio-temporal image patches

Orthographic camera:

- **Easier to train than perspective camera, but…**
- **PROBLEM: linear ambiguity *(GBR transformation)***



Example of generalized bas-relief ambiguity. From left to right, two versions of the same surface, of which the second one is a GBR transformed version of the first one, and their corresponding views from an orthographic camera located at $(0, 0, 1)$. The GBR transformation changes the orientation of the surface normals, which in turn slightly changes the albedo pattern of the surface. However, the second image can be mistakenly interpreted as its non-transformed version rendered with the same texture with slightly modulated pixel intensities.

Learning non-rigid surface reconstruction from spatio-temporal image patches

Orthographic camera:

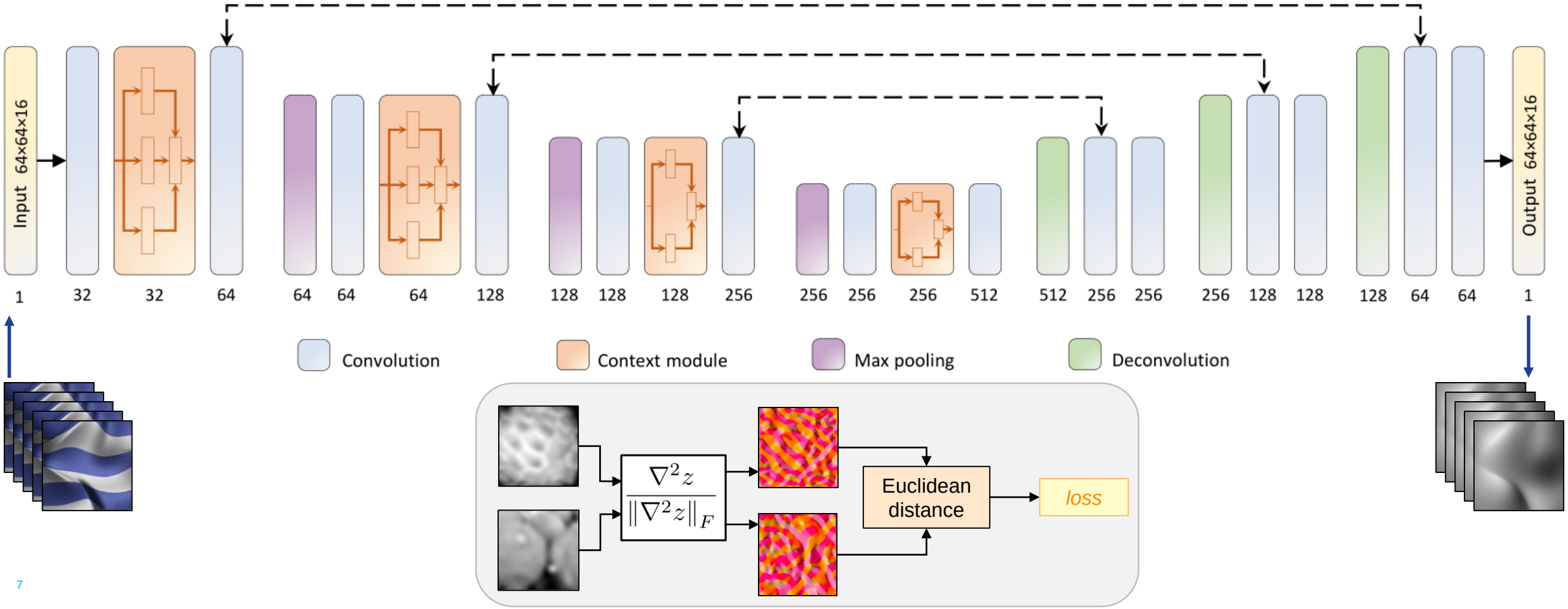
- **Easier to train than perspective camera, but…**
- **PROBLEM: linear ambiguity *(GBR transformation)***
- ***Proposed solution: represent surfaces with GBR-invariants***

$$\frac{\nabla^2 z}{\|\nabla^2 z\|_F}$$

The normalized Hessian of a depth map *z* is a *complete differential invariant* to generalized bas-relief transformations.

Learning non-rigid surface reconstruction from spatio-temporal image patches

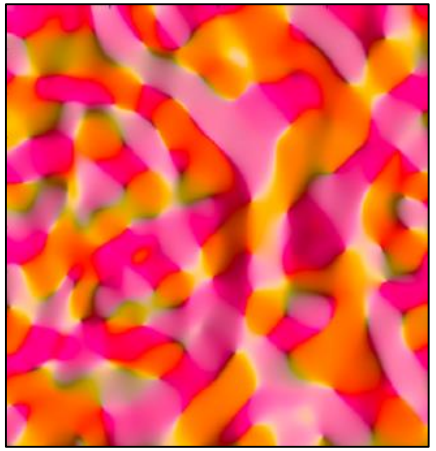
Network architecture (based on 3D U-net)



Learning non-rigid surface reconstruction from spatio-temporal image patches

GBR-invariant loss function

$$\frac{\nabla^2 z}{\|\nabla^2 z\|_F} = \frac{1}{\sqrt{{z_{xx}}^2 + {z_{xy}}^2 + {z_{yx}}^2 + {z_{yy}}^2}} \begin{pmatrix} z_{xx} & z_{xy} \\ z_{yx} & z_{yy} \end{pmatrix}$$

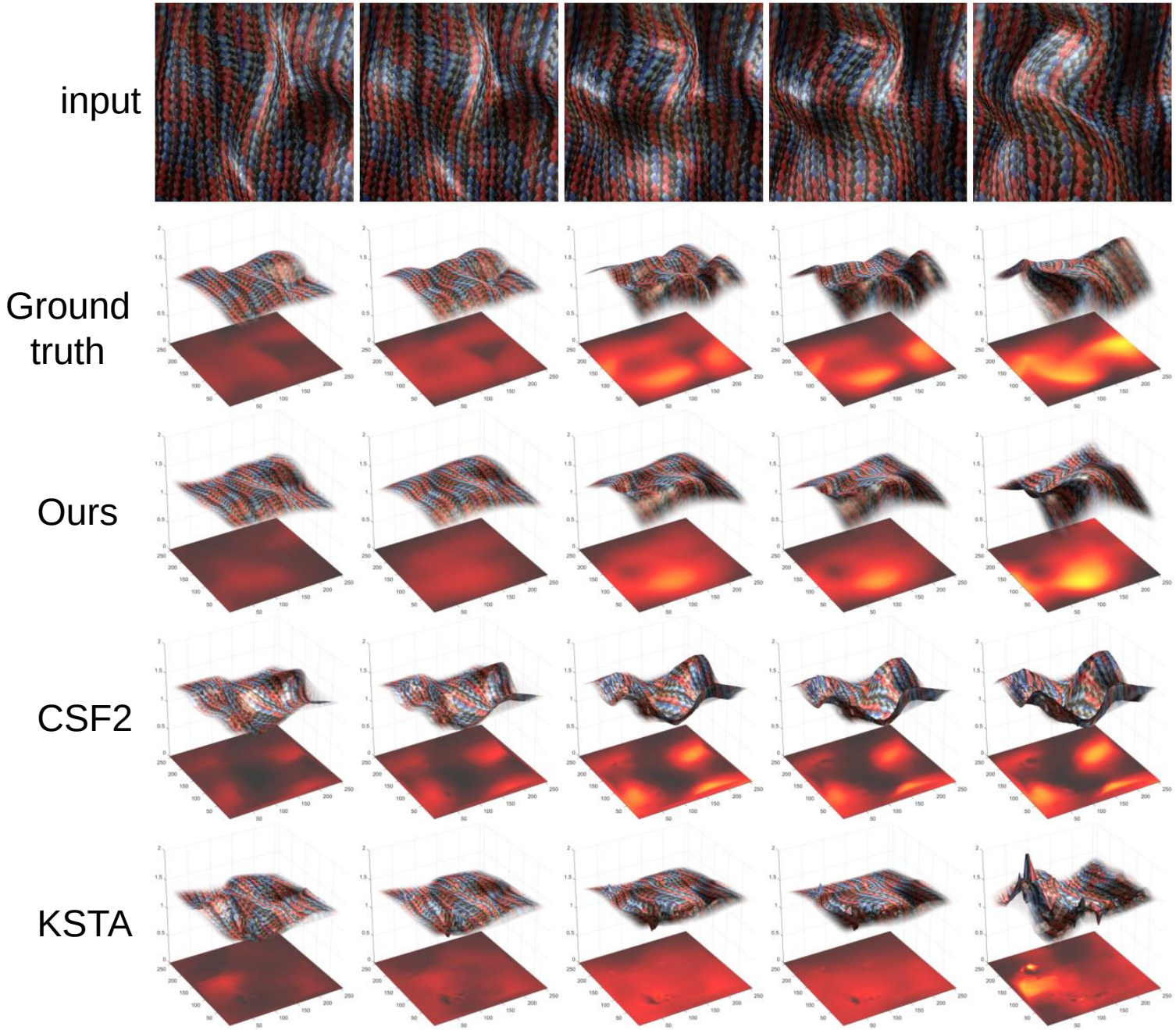


Visualized in *Lab* color space

- Each pixel of the GBR-invariant depth map has **two degrees of freedom** (…can be seen as points on the unit sphere in 3D space)
- Euclidean distance corresponds to chordal distance between points on the sphere

Results

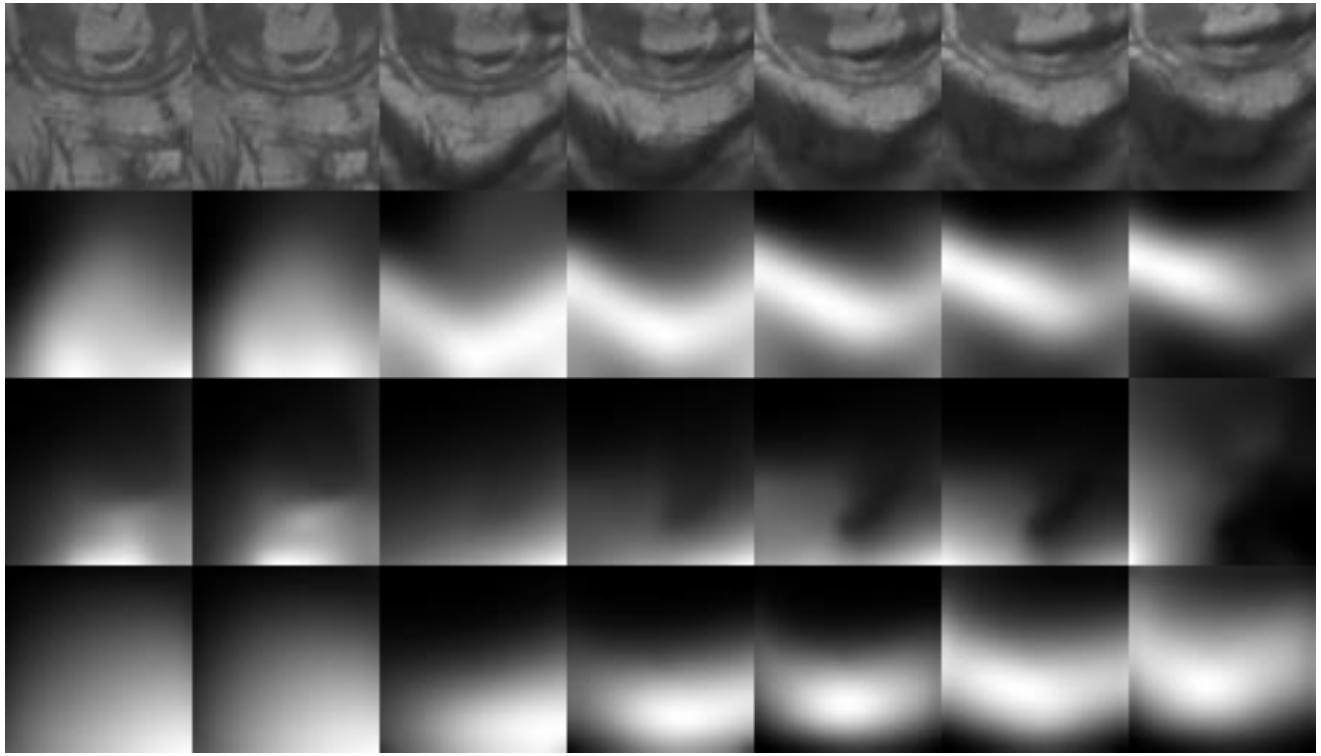
- **Synthetic data**
- **Different motion parameters than in training**
- **Comparison with two state-of-the-art NRSfM methods**



CSF2 P. F. U. Gotardo and A. M. Martinez. Kernel non-rigid structure from motion. In *2011 International Conference on Computer Vision*, pages 802–809, November 2011. doi: 10.1109/ICCV.2011.6126319.

KSTA Paulo F. U. Gotardo and Aleix M. Martinez. Kernel non-rigid structure from motion. In *2011 International Conference on Computer Vision*, pages 802–809, Barcelona, Spain, November 2011. IEEE. ISBN 978-1-4577-1102-2 978-1-4577-1101-5 978-1-4577-1100-8. doi: 10.1109/ICCV.2011.6126319.

Results (real data)



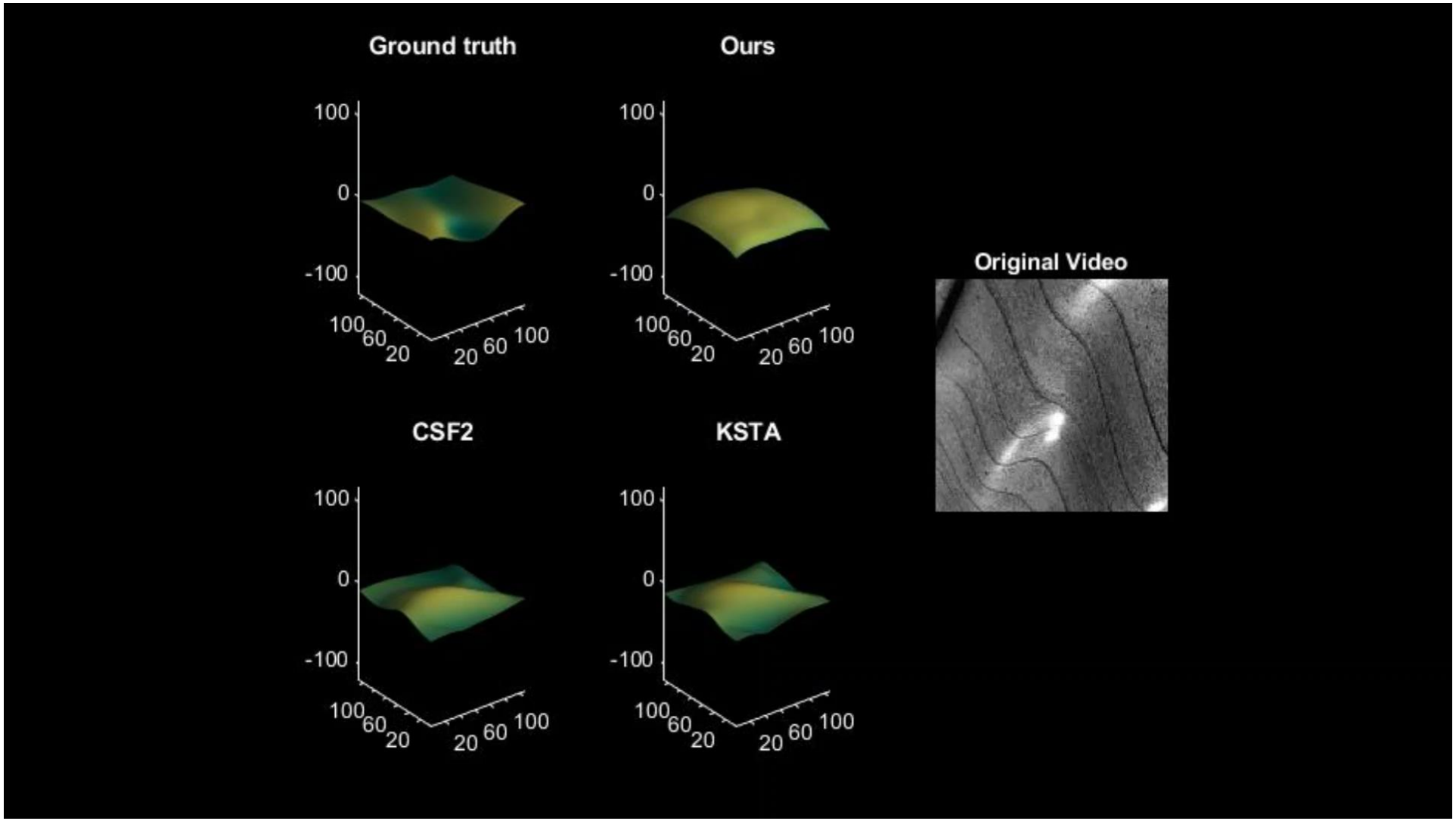
Kinect RGB sequence

Kinect Depth maps

CSF2

Ours

More results…



Quantitative results

Synthetically generated sequences		
Ours	CSF2	KSTA
0.5907 ± 0.4536	0.8746 ± 0.6372	0.8738 ± 0.6369

Average and standard deviation of *spatially normalized MAE* calculated from 1000 videos

Kinect sequences		
Ours	CSF2	KSTA
3.7 mm	4.6 mm	4.3 mm

Average *MAE* calculated from two Kinect sequences