3D Scene Representation Learning

Martin Oswald

Computer Vision Group, University of Amsterdam

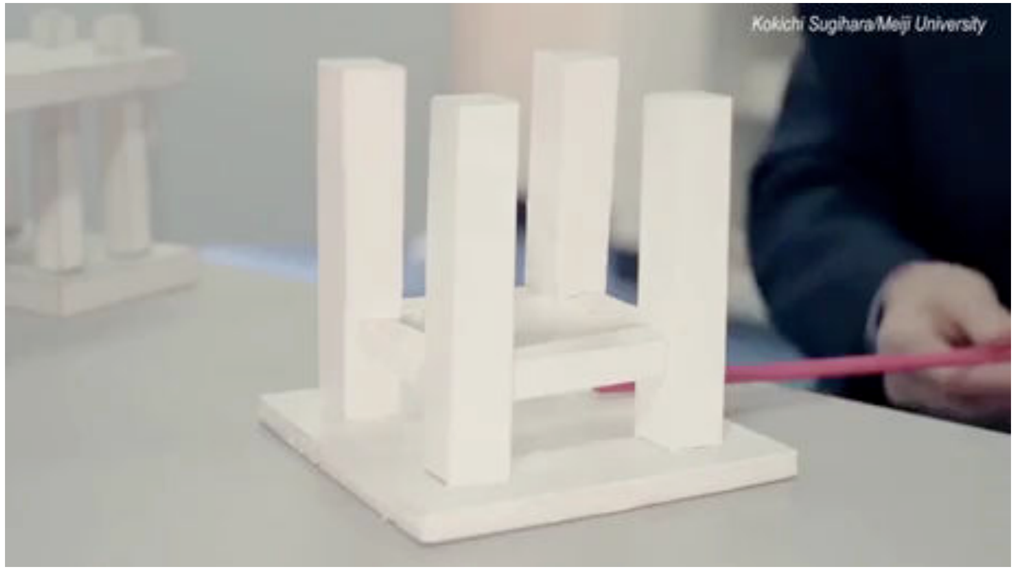
UNIVERSITY OF AMSTERDAM

ATLAS LAB.

Motivation: 3D reconstruction is hard!

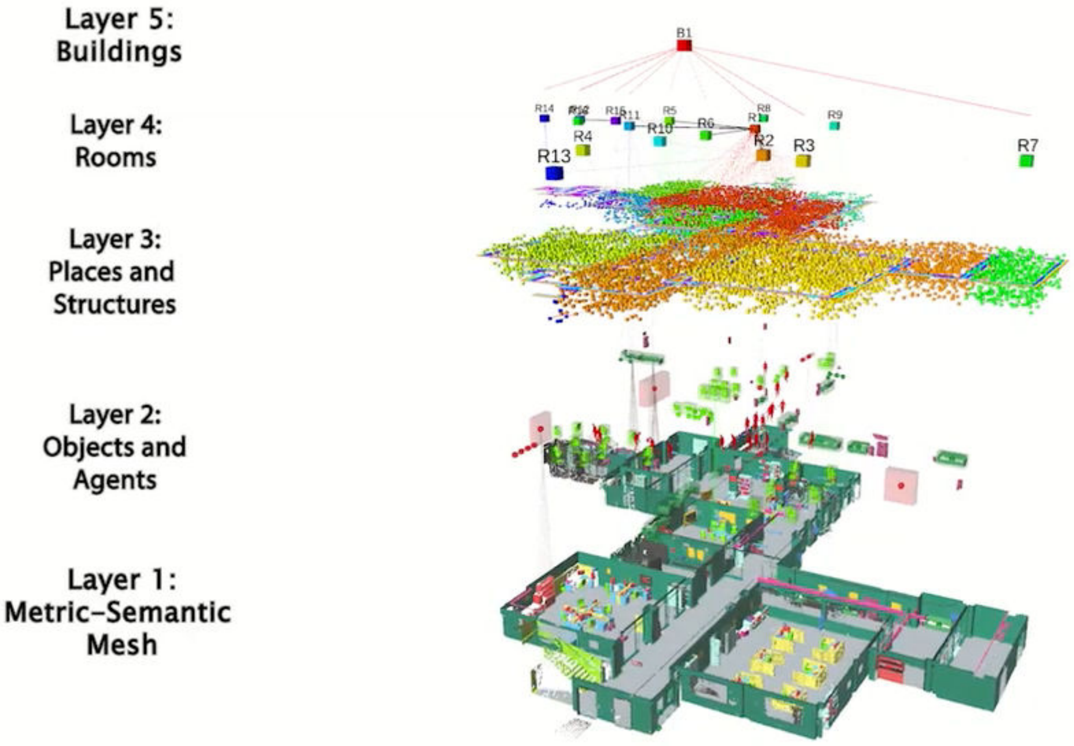
Motivation: 3D reconstruction is hard!

Motivation: 3D reconstruction is hard!



Scene Representation - Goals

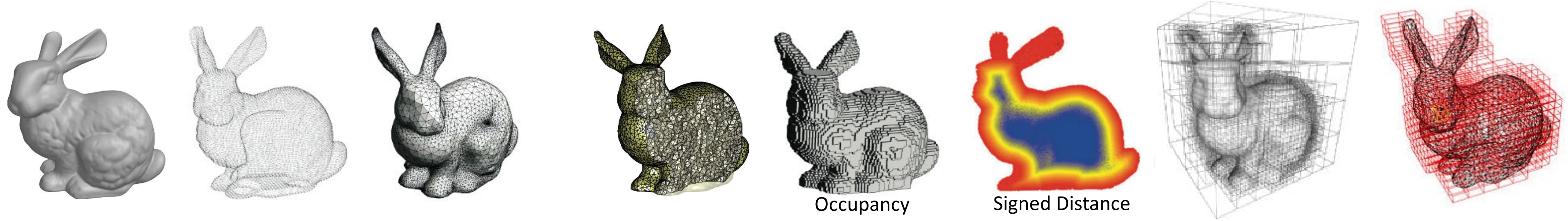
- Novel, learned scene representations for large-scale mapping
  - Flexible, scalable, efficient (storage / access)
  - Richer content (geometry, appearance, lighting, semantics, instances, physics, materials, dynamics, functionality, actions, natural language)
  - Hierarchical content
  - Task agnostic, multi-task use
  - Suitable for online updates
  - Suitable for sensor fusion / collaborative editing



[Rosinol et al. 3D Dynamic Scene Graphs: Actionable SpatialPerception with Places, Objects, and Humans, RSS 2020]

Scene Representations

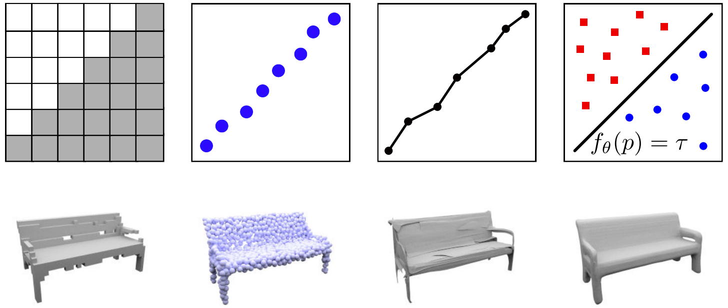
[https://arxiv.org/pdf/1803.03352.pdf]



Spline/NURBS | Point Cloud | Surface Mesh | Tetrahedral Mesh | Voxel Grid | Voxel Octree | Voxel Hashing

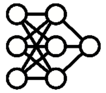
explicit **(topology change=hard)** | **implicit** **(topology change=simple)**

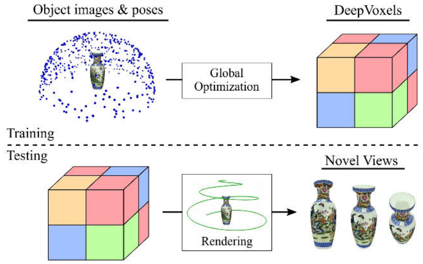
Traditional



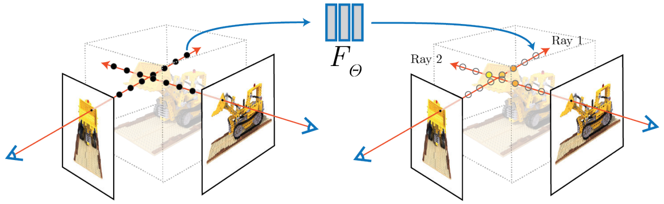
Voxel Grid | Point Cloud | Mesh | (Neural) Classifier

Learned / Deep Representations:
OccNet [https://arxiv.org/pdf/1812.03828.pdf]
DeepSDF [https://arxiv.org/pdf/1901.05103.pdf]
IM-Net [https://arxiv.org/pdf/1812.02822.pdf]

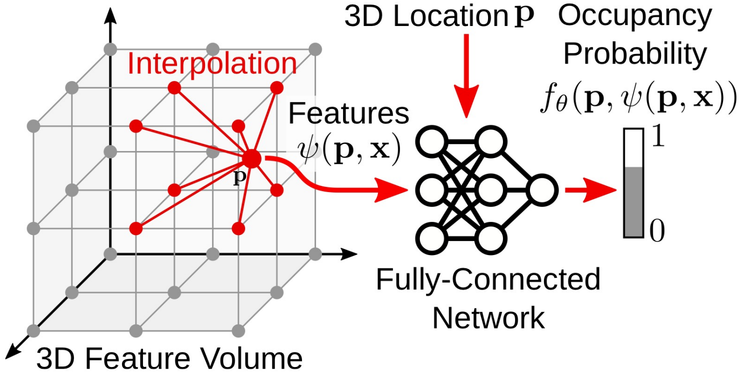




[DeepVoxels CVPR 2019]



[NeRF - Neural Radiance Fields, ECCV 2020]



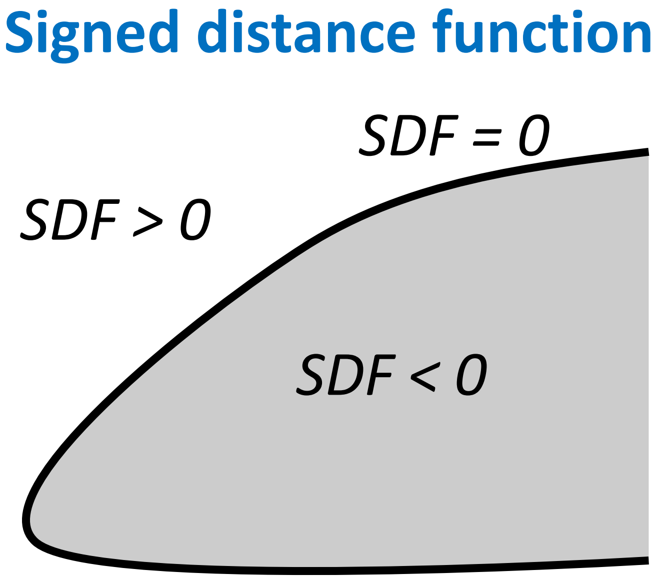
[Peng et al., Convolutional Occupancy Networks, ECCV 2020]

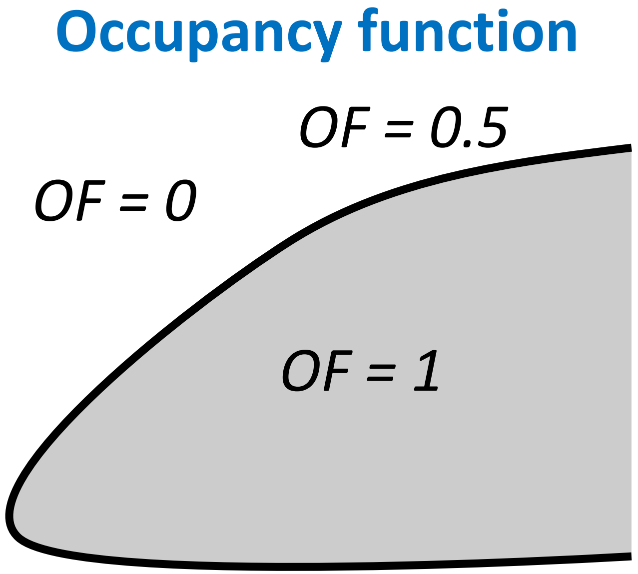
Learned

Neural implicit

Implicit Volumetric Representation

- *Voxel grid*: sample a volume containing the surface of interest uniformly
- Label each grid point as lying *inside* or *outside* the surface

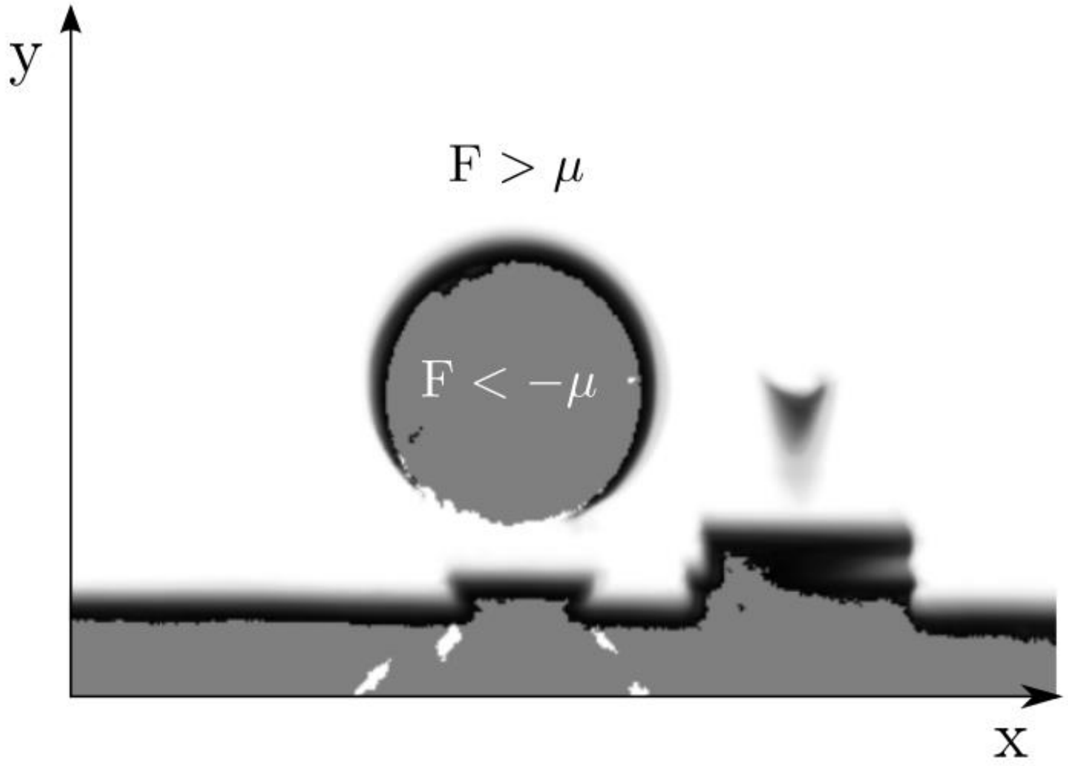


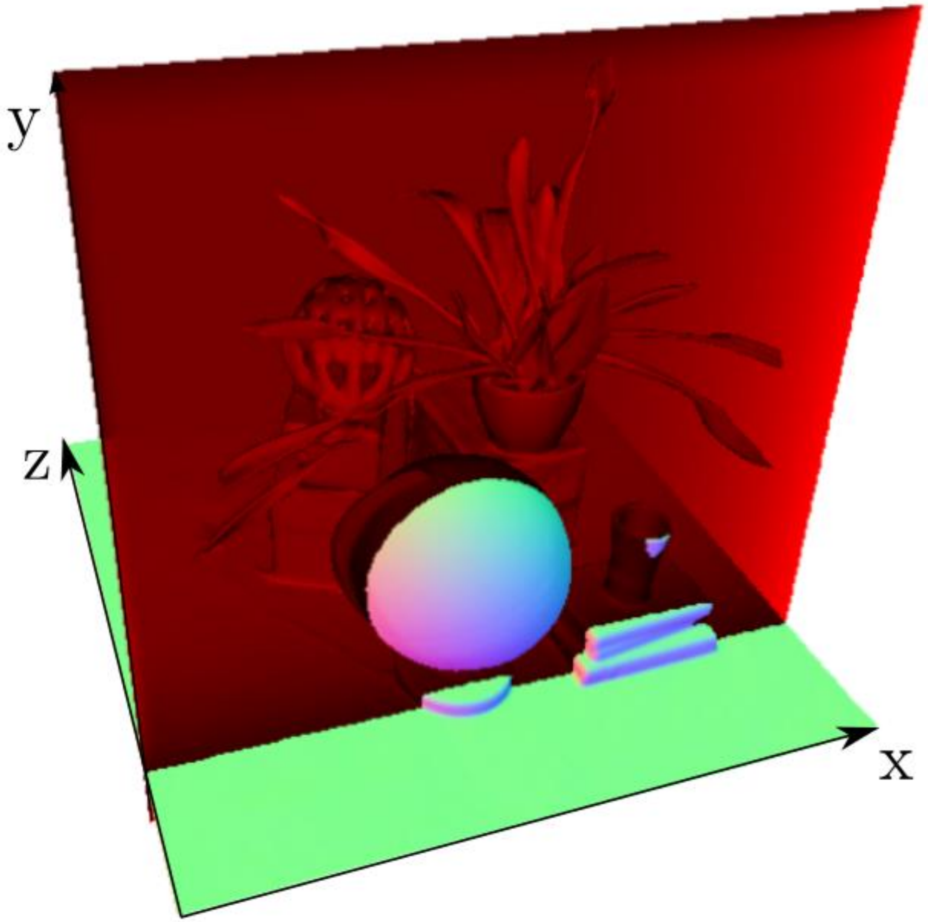


- The modeled surface is represented as an *isosurface* (e.g. SDF = 0 or OF = 0.5) of the labeling (implicit) function

- Advantages: simple handling of topological changes, watertight surfaces, no self-occlusions
  Disadvantages: Large memory requirement, bad scalability to large scenes (cubic growth)

Represent Scenes with TSDFs

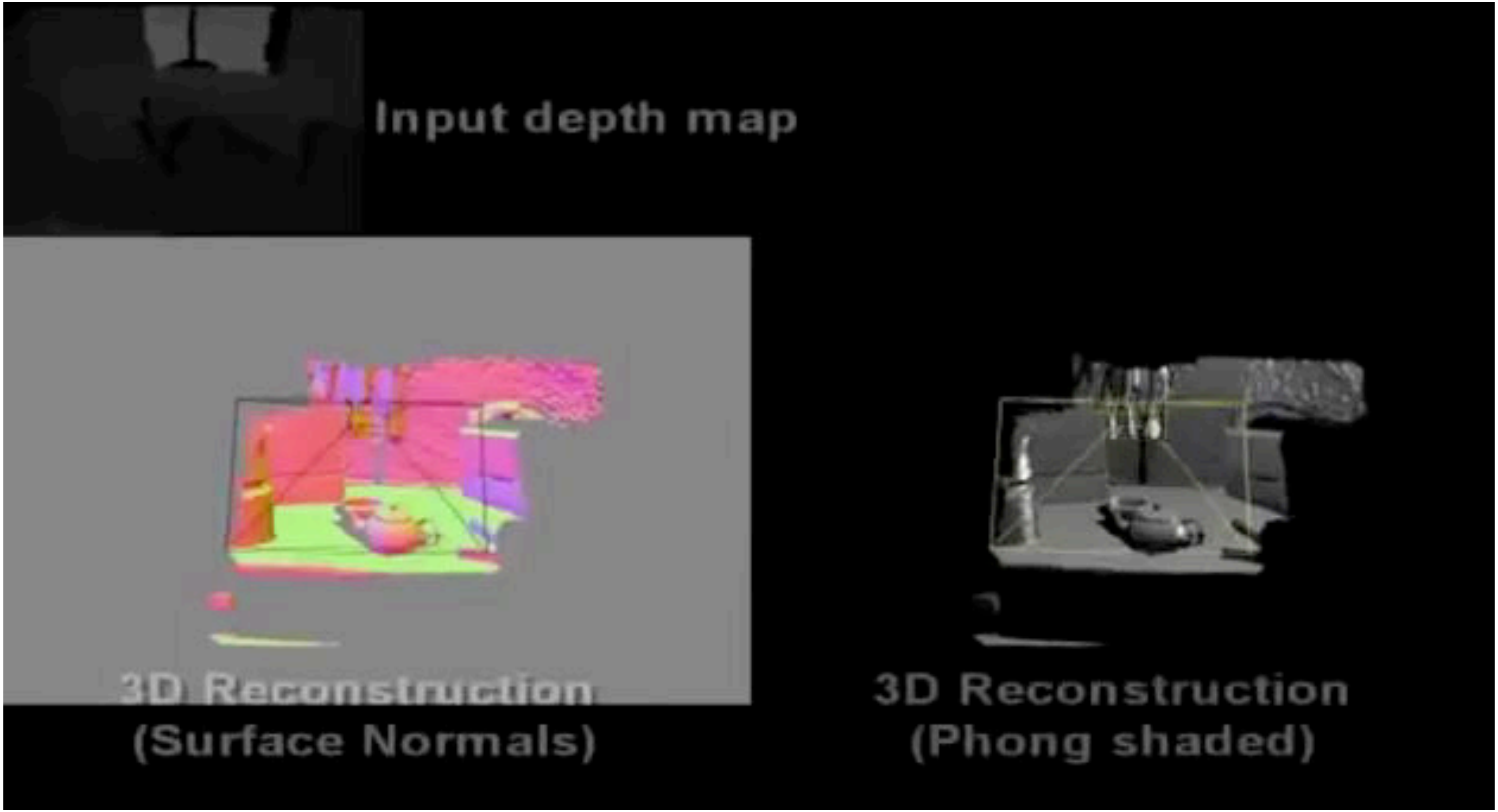




[Newcombe & Lovegrove, Geometric Reconstruction Lecture]

Real-time Mapping - KinectFusion

[Newcombe et al, ISMAR 2011]



Real-time Mapping

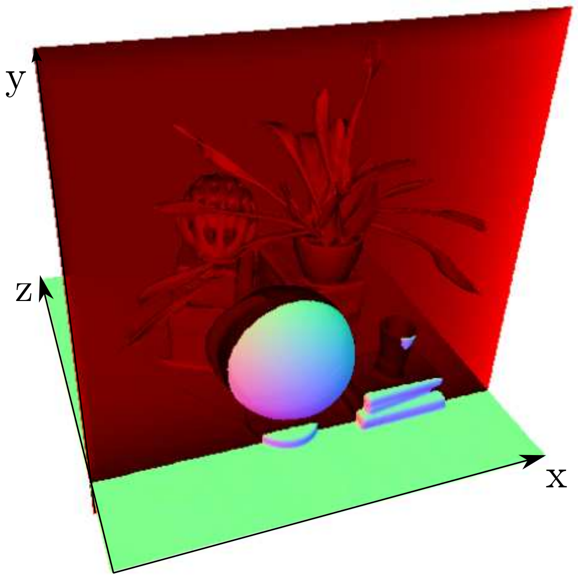
Baseline: Depth fusion with Truncated signed distance functions (TSDFs)

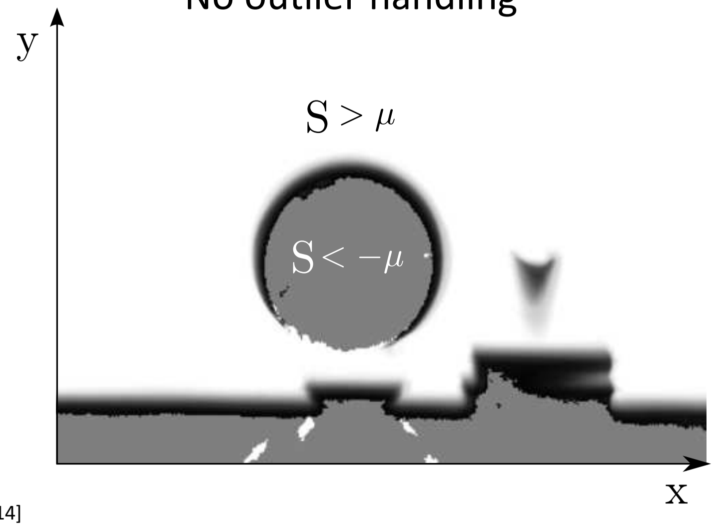
Advantages

- Local updates on grid (const. time)
- Simple online updates, noise removal
- Highly parallelizable
- Real-time

Disadvantages

- Not well suited for non-Gaussian, non-zero-mean noise (often depth dependent)
- Minimal surface thickness according to expected noise level
- Does not work for thin surfaces, updates might cancel out each other for opposing views
- Noise level is assumed to be directional independent (but depends on viewing direction)
- No outlier handling





[Richard Newcombe, PhD Thesis, Imperial College London, 2014]

TSDF Fusion

Learned Fusion

TSDF Fusion

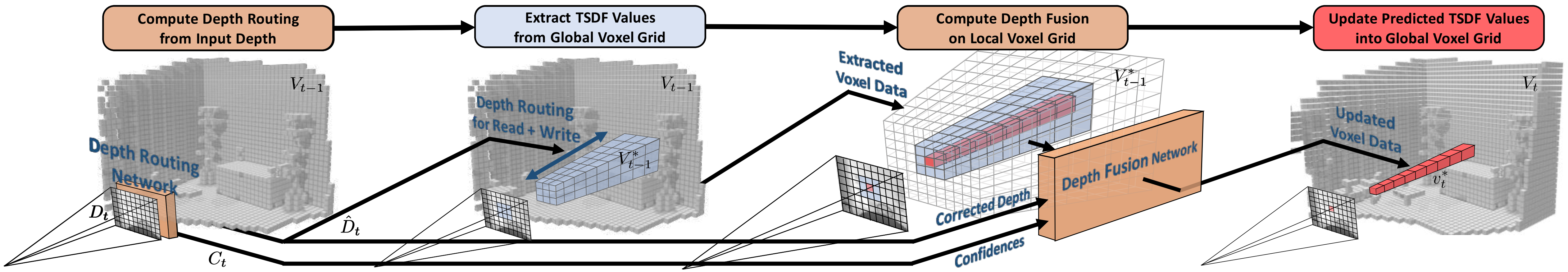
$$\boldsymbol{V}_t(\boldsymbol{x}) = \frac{\boldsymbol{W}_{t-1}(\boldsymbol{x}) \cdot \boldsymbol{V}_{t-1}(\boldsymbol{x}) + w_t(\boldsymbol{x}) \cdot v_t(\boldsymbol{x})}{\boldsymbol{W}_{t-1}(\boldsymbol{x}) + w_t(\boldsymbol{x})}$$

$$\boldsymbol{W}_t(\boldsymbol{x}) = \boldsymbol{W}_{t-1}(\boldsymbol{x}) + w_t(\boldsymbol{x}) \ ,$$

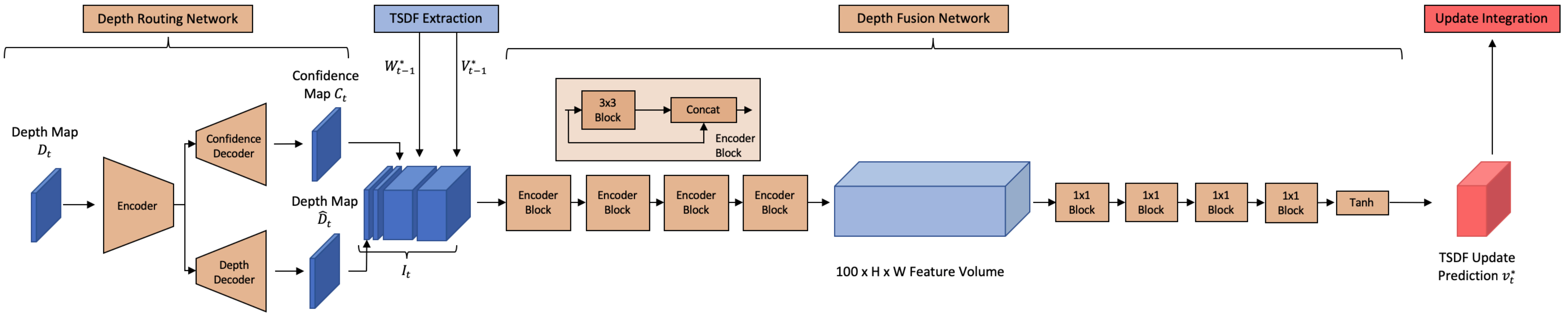
RoutedFusion: Learning Depth Map Fusion

[Weder, Schoenberger, Pollefeys, Oswald, CVPR 2020]

System Overview

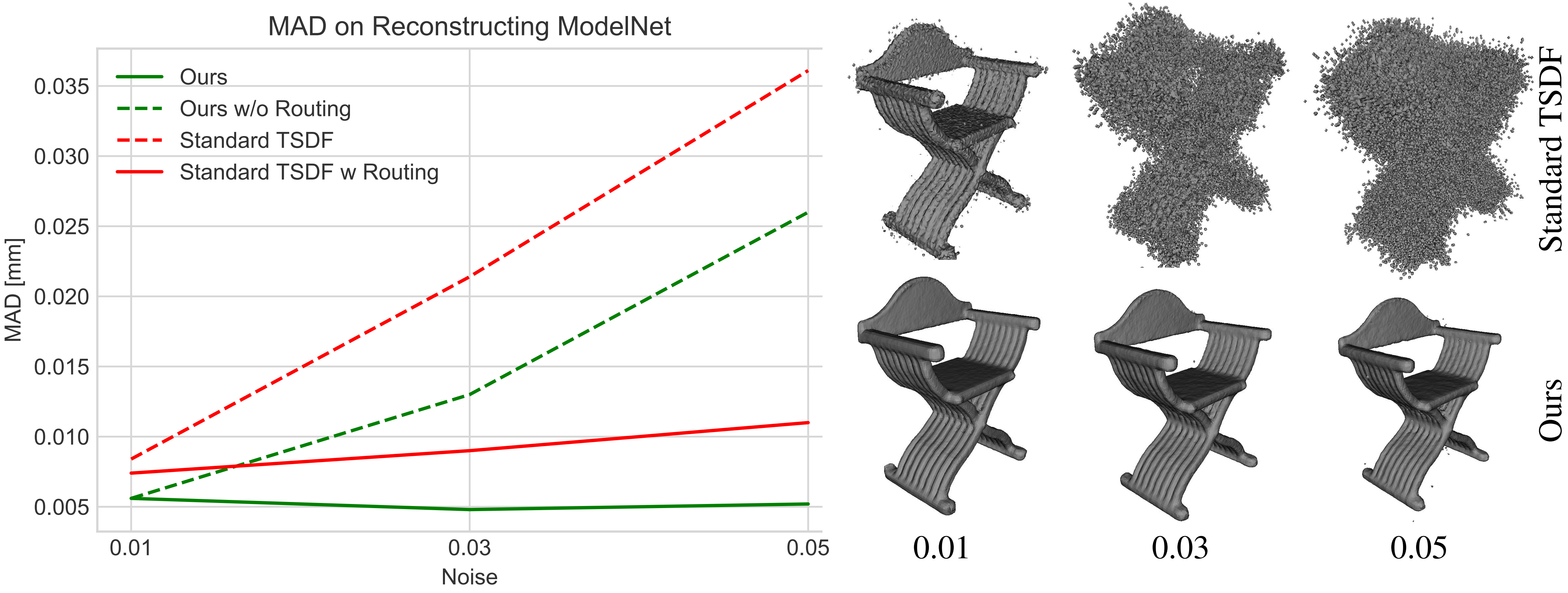


Network Architecture



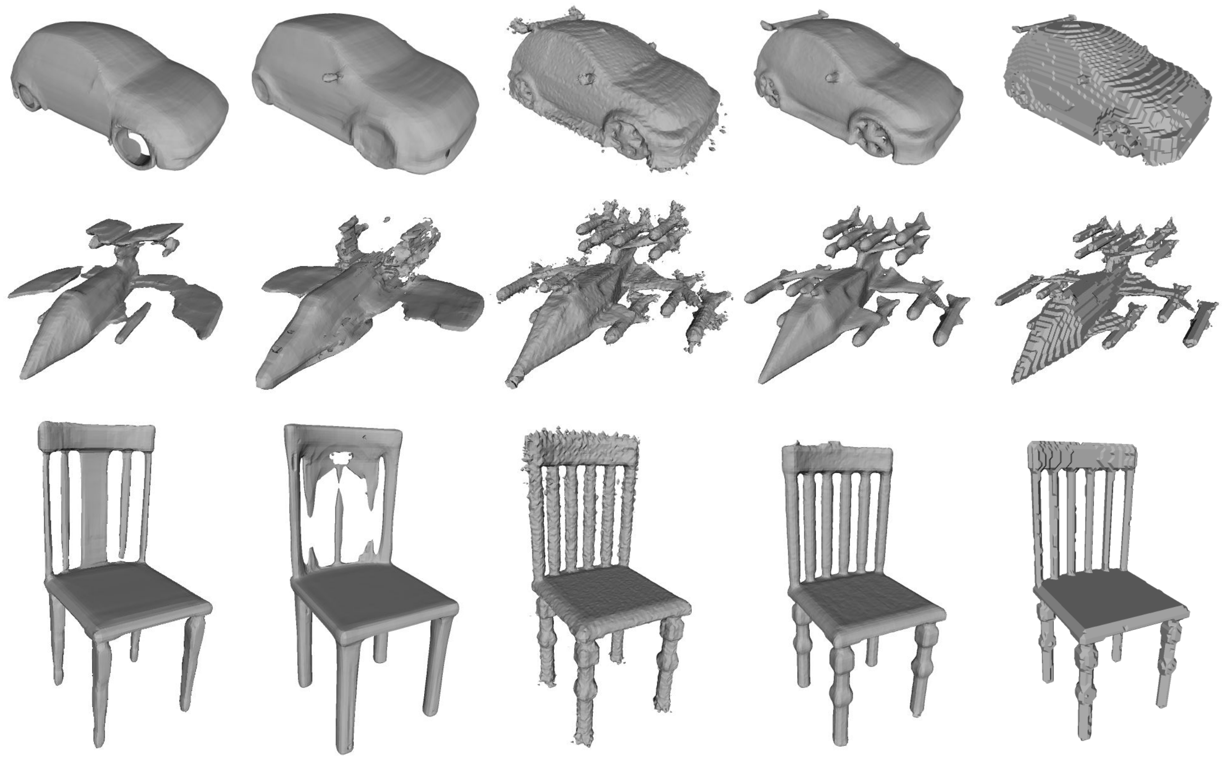
RoutedFusion: Learning Depth Map Fusion

[Weder, Schoenberger, Pollefeys, Oswald, CVPR 2020]

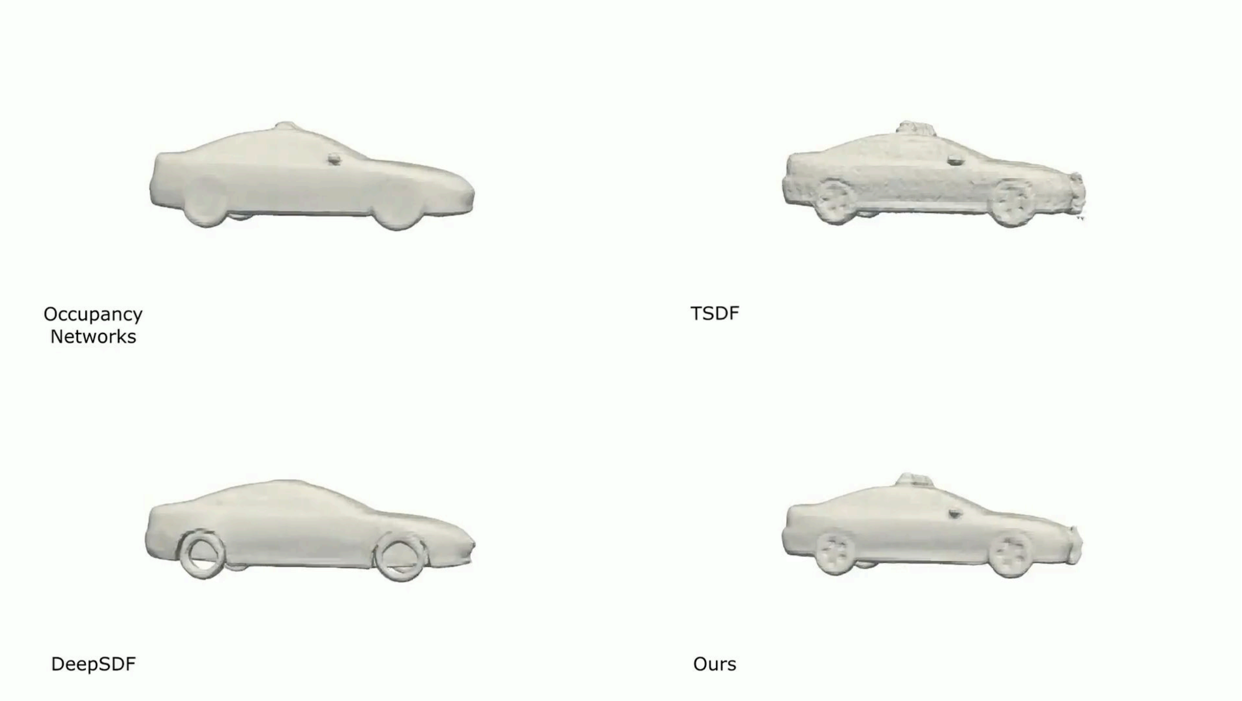


RoutedFusion: Learning Depth Map Fusion

[Weder, Schoenberger, Pollefeys, Oswald, CVPR 2020]



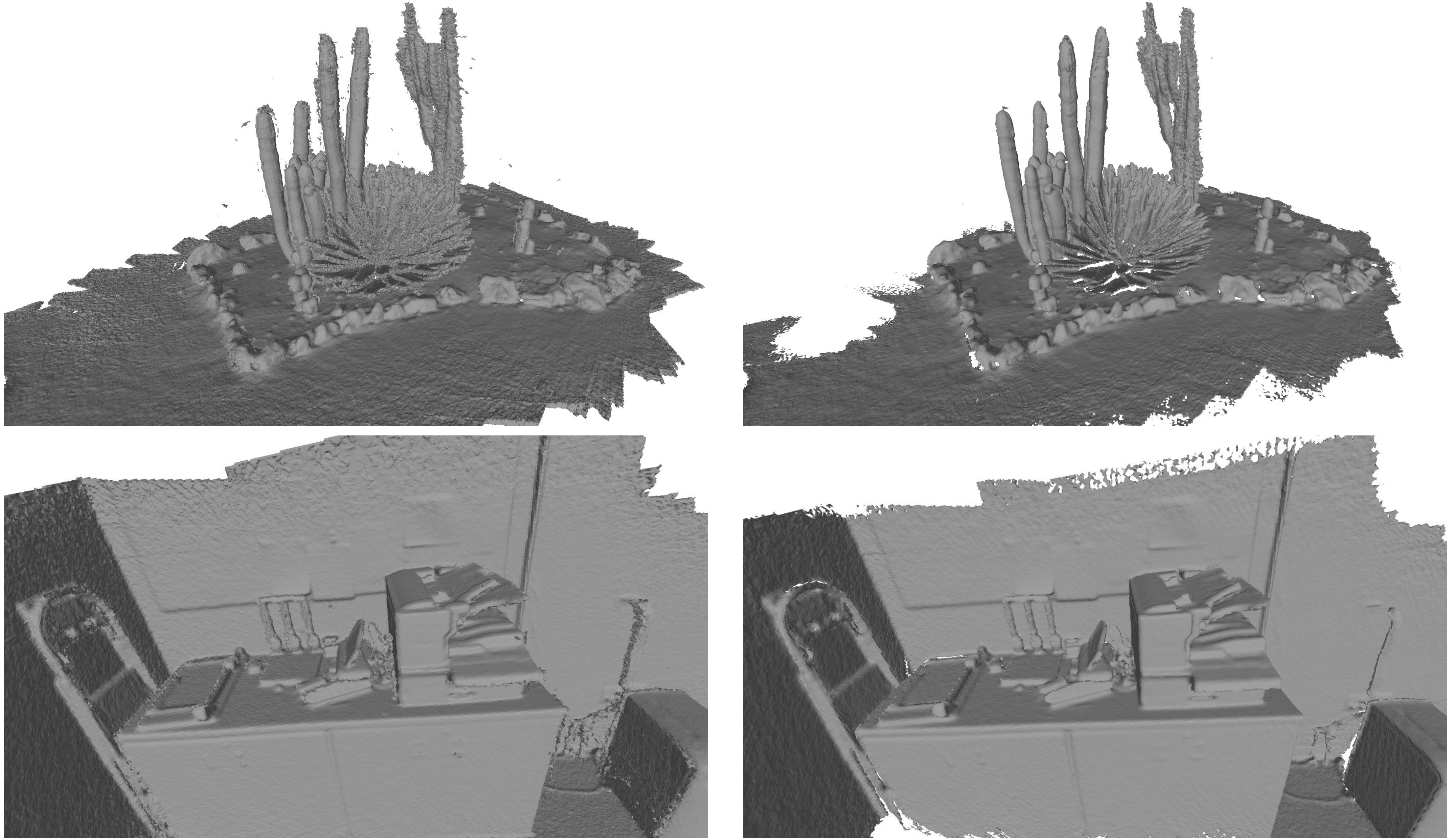
DeepSDF [38] Occ.Net. [35] TSDF [9] Ours GT



Method	MSE [e-05]	MAD	Acc. [%]	IoU [0, 1]
DeepSDF [38]	464.0	0.0499	66.48	0.538
OccupancyNetworks [35]	56.8	0.0166	85.66	0.484
TSDF Fusion [9]	11.0	0.0078	88.06	0.659
TSDF Fusion + Routing	27.0	0.0084	87.48	0.650
Ours w/o Routing	**5.9**	0.0051	93.91	0.765
Ours	**5.9**	**0.0050**	**94.77**	**0.785**

RoutedFusion: Learning Depth Map Fusion

[Weder, Schoenberger, Pollefeys, Oswald, CVPR 2020]

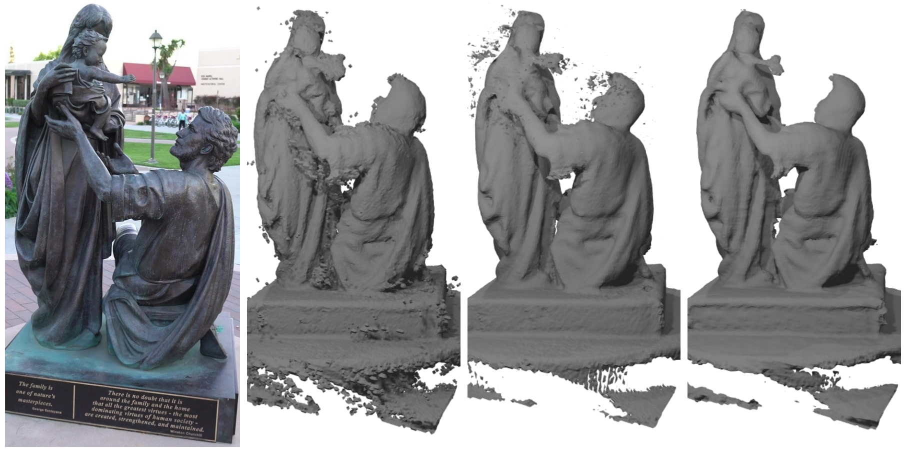


NeuralFusion: Depth Fusion in Latent Space

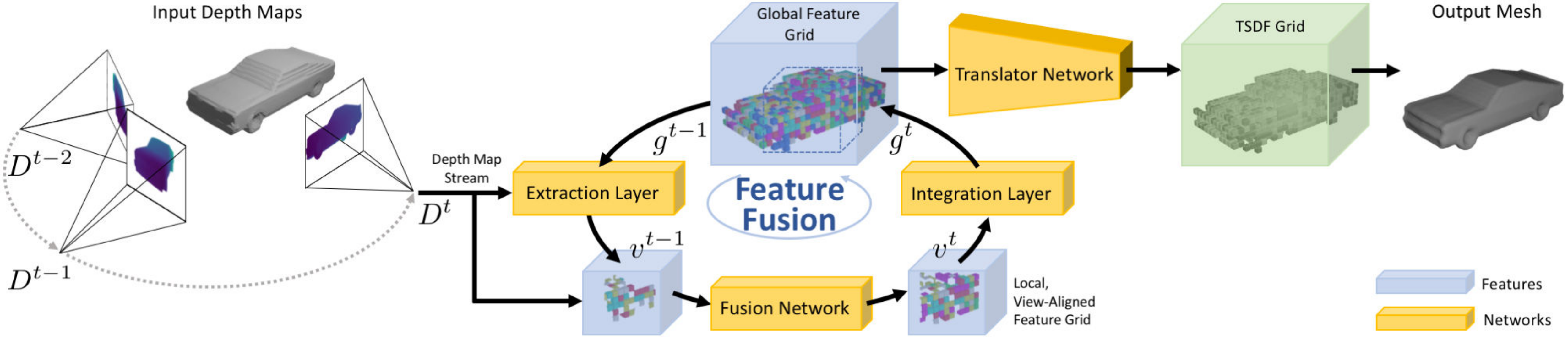
[Weder, Schoenberger, Pollefeys, Oswald, CVPR 2021]

Major Problem: Outlier handling

Difficulty in **online** fusion:
1st measurement vs. outlier



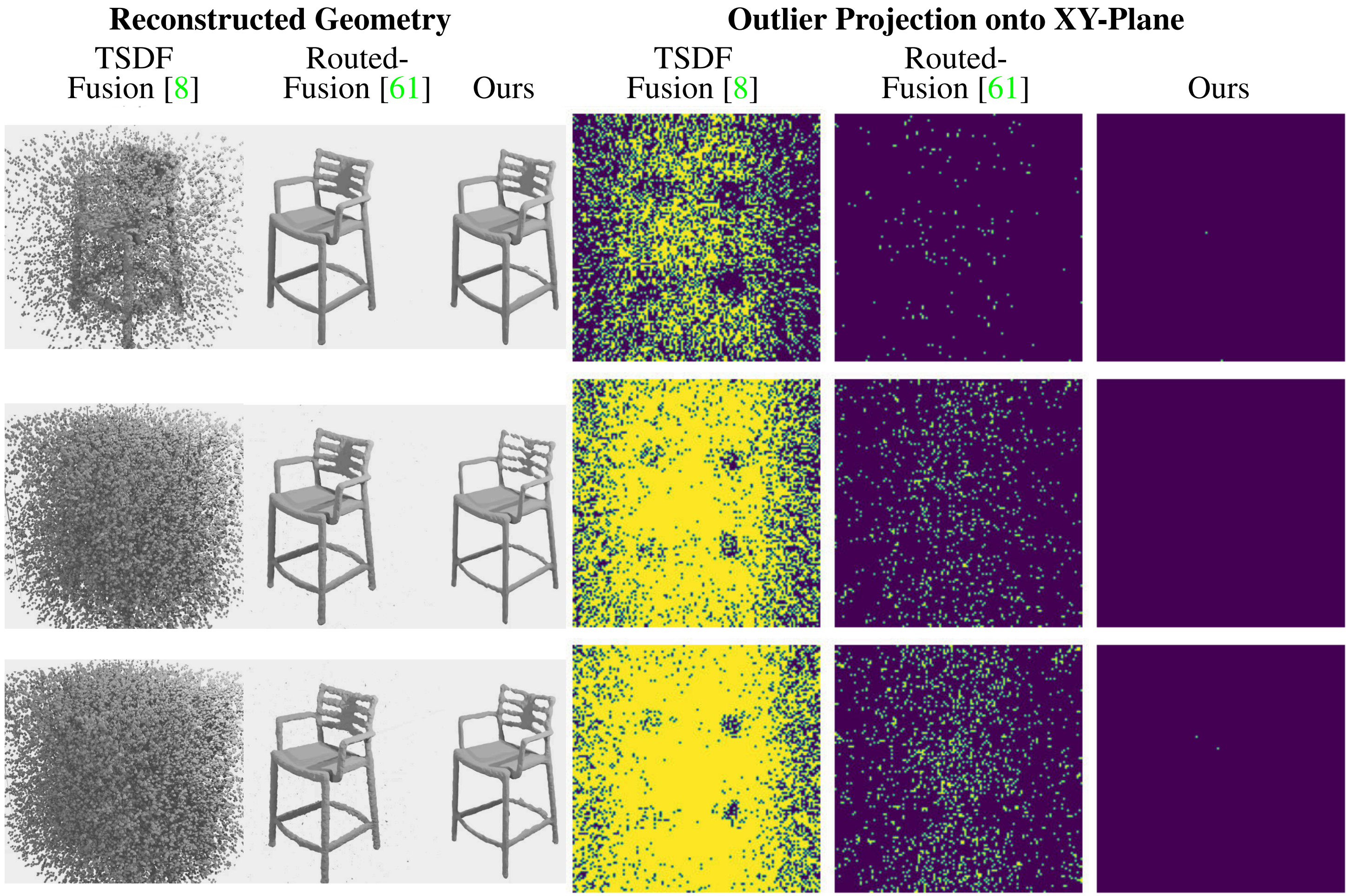
Input TSDF Fusion RoutedFusion NeuralFusion



NeuralFusion: Depth Fusion in Latent Space

[Weder, Schoenberger, Pollefeys, Oswald, CVPR 2021]

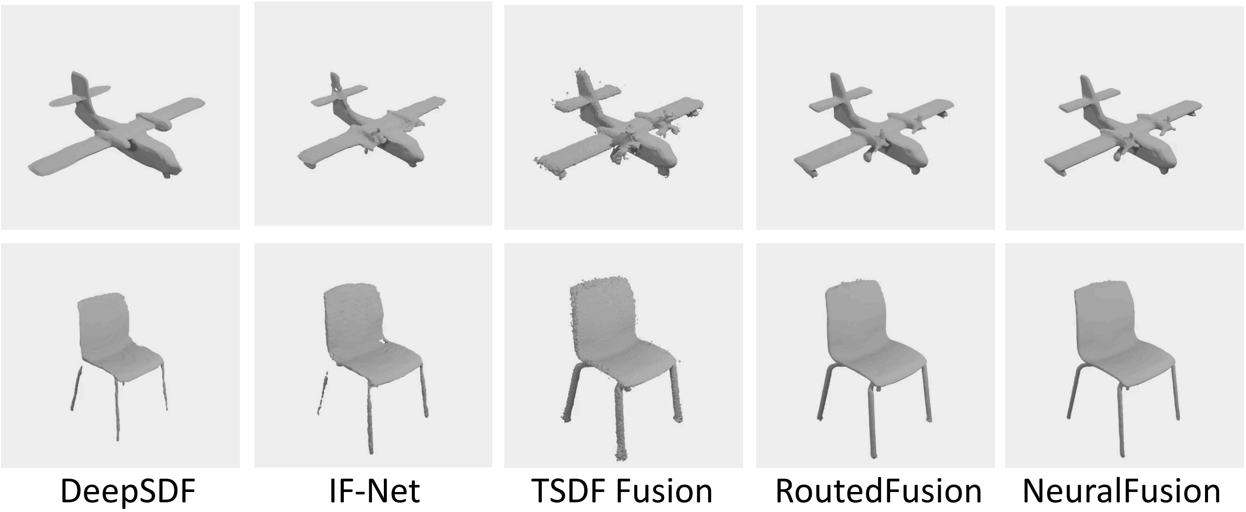
Outlier Fraction	Method	MSE↓ [e-05]	MAD↓ [e-02]	Acc.↑ [%]	IoU↑ [0,1]
0.01	TSDF Fusion	34.51	1.17	85.17	0.645
	Routed-Fusion	5.43	0.57	95.21	0.837
	Ours	**2.27**	**0.29**	**97.57**	**0.884**
0.05	TSDF Fusion	80.72	2.02	73.86	0.432
	Routed-Fusion	9.84	0.68	94.46	0.803
	Ours	**4.91**	**0.22**	**98.05**	**0.851**
0.1	TSDF Fusion	102.50	2.43	67.47	0.341
	Routed-Fusion	14.25	0.77	92.95	0.764
	Ours	**3.35**	**0.22**	**98.48**	**0.865**

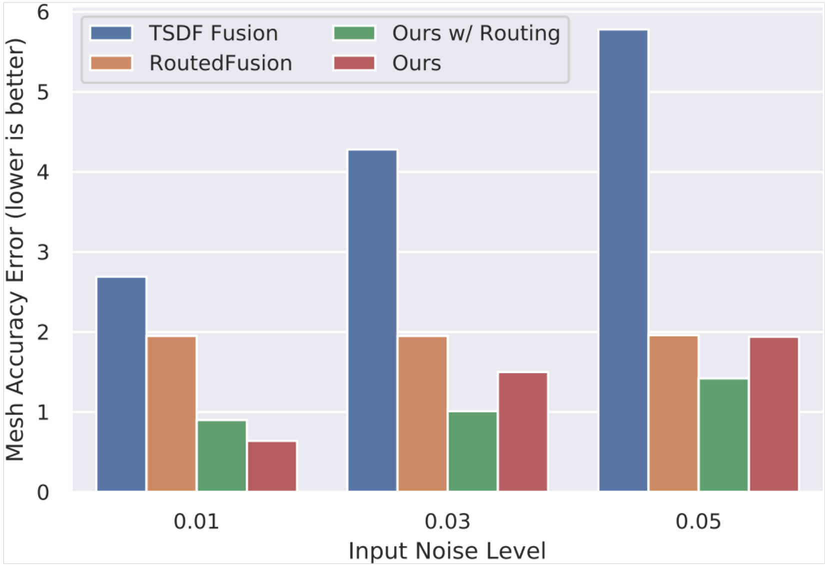


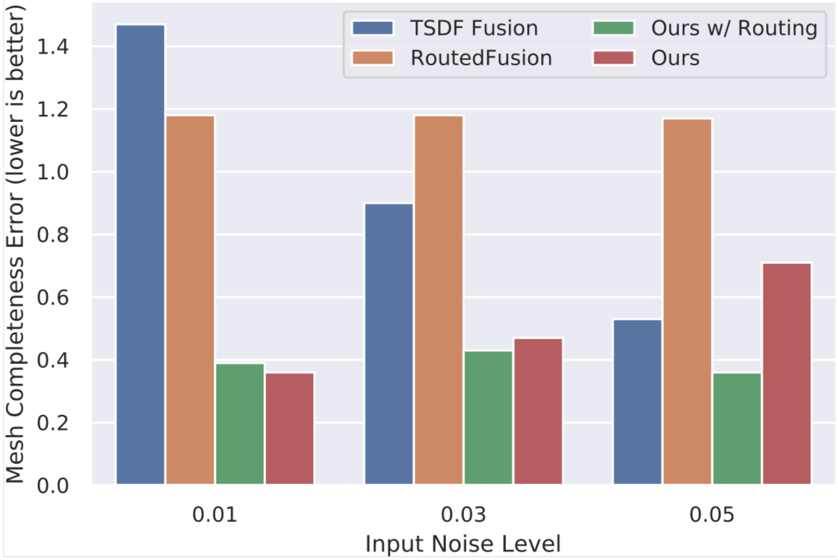
NeuralFusion: Depth Fusion in Latent Space

[Weder, Schoenberger, Pollefeys, Oswald, CVPR 2021]

Method	MSE↓ [e-05]	MAD↓ [e-02]	Acc.↑ [%]	IoU↑ [0,1]	F1↑ [0,1]
DeepSDF [42]	464.0	4.99	66.48	0.538	0.66
Occ.Net. [37]	56.8	1.66	85.66	0.484	0.62
IF-Net [7]	6.2	0.47	93.16	0.759	0.86
TSDF Fusion [8]	11.0	0.78	88.06	0.659	0.79
TSDF + 2D denoising	27.0	0.84	87.48	0.650	0.78
TSDF + 3D denoising	8.2	0.61	94.76	0.816	0.89
RoutedFusion [61]	5.9	0.50	94.77	0.785	0.87
Ours	**2.9**	**0.27**	**97.00**	**0.890**	**0.94**



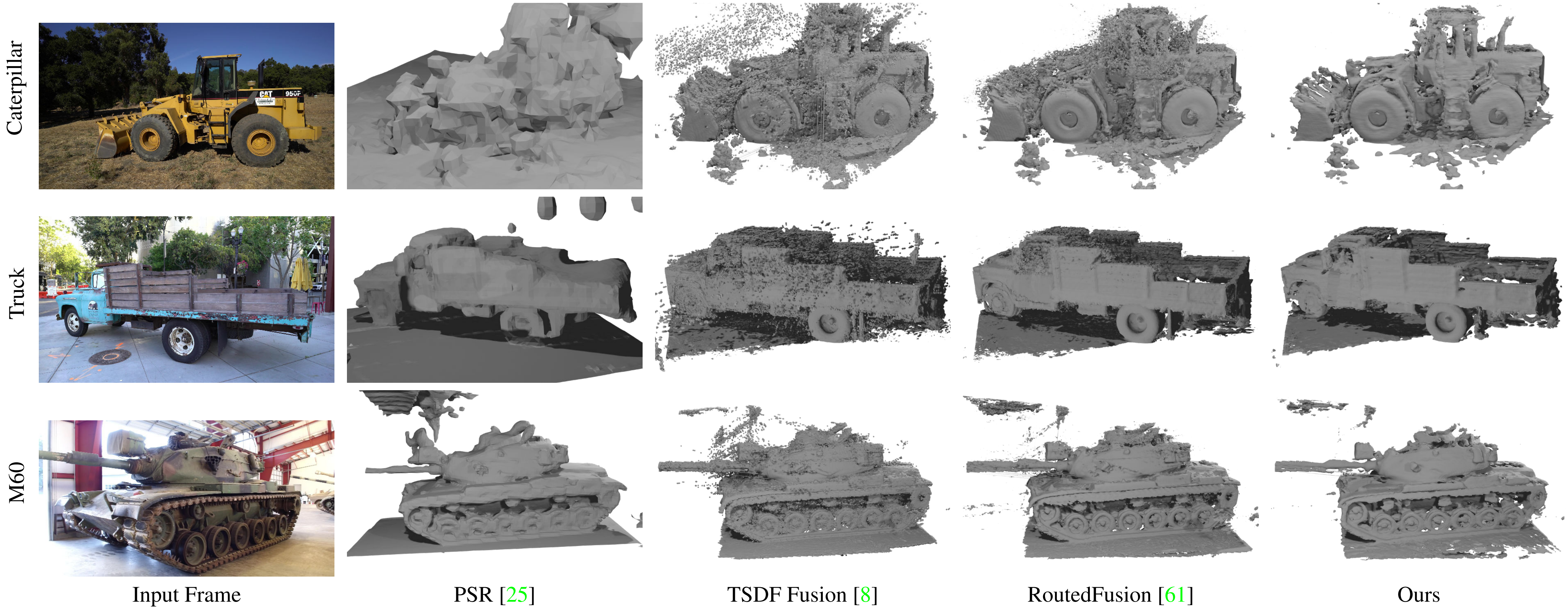




N	MSE↓ [e-05]	MAD↓ [e-02]	Acc.↑ [%]	IoU↑ [0,1]
1	-	-	-	-
2	9.45	0.64	94.67	0.717
4	4.03	0.30	97.51	0.863
8	3.99	0.29	97.46	0.862
16	3.91	0.29	97.50	0.863

NeuralFusion: Depth Fusion in Latent Space

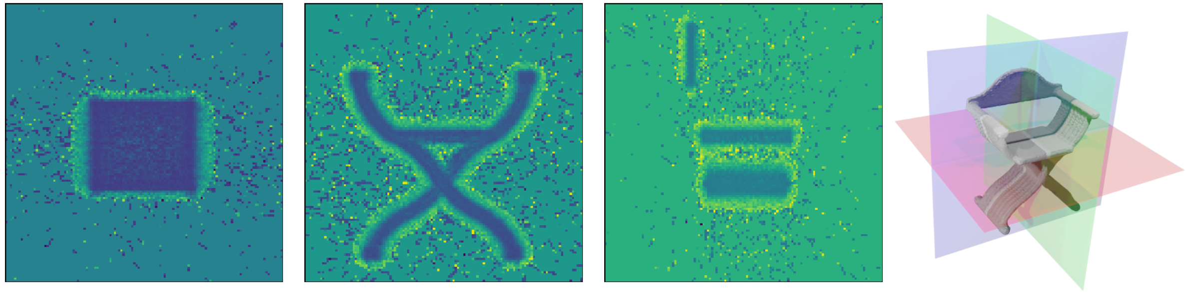
[Weder, Schoenberger, Pollefeys, Oswald, CVPR 2021]



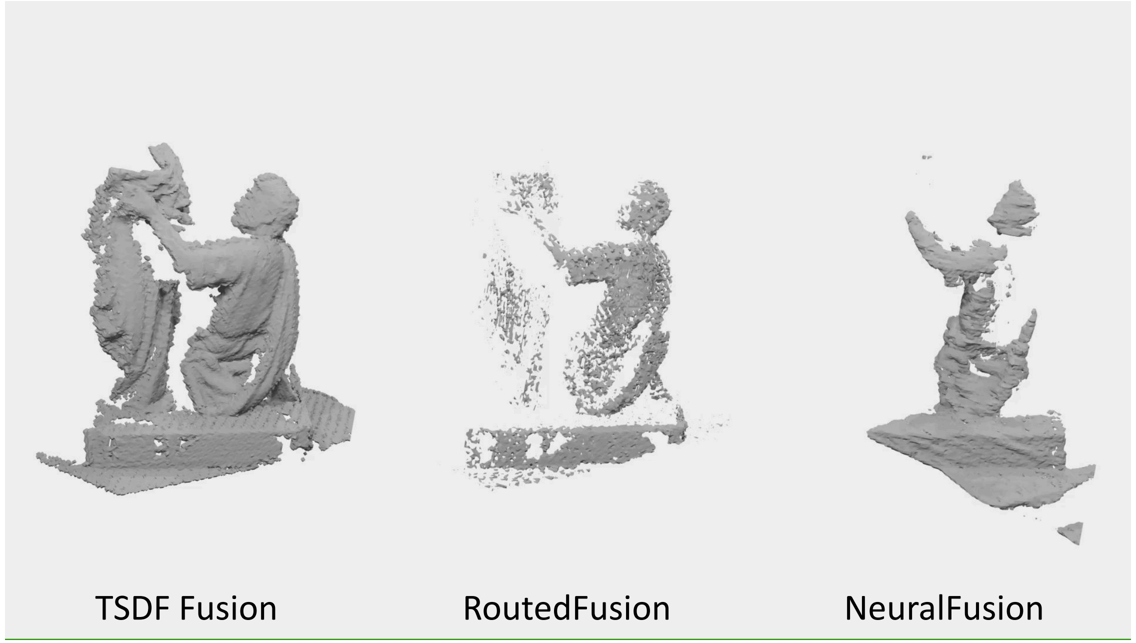
NeuralFusion: Depth Fusion in Latent Space

[Weder, Schoenberger, Pollefeys, Oswald, CVPR 2021]

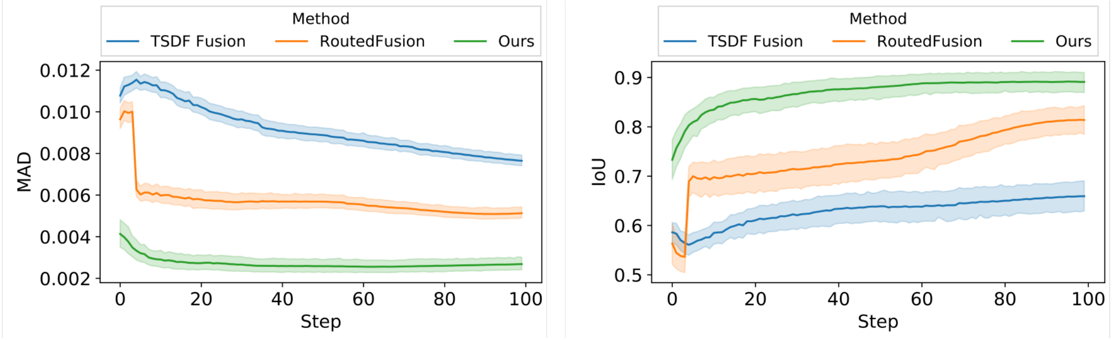
Latent Space



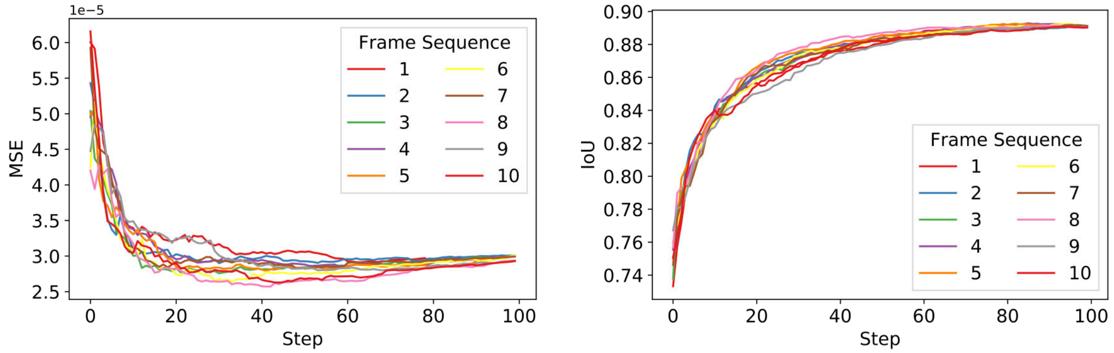
XY-Plane XZ-Plane YZ-Plane Output Mesh



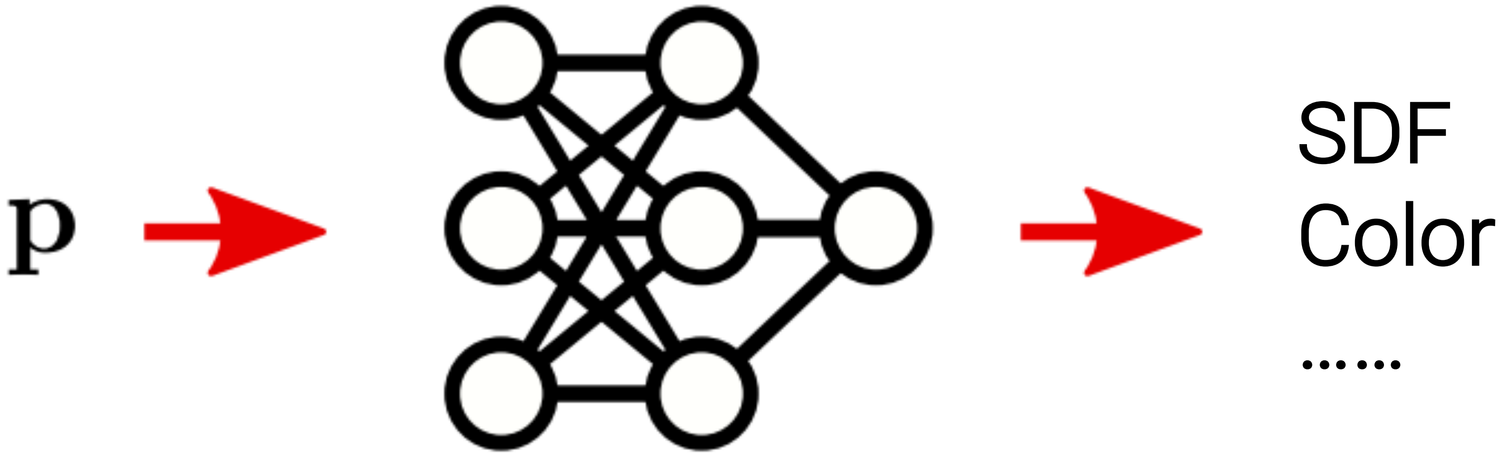
Iterative fusion performance



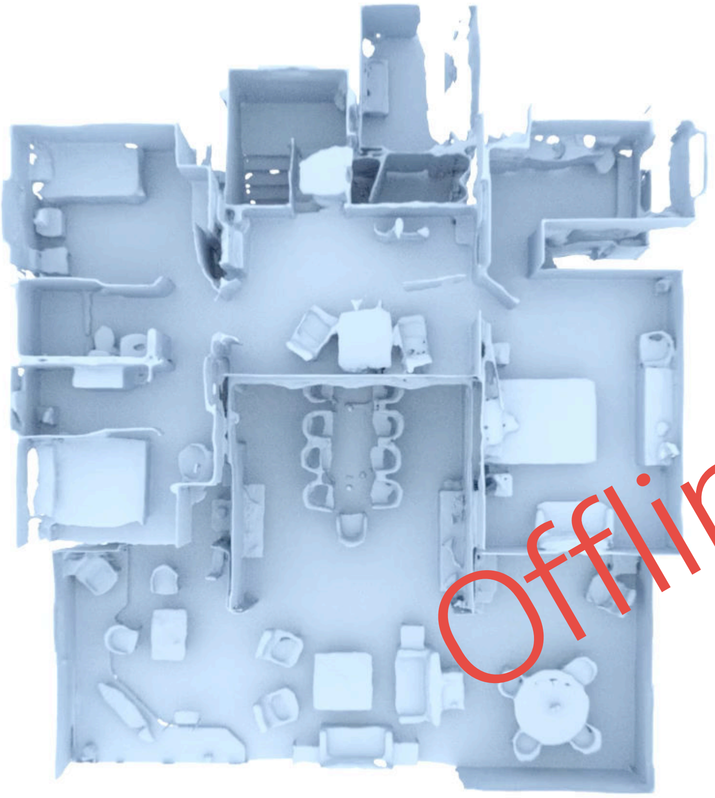
Random frame order



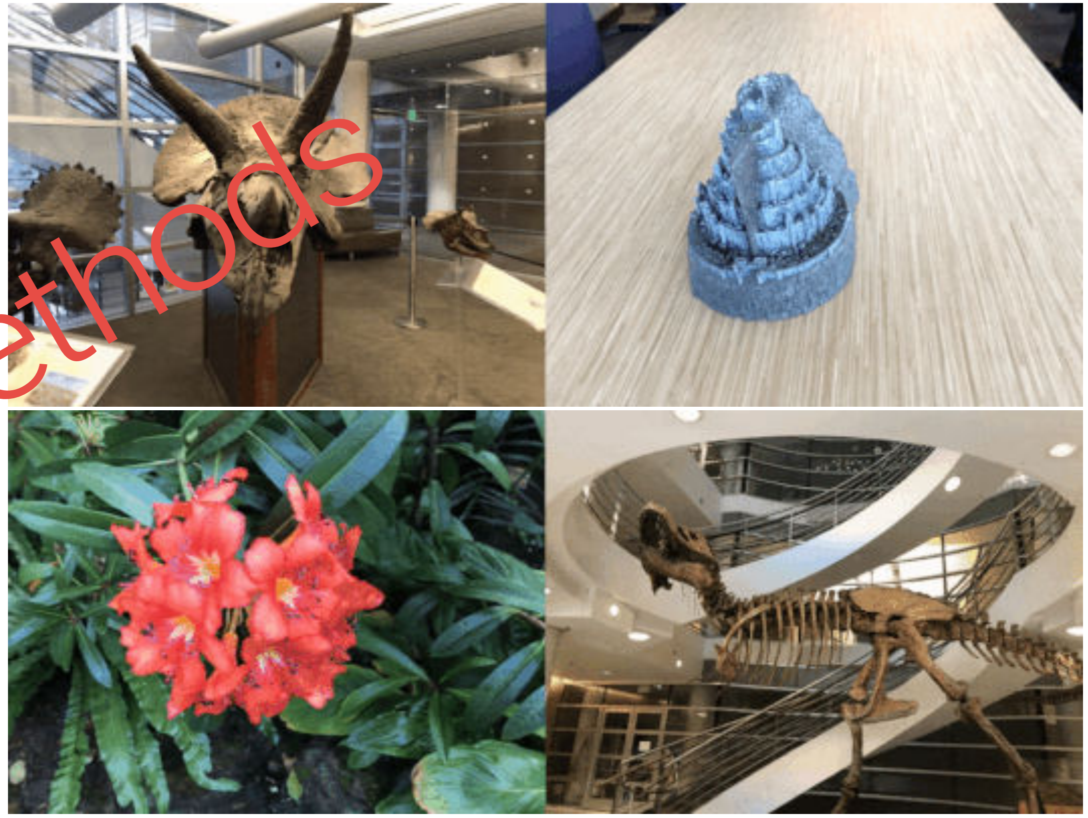
Neural Implicit Representations



Neural Implicit Representations



Offline Methods



ConvONet [Peng et al., ECCV'20]

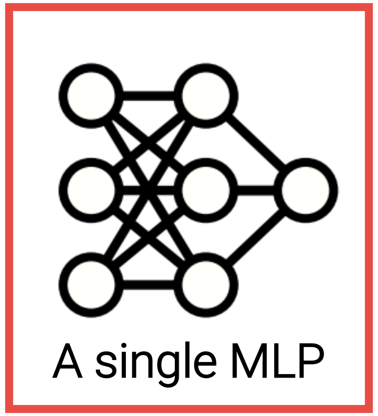
NeRF [Mildenhall et al., ECCV'20]

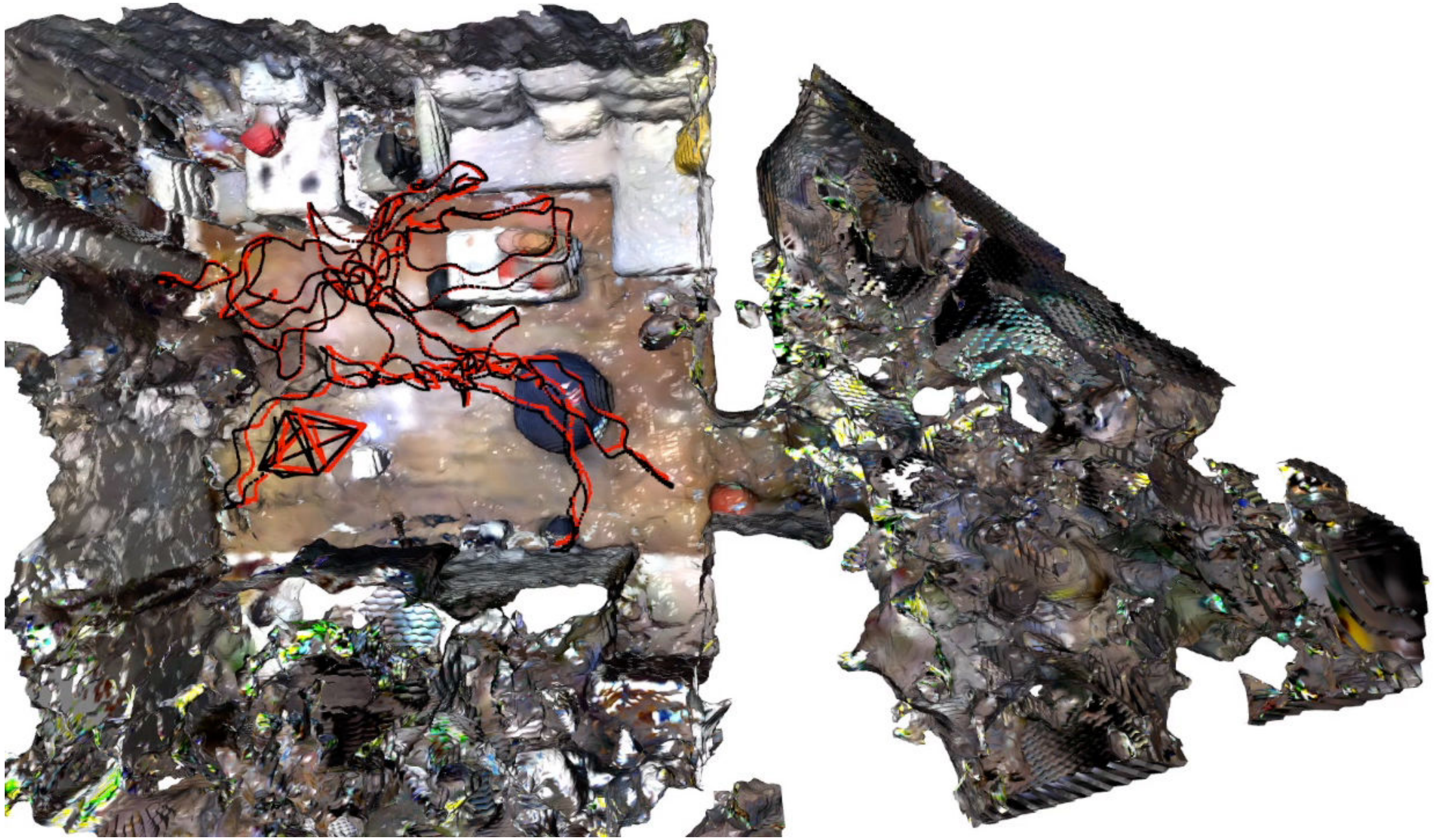
Neural Implicit SLAM: iMAP

[Sucar et al., ICCV'21]

Neural Implicit SLAM: iMAP

[Sucar et al., ICCV'21]



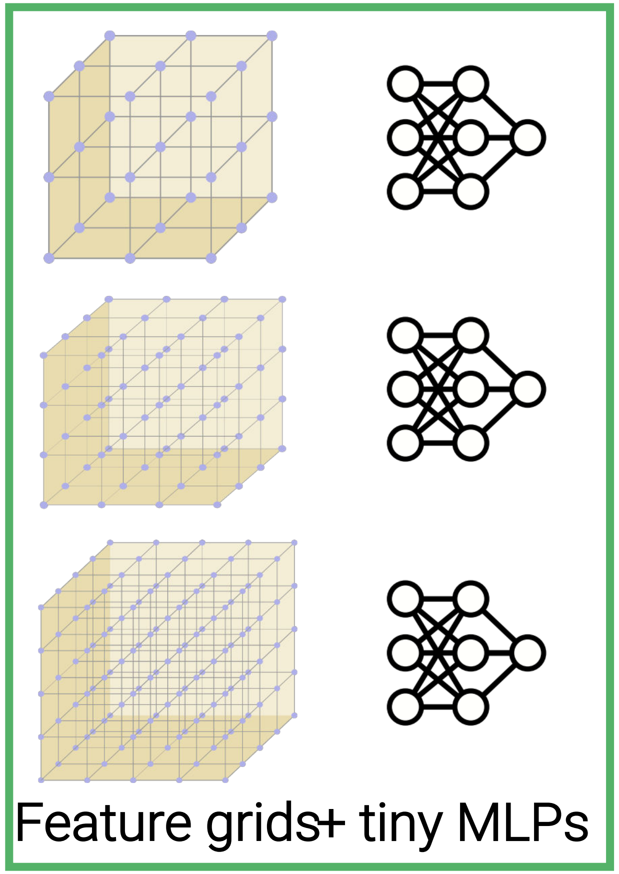


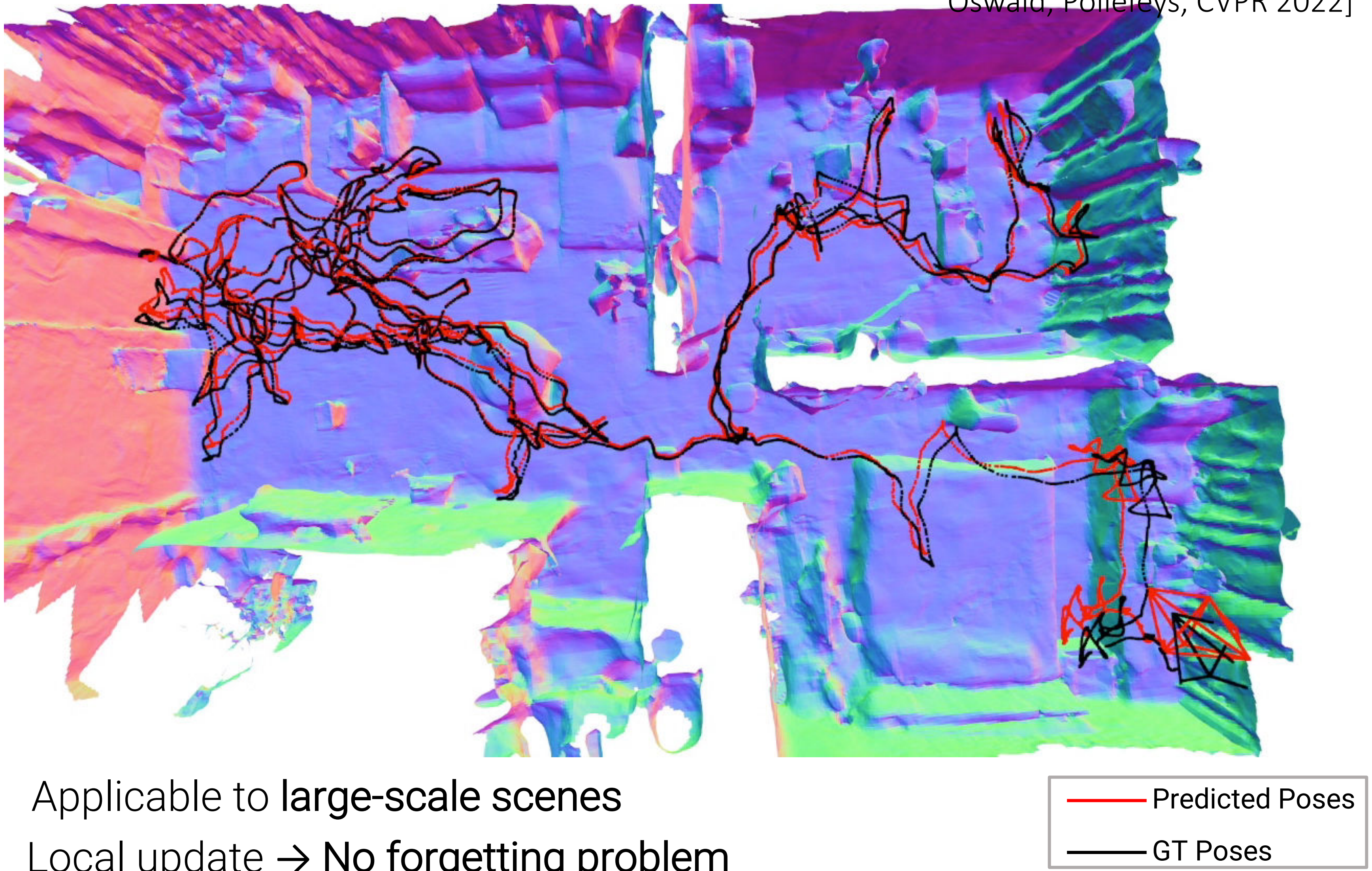
- Fail when scaling up to larger scenes
- Global update → Catastrophic forgetting
- Slow convergence

Predicted Poses
GT Poses

NICE-SLAM

[NICE-SLAM, Zhu, Peng, Larsson, Xu, Bao, Cui, Oswald, Pollefeys, CVPR 2022]

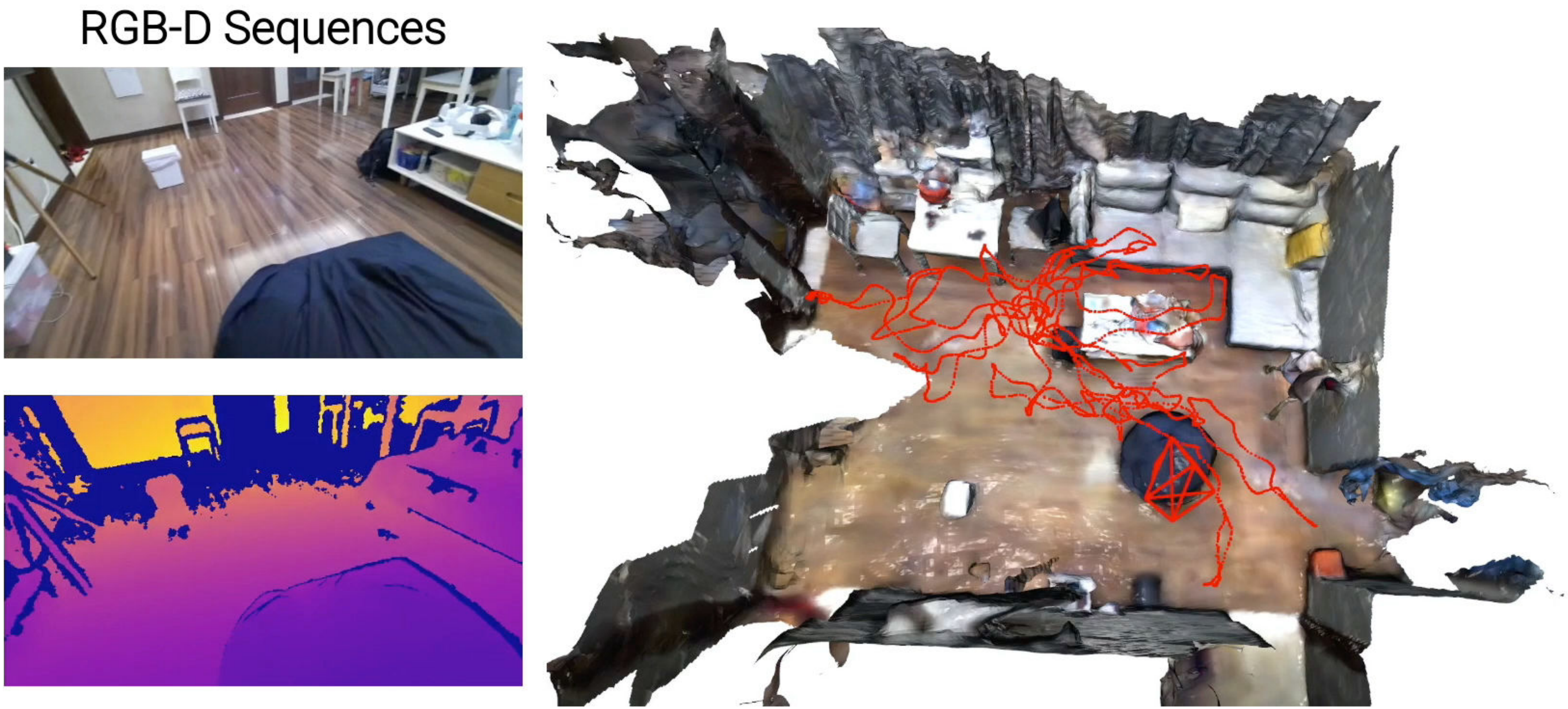




- Applicable to **large-scale scenes**
- Local update → **No forgetting problem**
- **Fast** convergence

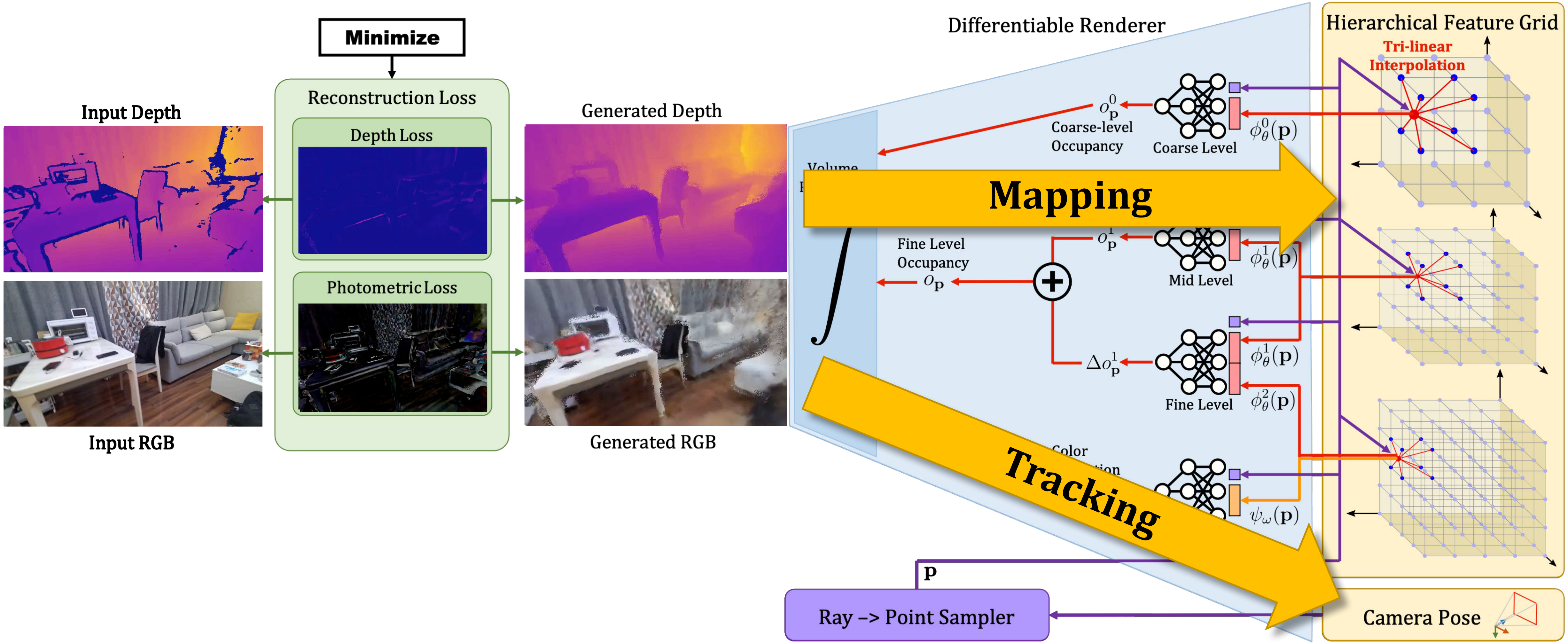
NICE-SLAM

[NICE-SLAM, Zhu, Peng, Larsson, Xu, Bao, Cui, Oswald, Pollefeys, CVPR 2022]



NICE-SLAM

[NICE-SLAM, Zhu, Peng, Larsson, Xu, Bao, Cui, Oswald, Pollefeys, CVPR 2022]



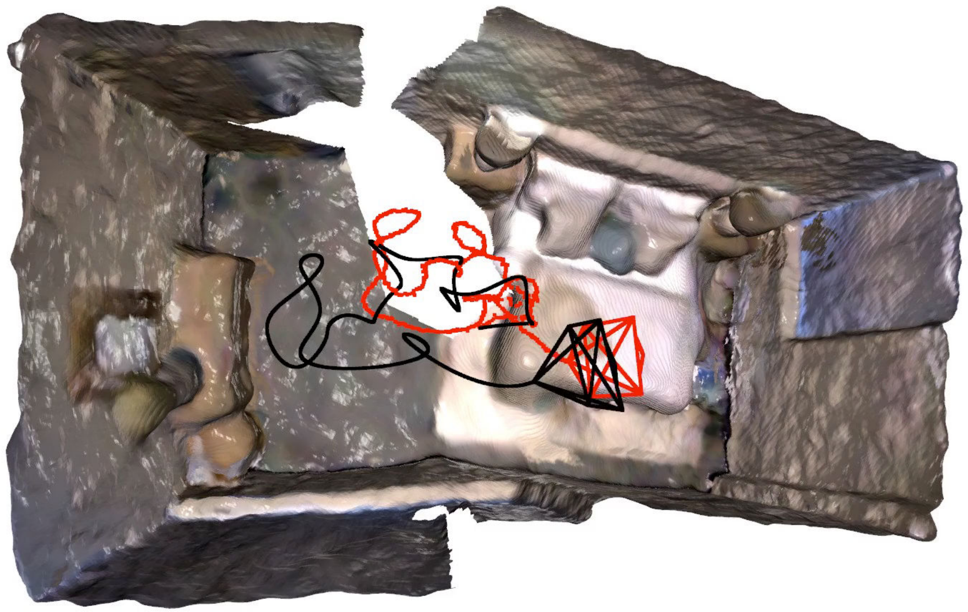
Results

[NICE-SLAM, Zhu, Peng, Larsson, Xu, Bao, Cui, Oswald, Pollefeys, CVPR 2022]

iMAP*

(our re-implementation of iMAP)

NICE-SLAM



Predicted Poses

GT Poses

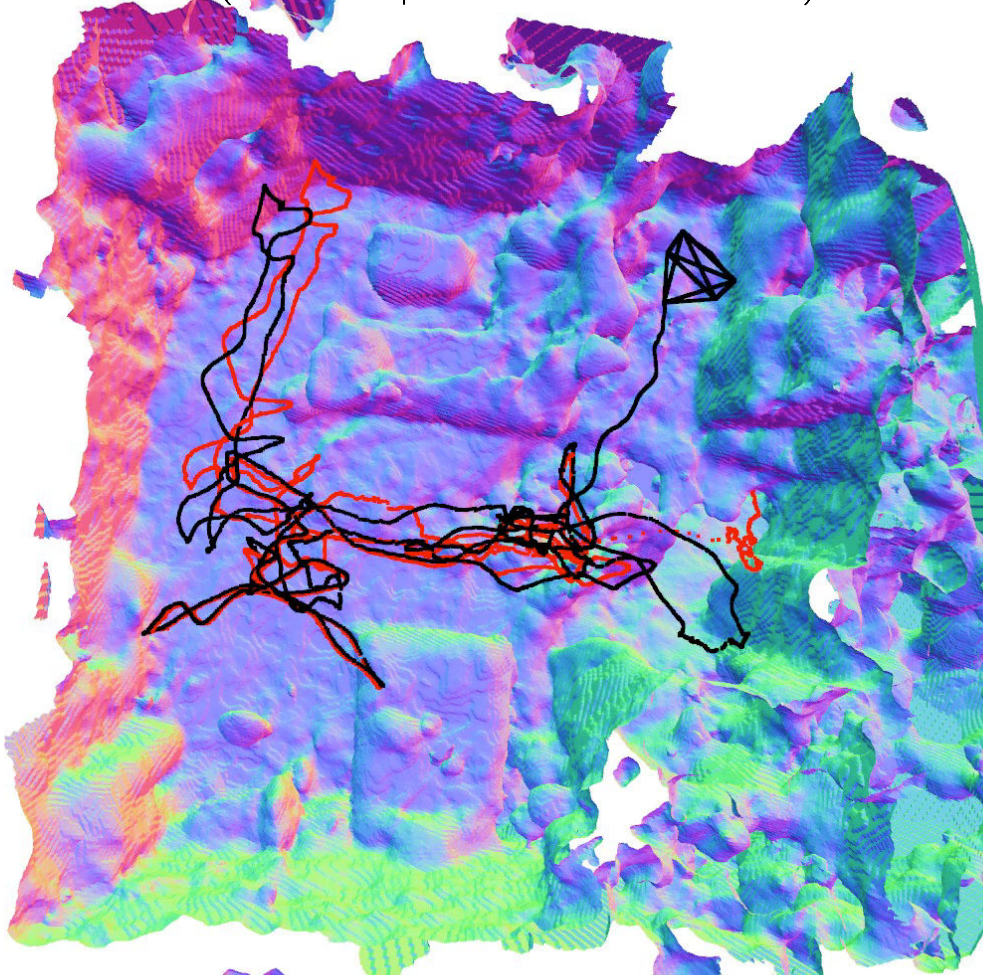
4x Speed

Results

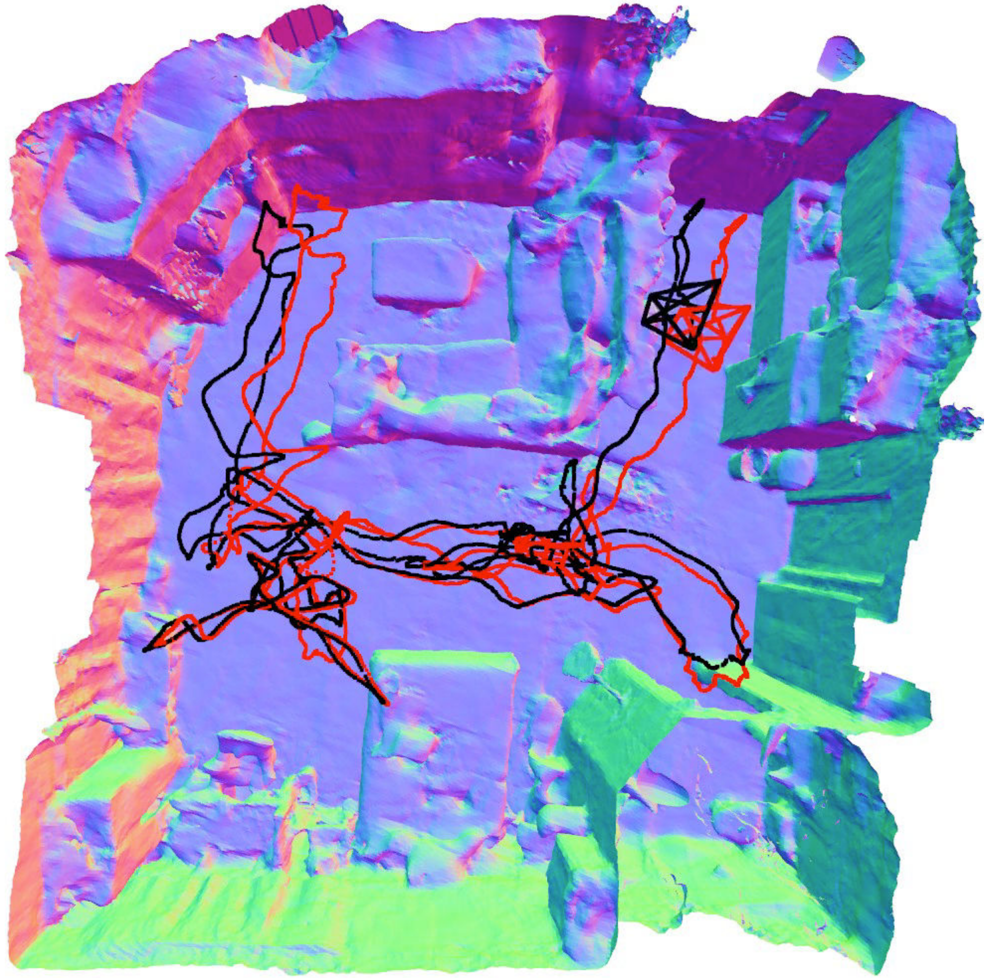
[NICE-SLAM, Zhu, Peng, Larsson, Xu, Bao, Cui, Oswald, Pollefeys, CVPR 2022]

iMAP*

(our re-implementation of iMAP)



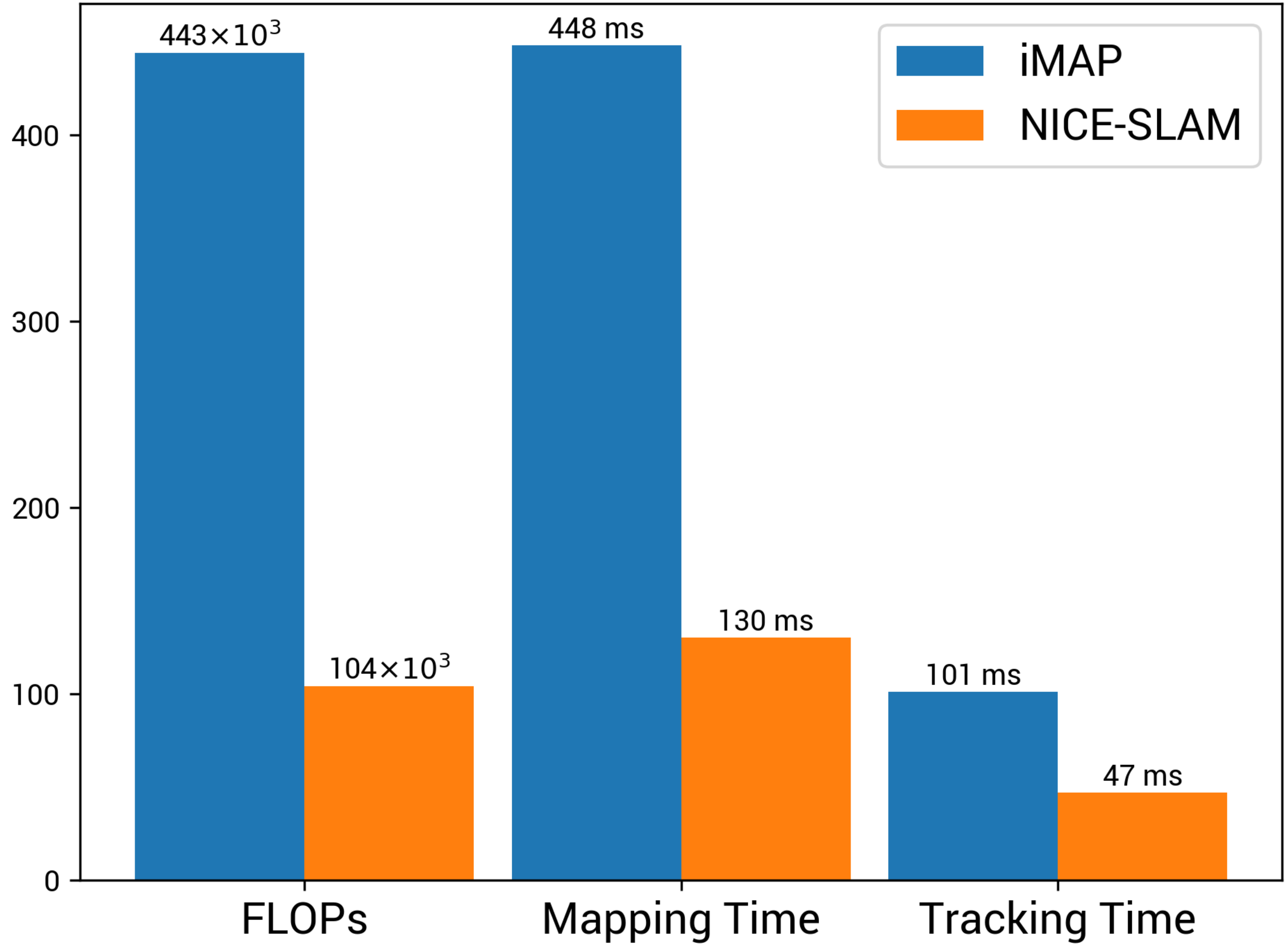
NICE-SLAM



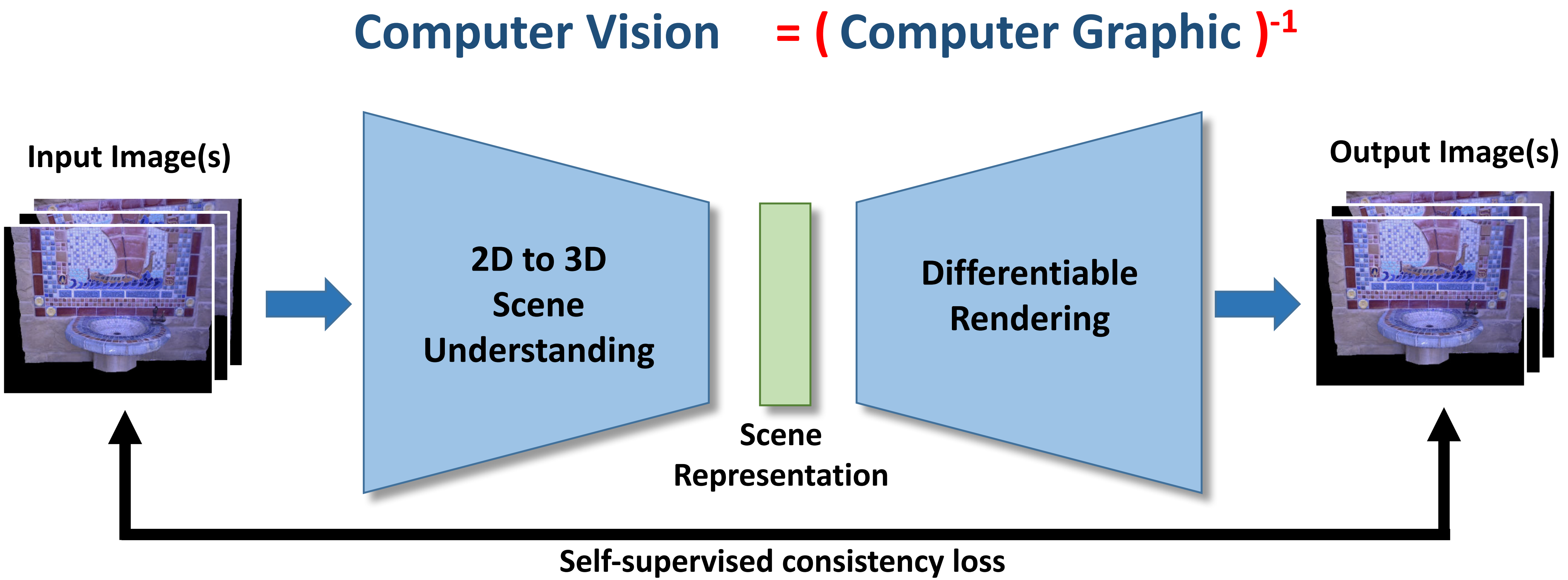
10x Speed

Results

[NICE-SLAM, Zhu, Peng, Larsson, Xu, Bao, Cui, Oswald, Pollefeys, CVPR 2022]

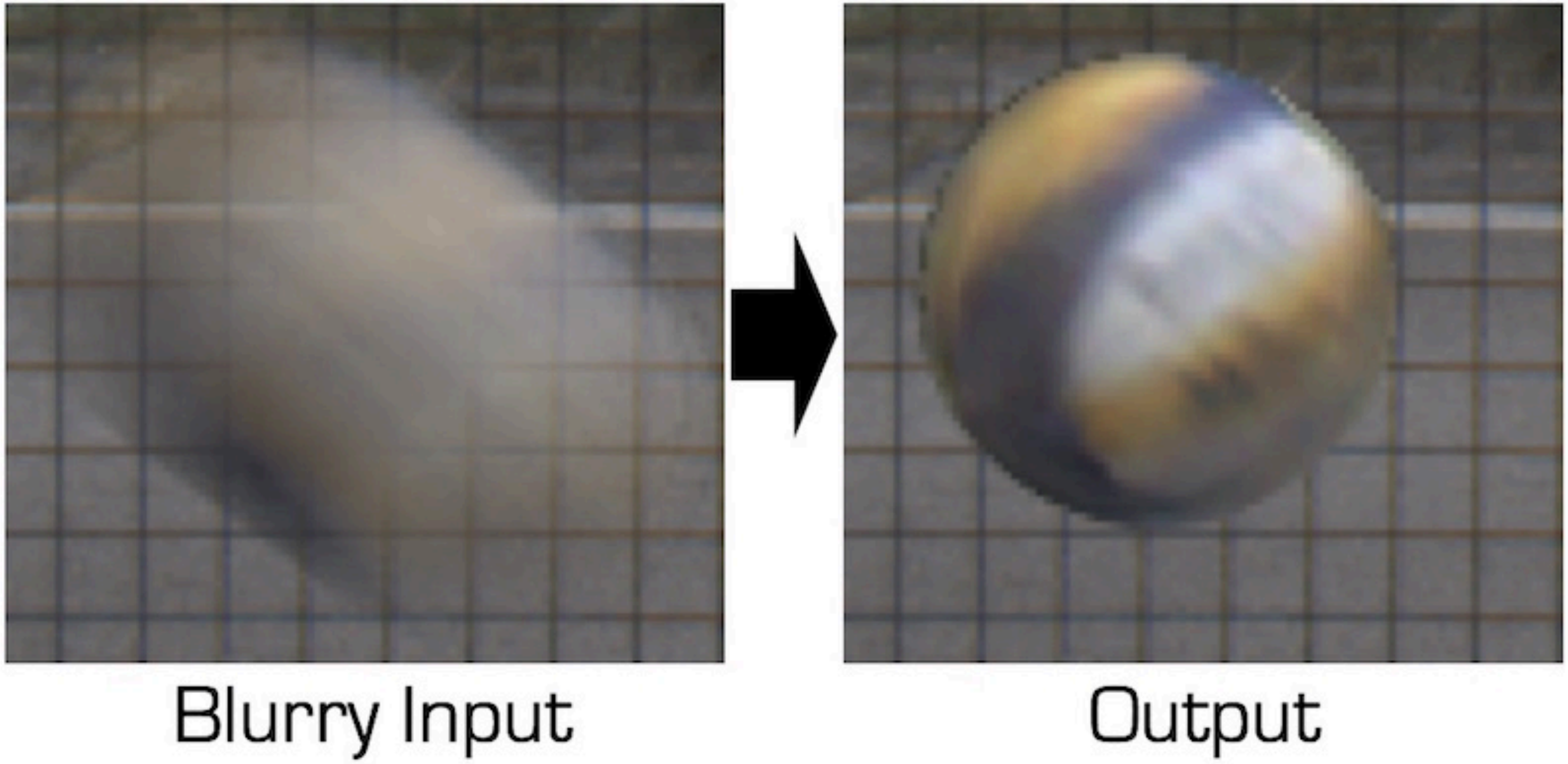


Self-Supervised Learning



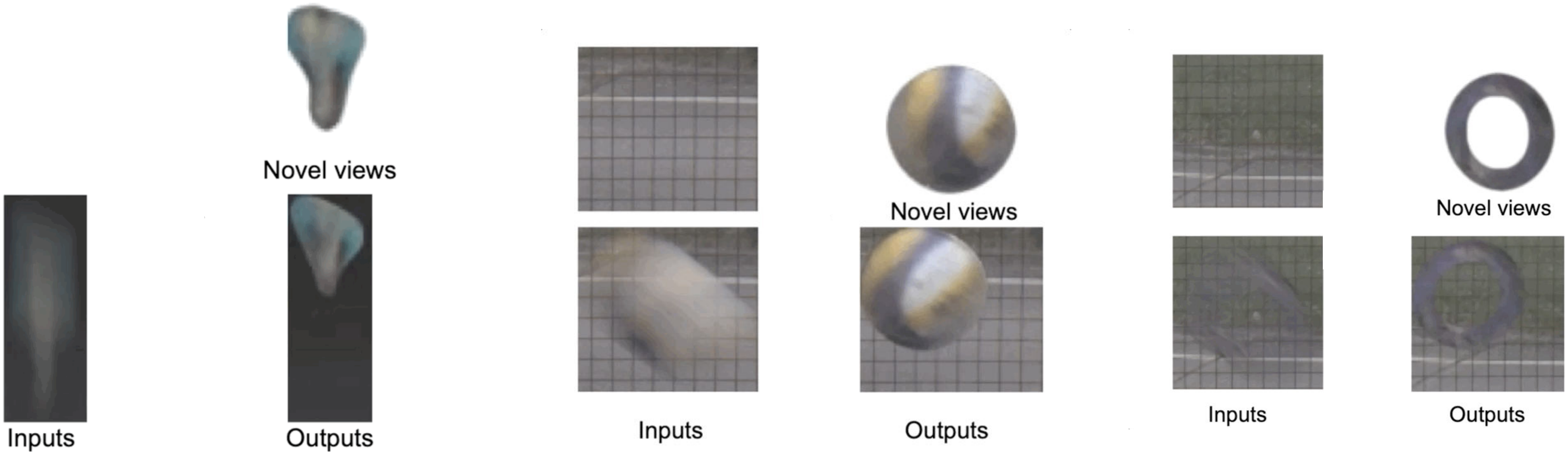
Shape From Blur

[Rozumnyi, Oswald, Ferrari, Pollefeys, NeurIPS 2021]



Shape From Blur

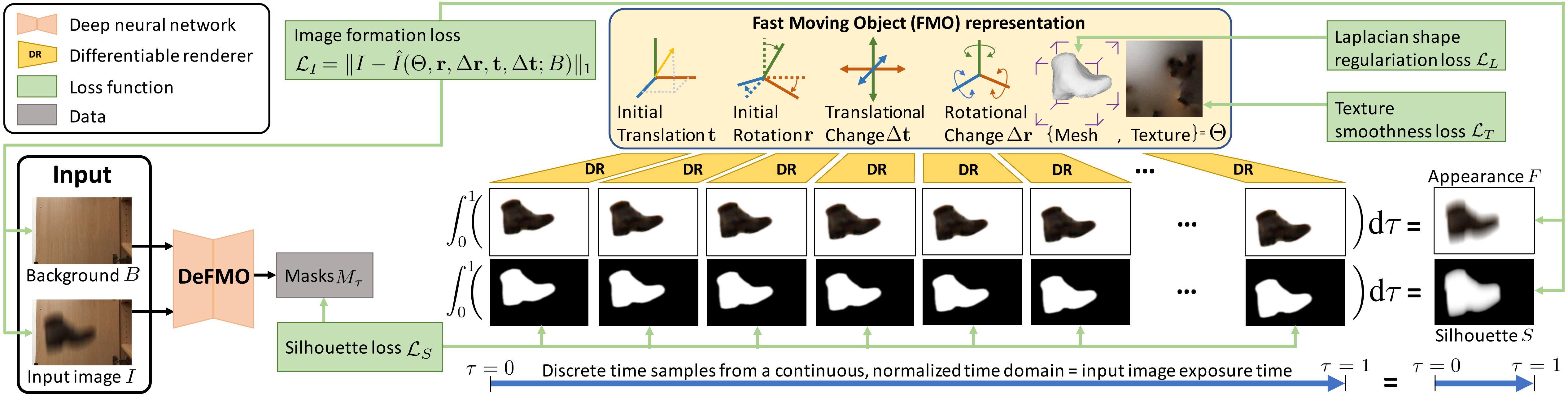
[Rozumnyi, Oswald,
Ferrari, Pollefeys, NeurIPS 2021]



2D → 3D

Method overview

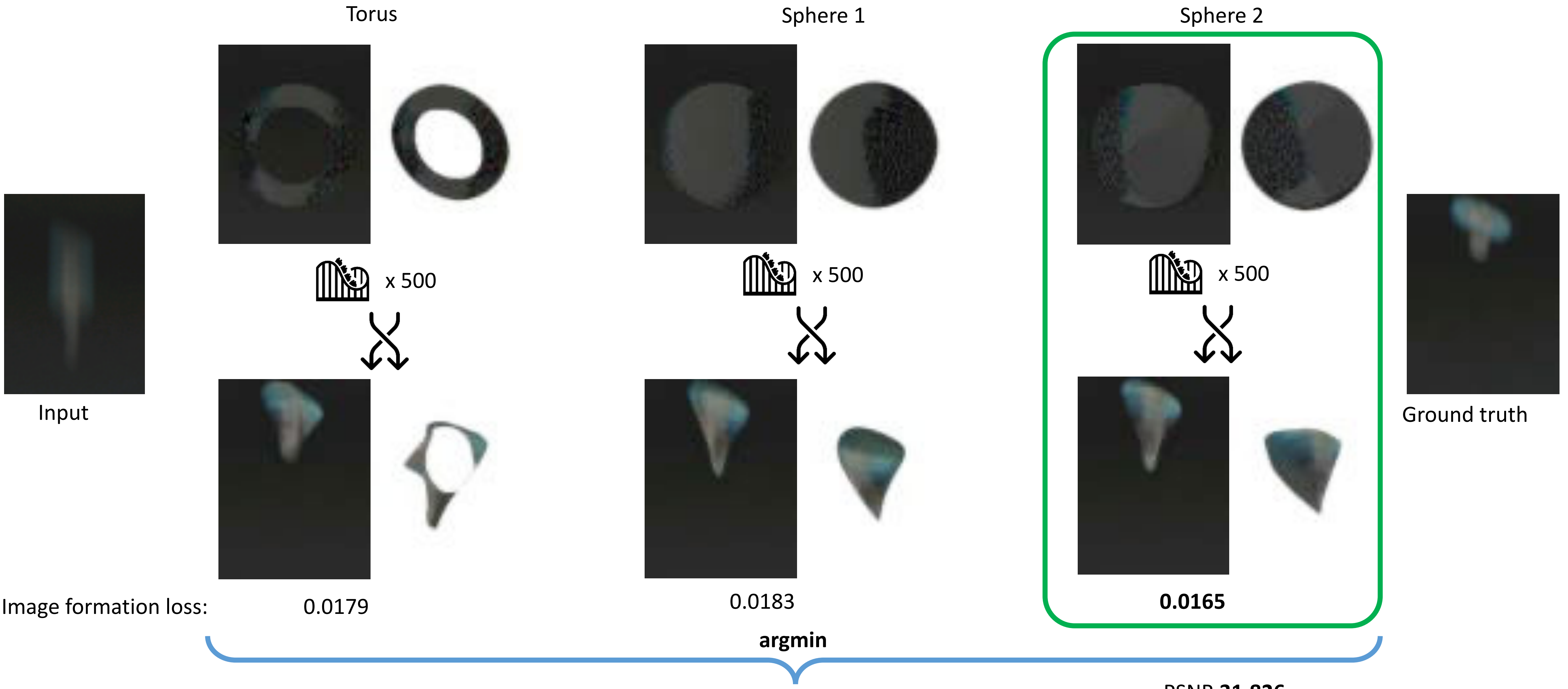
[Rozumnyi, Oswald, Ferrari, Pollefeys, NeurIPS 2021]



[Rozumnyi et al. “DeFMO: Deblurring and Shape Recovery of Fast Moving Objects”, CVPR 2021]

Loss Optimization

[Rozumnyi, Oswald, Ferrari, Pollefeys, NeurIPS2021]



Scores:

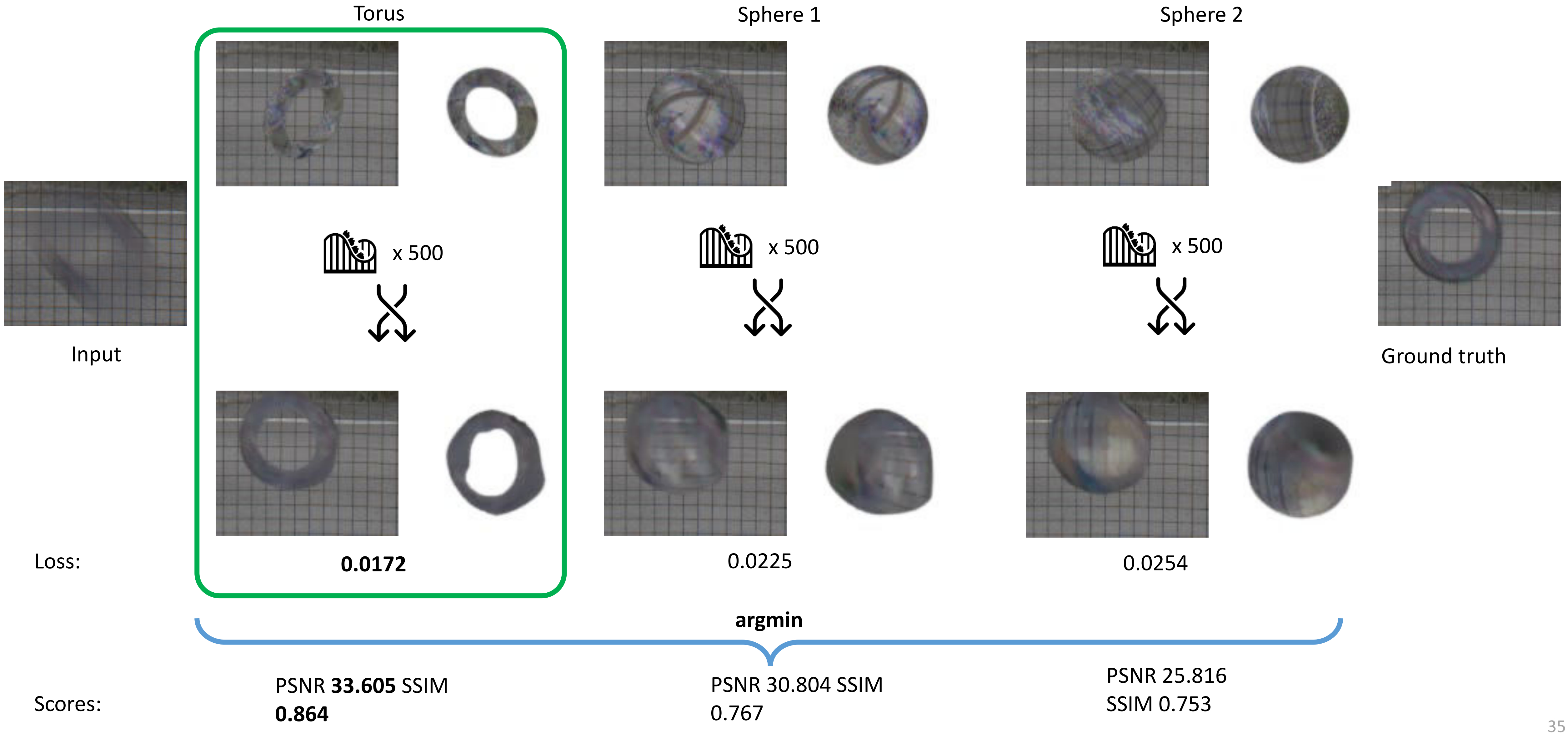
PSNR 30.132
SSIM 0.772

PSNR 28.265
SSIM 0.768

PSNR **31.826**
SSIM **0.794**

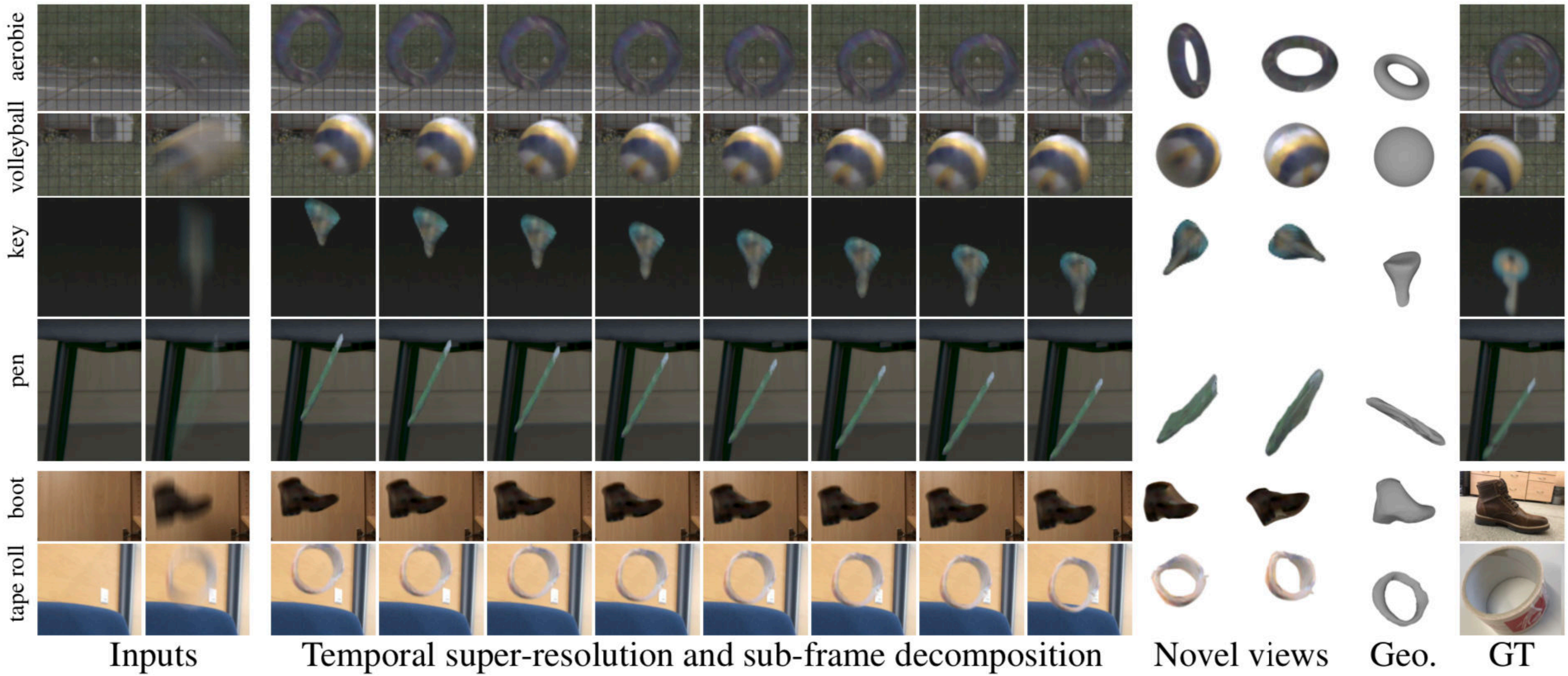
Loss Optimization

[Rozumnyi, Oswald, Ferrari, Pollefeys, NeurIPS2021]



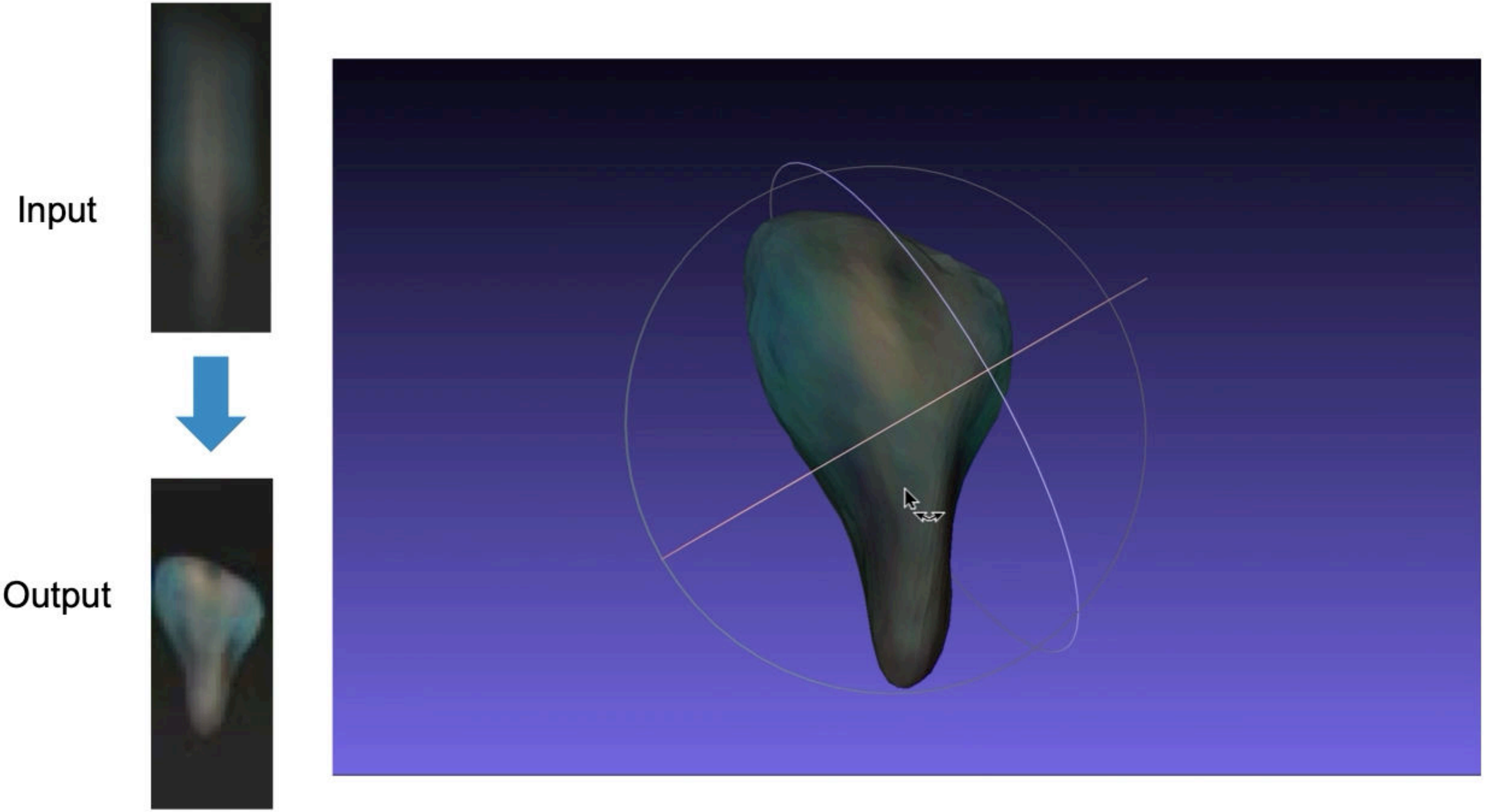
Shape From Blur Results

[Rozumnyi, Oswald, Ferrari, Pollefeys, NeurIPS2021]



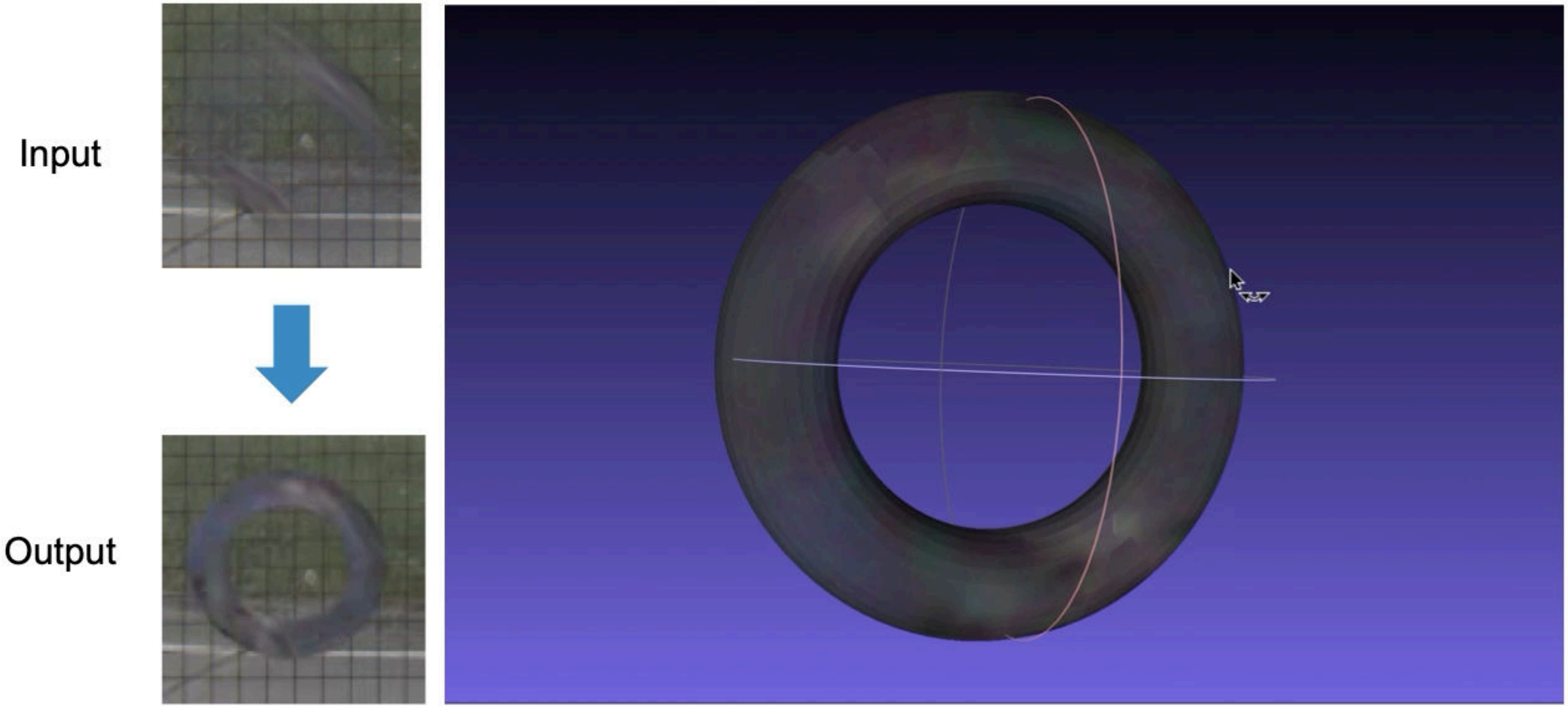
Shape From Blur Results

[Rozumnyi, Oswald, Ferrari, Pollefeys, NeurIPS2021]



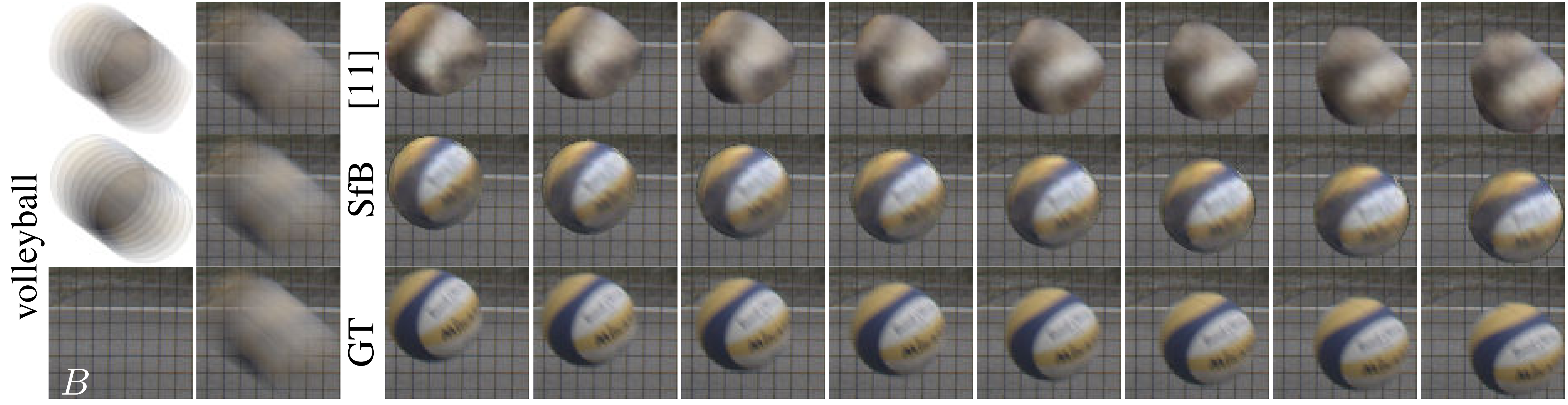
Shape From Blur Results

[Rozumnyi, Oswald, Ferrari, Pollefeys, NeurIPS2021]



Shape From Blur Results

[Rozumnyi, Oswald, Ferrari, Pollefeys, NeurIPS2021]

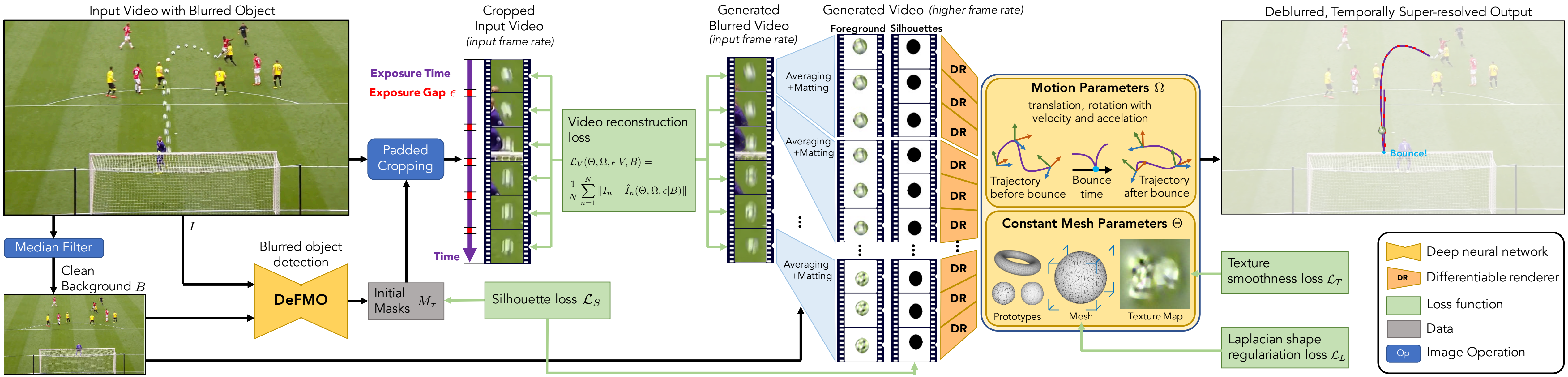


Can we further improve these results?

Yes!

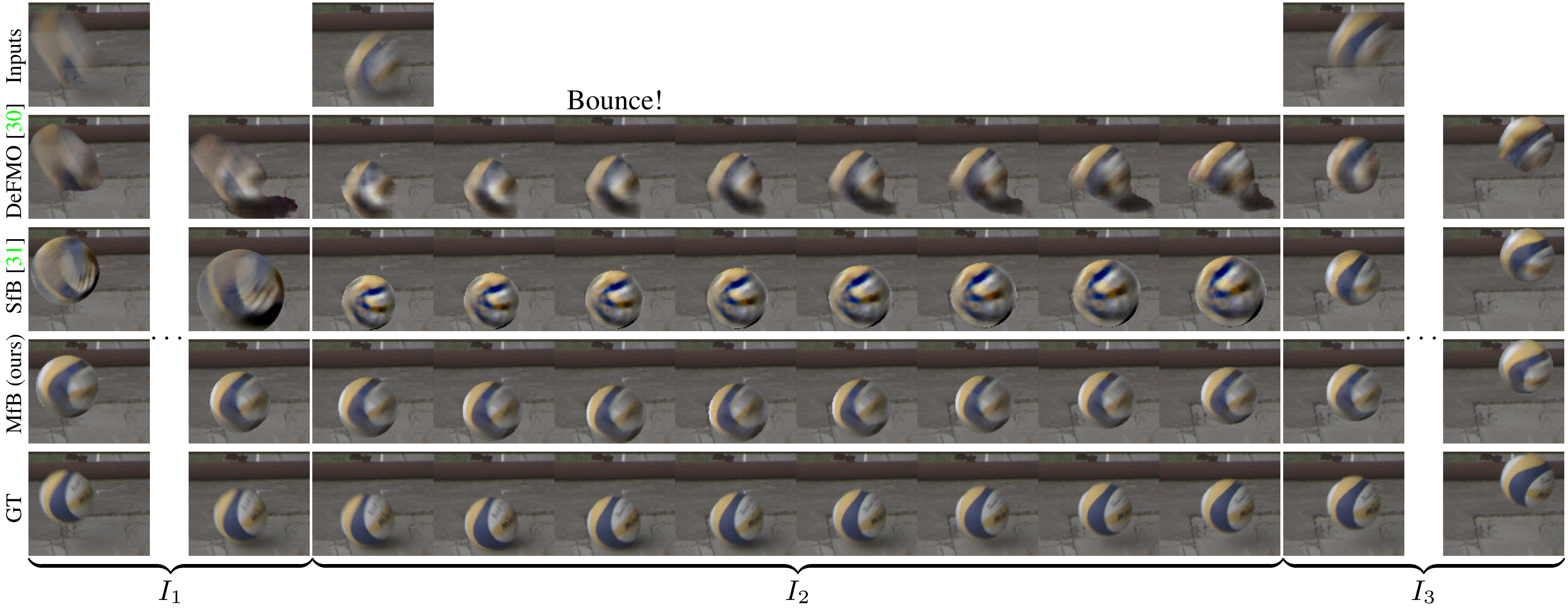
Motion From Blur

[Rozumnyi, Oswald, Ferrari, Pollefeys, CVPR 2022]



Motion From Blur

[Rozumnyi, Oswald, Ferrari, Pollefeys, CVPR 2022]



Motion From Blur

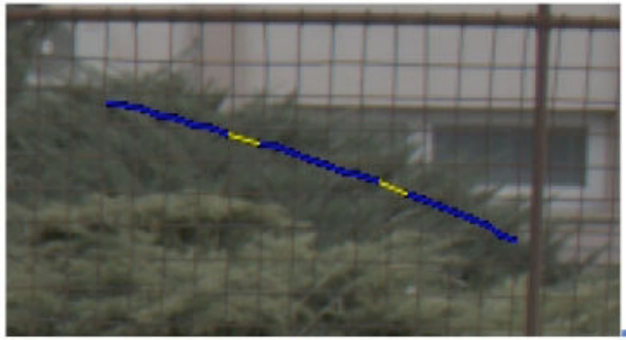
[Rozumnyi, Oswald, Ferrari, Pollefeys, CVPR 2022]

Input

Ground truth

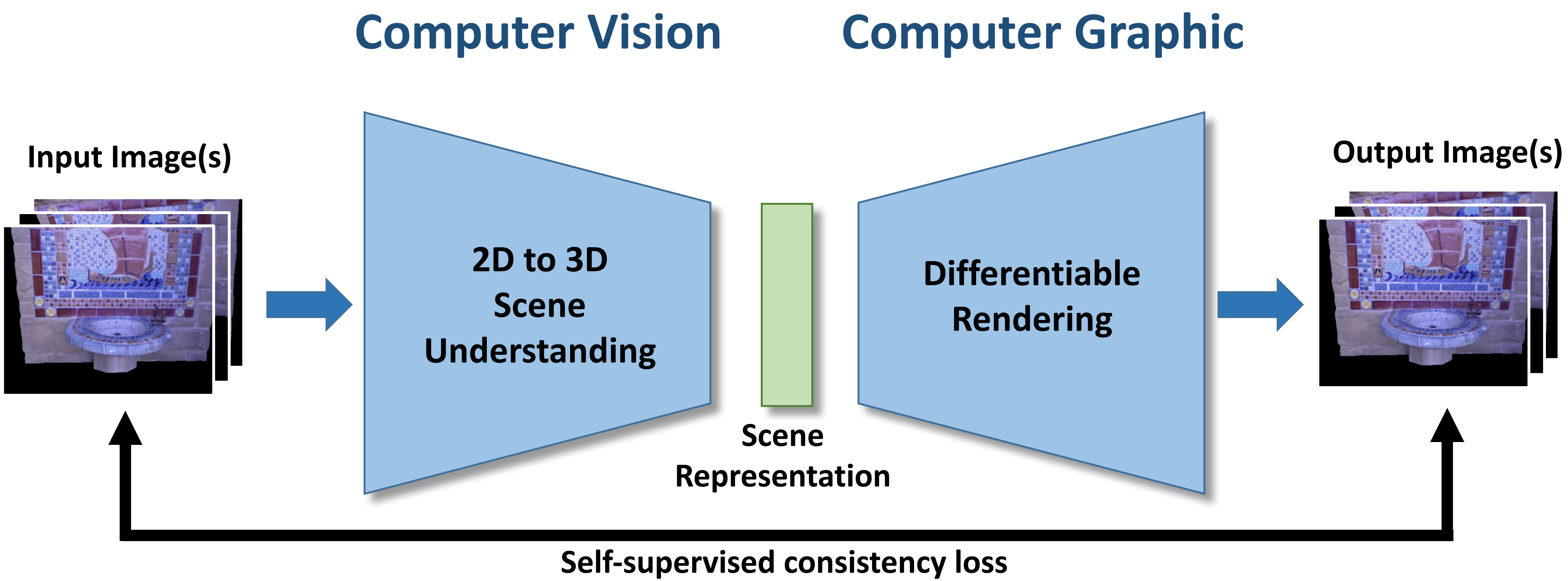
DeFMO [30]

SfB [31]



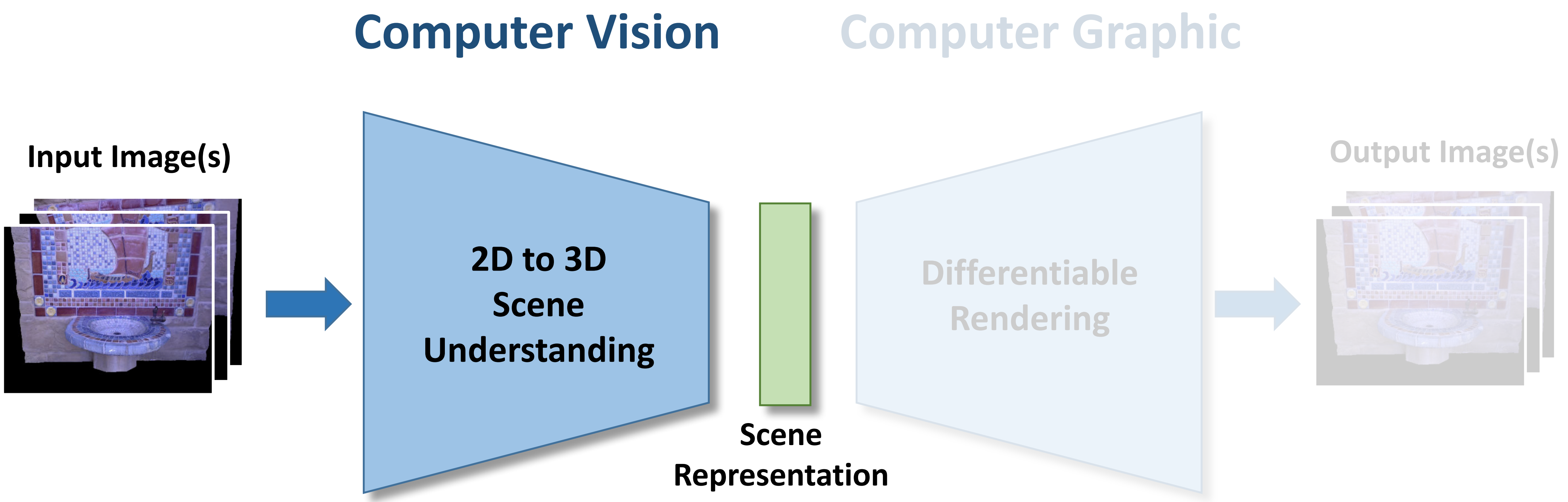
MfB (ours)

Self-Supervised Learning



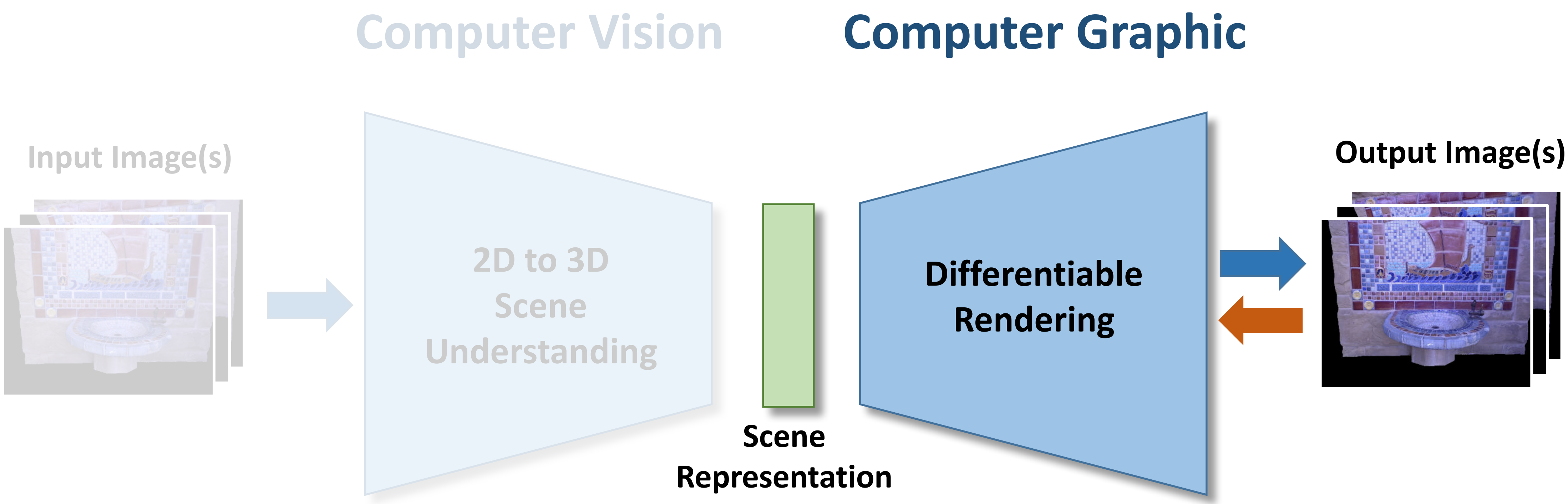
Training-Time

Self-Supervised Learning



Test-Time

Self-Supervised Learning



Test-Time

Take Home Messages

- Combining learning approaches and classical geometry improves learning

- Image reprojection error with differentiable rendering is a powerful supervisory signal - at training + test time -> self-supervised learning!

- Test time optimization is powerful
  -> clever combination of training+test-time optimization