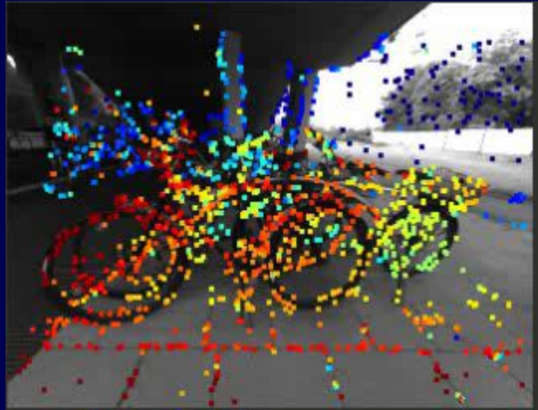
Chapter 10: Recent Developments in Multi-view Reconstruction

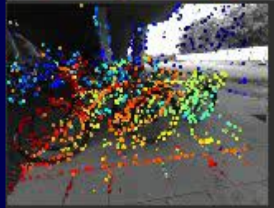
Daniel Cremers
Chair of Computer Vision and AI, TU Munich
Munich Center for Machine Learning

LSD SLAM: Large-Scale Direct SLAM

Engel, Schöps, Cremers, ECCV 2014:
LSD SLAM

Engel, Koltun, Cremers, PAMI 2018:
Direct Sparse Odometry





Engel, Koltun, Cremers, PAMI 2018

Mobile Mapping System:

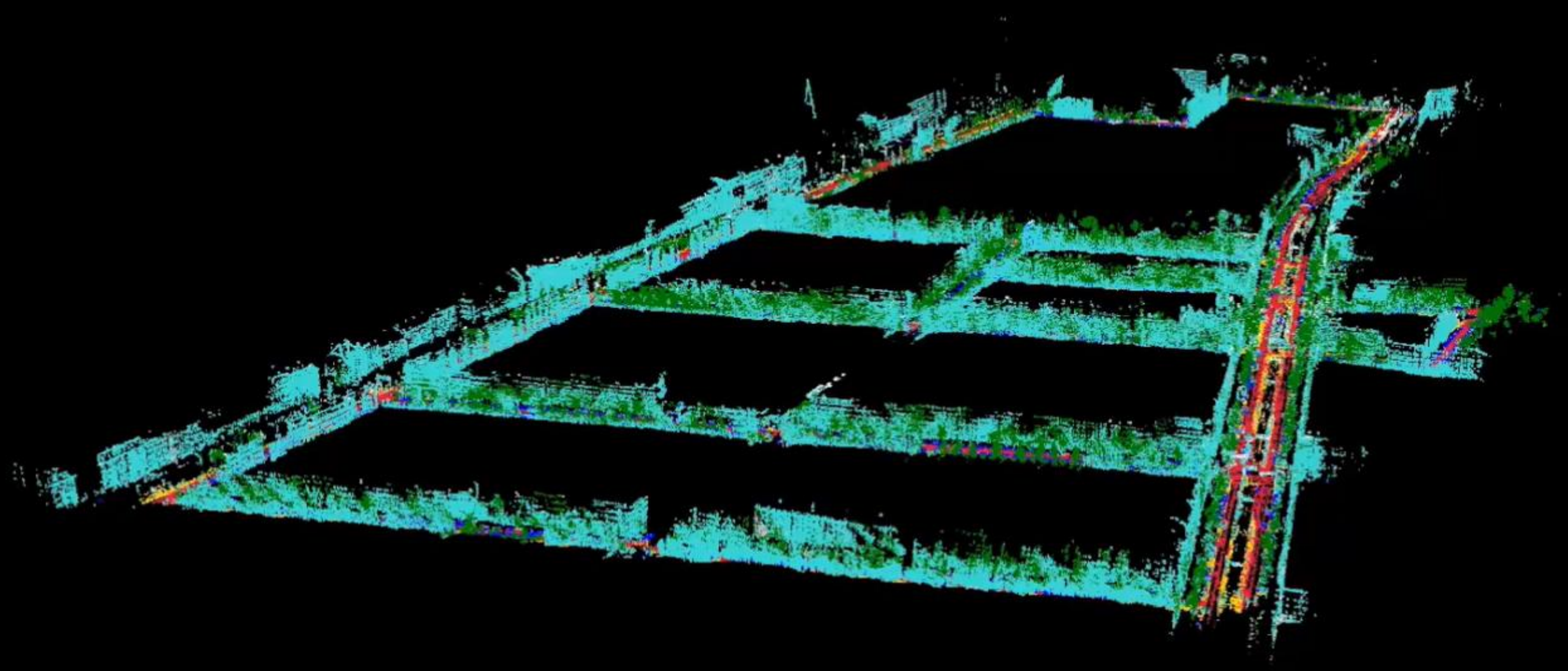
- GPS-independent realtime localization
- Leveraging off-the-shelf cameras

Spatial AI Cloud Platform:

- Access to navigation and map data
- Semantic scene understanding

Visual-inertial Odometry

Von Stumberg, Cremers, “Delayed Marginalization VI Odometry”, ICRA 2022



drivable

sidewalk

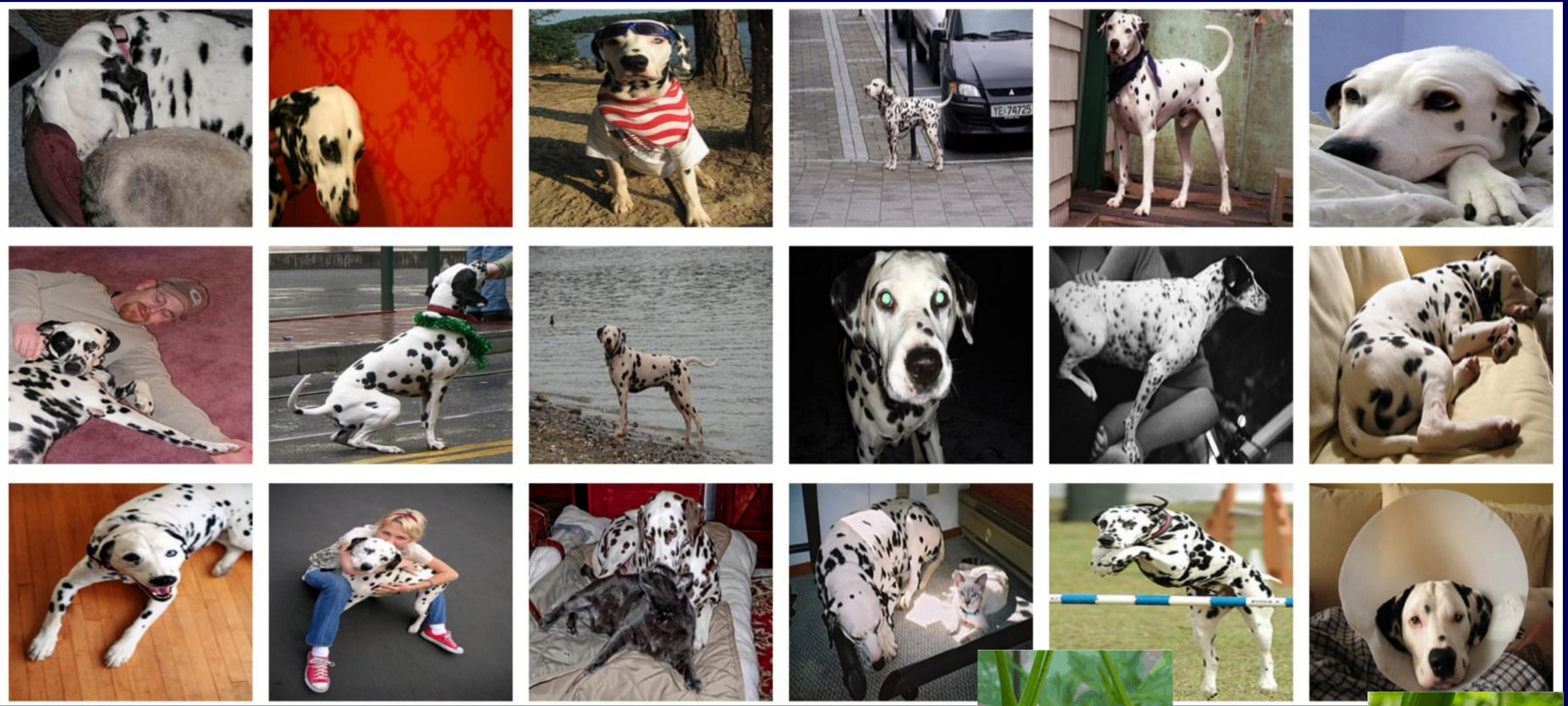
cars

buildings

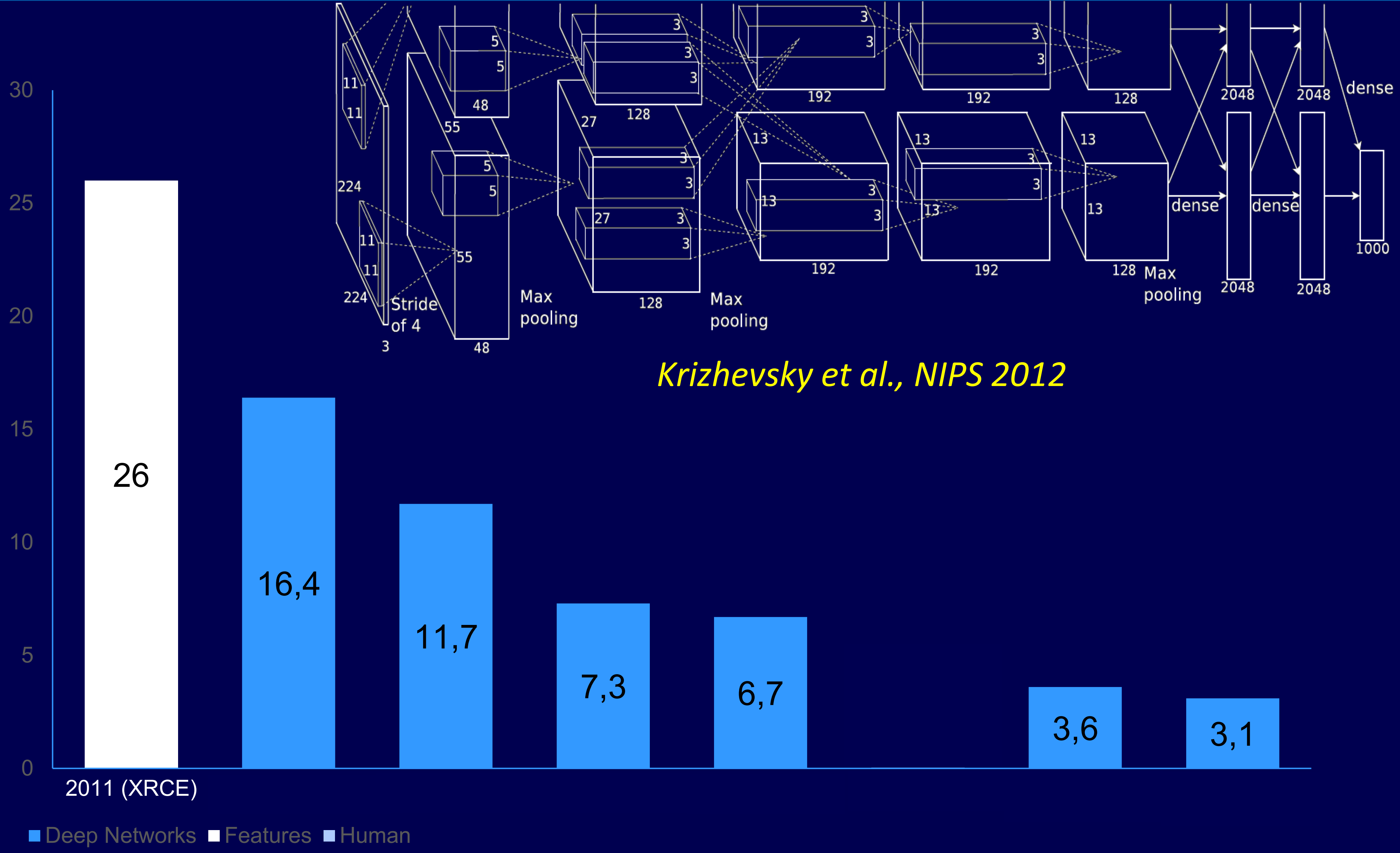
vegetation

...

ImageNet: Objects in 14 Mio Images



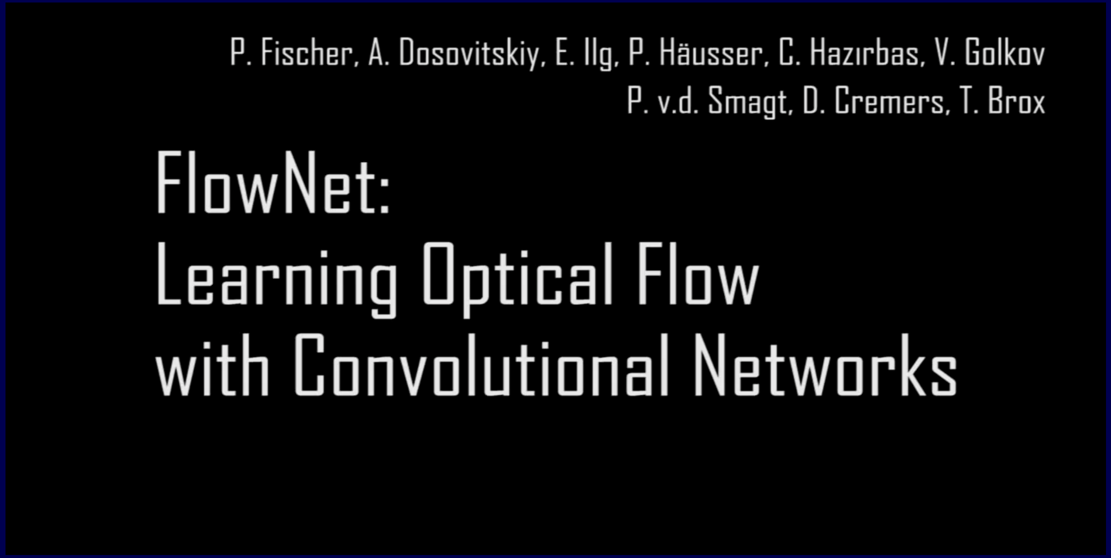
The Advent of Deep Networks



Krizhevsky et al., NIPS 2012

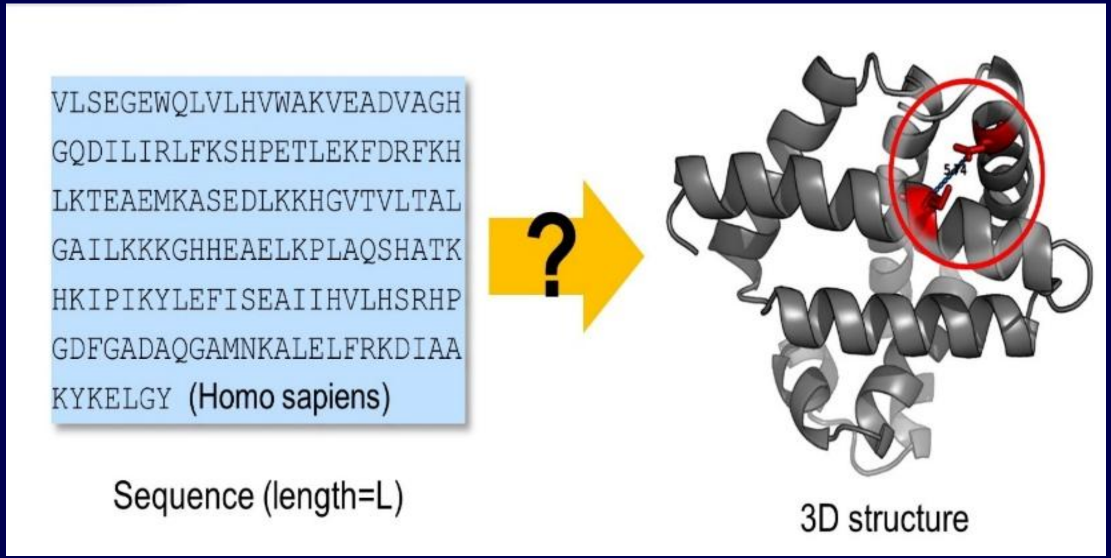
Performance Scores on the ImageNet Recognition Challenge

Deep Nets beyond Object Recognition



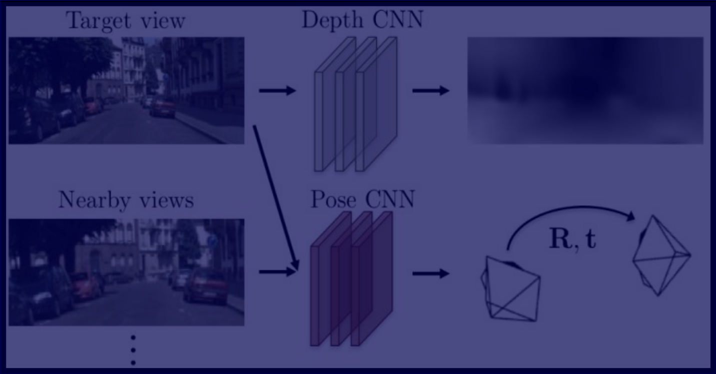
Dosovitsky et al., "FlowNet", ICCV 2015

Badrinarayanan et al., "SegNet", arxiv'15

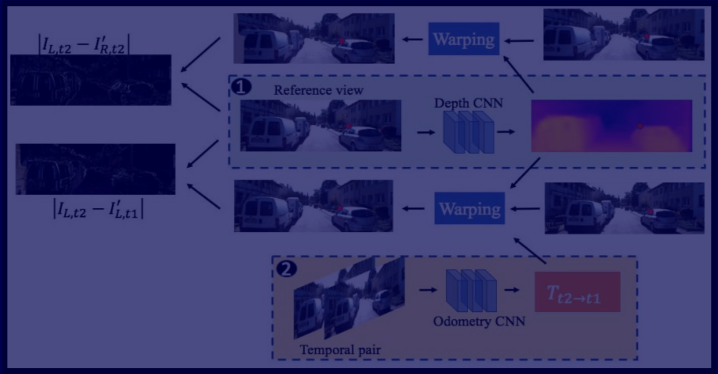


Golkov et al., NeurIPS '16

Caelles et al., CVPR 2018



Zhou et al. CVPR 17

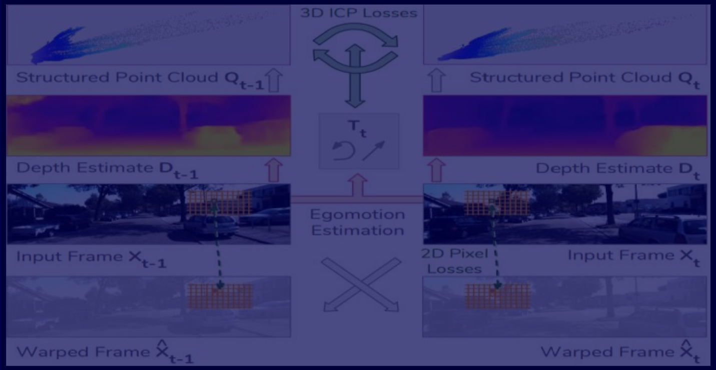


Zhan et al. CVPR 18

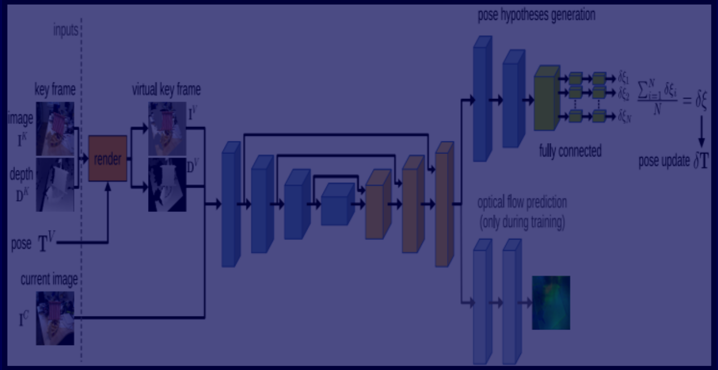
Initial methods were not state-of-the-art!

Ummenhofer et al. CVPR 17

Yin et al. CVPR 18



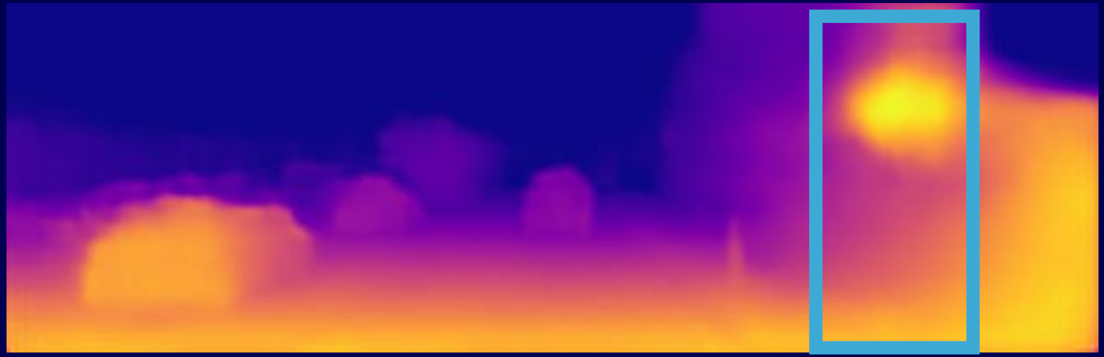
Mahjourian et al. CVPR 18



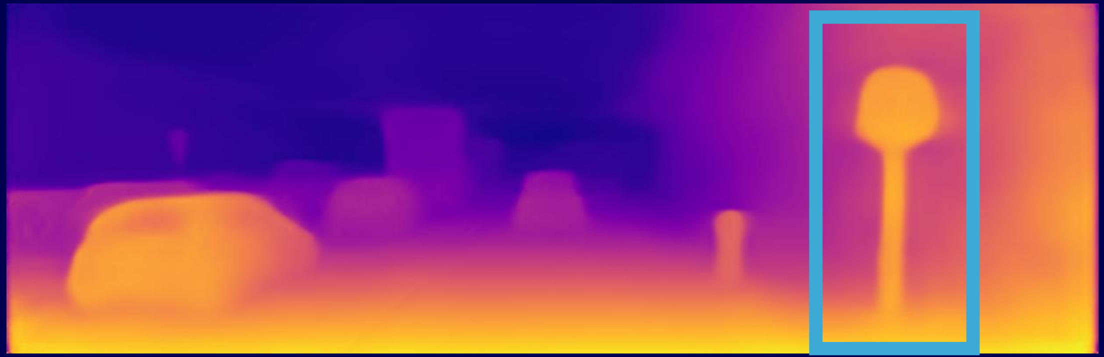
Zhou et al. ECCV 18

Deep Virtual Stereo Odometry

Deep
Neural
Network

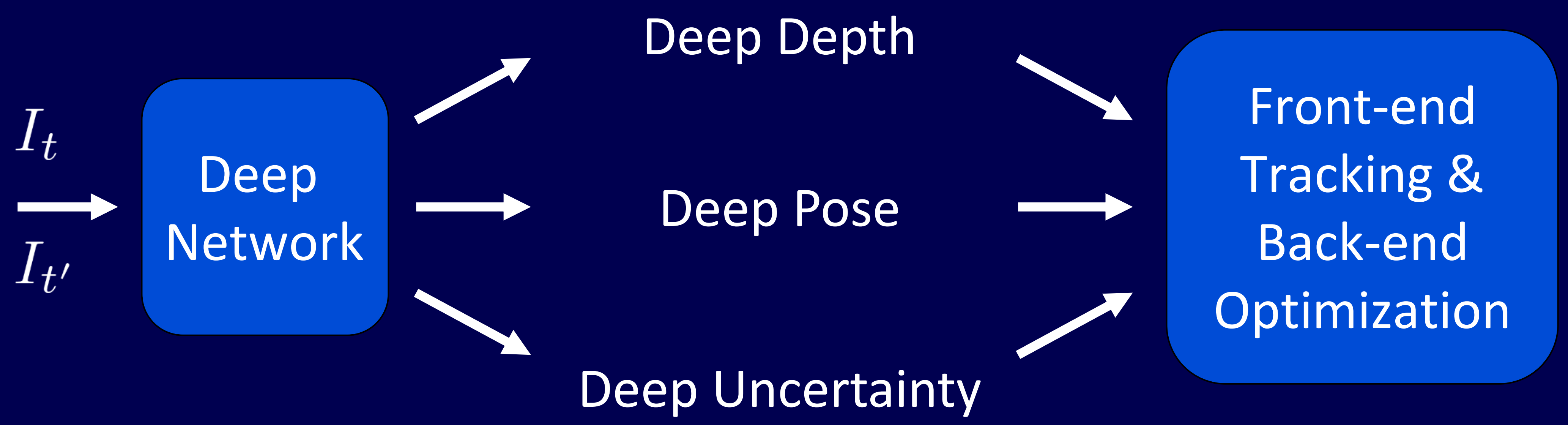


Kuznietsov et al. CVPR 2017



Yang, Wang, Stueckler, Cremers, ECCV 2018

D3VO: Depth, Pose & Uncertainty



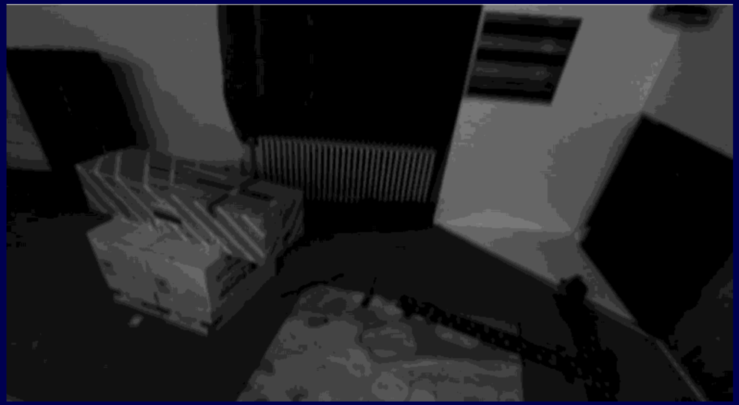
Yang et al., “D3VO”, CVPR ‘20

Deep Depth and Pose

I_t → Depth Net → D_t

I_t, $I_{t'}$ → Pose Net → $T_{t'}^{t}$

I_t



$I_{t'\to t}$

Self-supervised learning:

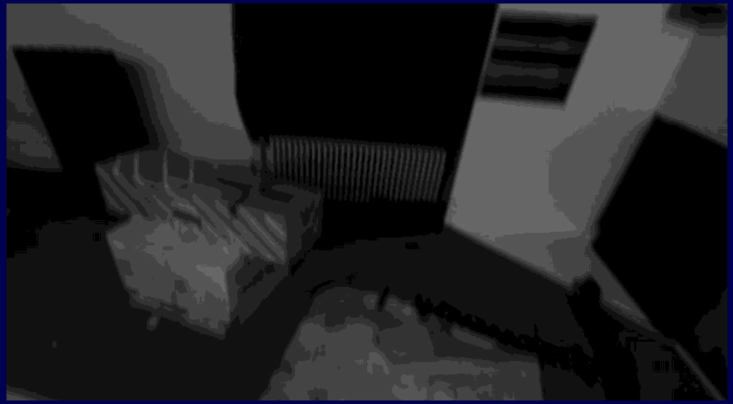
$$L_{self} = r(I_t, I_{t'\to t})$$

Yang et al., "D3VO", CVPR '20

Deep Affine Brightness Correction

$I_t \rightarrow$ Depth Net $\rightarrow D_t$

$I_t, I_{t'} \rightarrow$ Pose Net $\rightarrow T_{t'}^{t}$, $a_t^{t'}, b_t^{t'}$



$a_t^{t'} I_t + b_t^{t'}$

$I_{t' \rightarrow t}$

Self-supervised learning:

$$L_{self} = r(a_t^{t'} I_t + b_t^{t'}, I_{t' \rightarrow t})$$

Yang et al., "D3VO", CVPR '20

$I_t \rightarrow$ Depth Net $\rightarrow D_t,\ \Sigma_t$

$I_{t'} \rightarrow$ Pose Net $\rightarrow T_{t'}^{t},\ a_t^{t'}, b_t^{t'}$

I_t

Σ_t

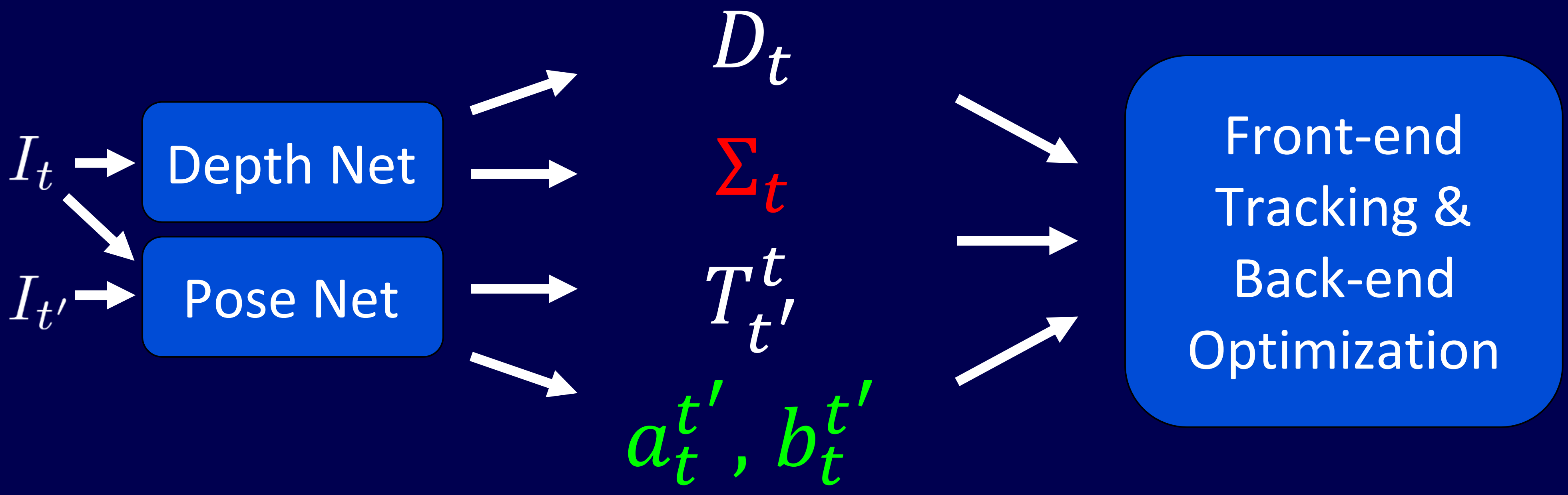
Self-supervised learning:

$$L_{self} = \frac{r(a_t^{t'} I_t + b_t^{t'}, I_{t' \to t})}{\Sigma_t} + \log \Sigma_t$$

Kendall et al., NeurIPS '17, Klodt et al. ECCV '18

Yang et al., "D3VO", CVPR '20

Predicting Depth, Pose & Uncertainty

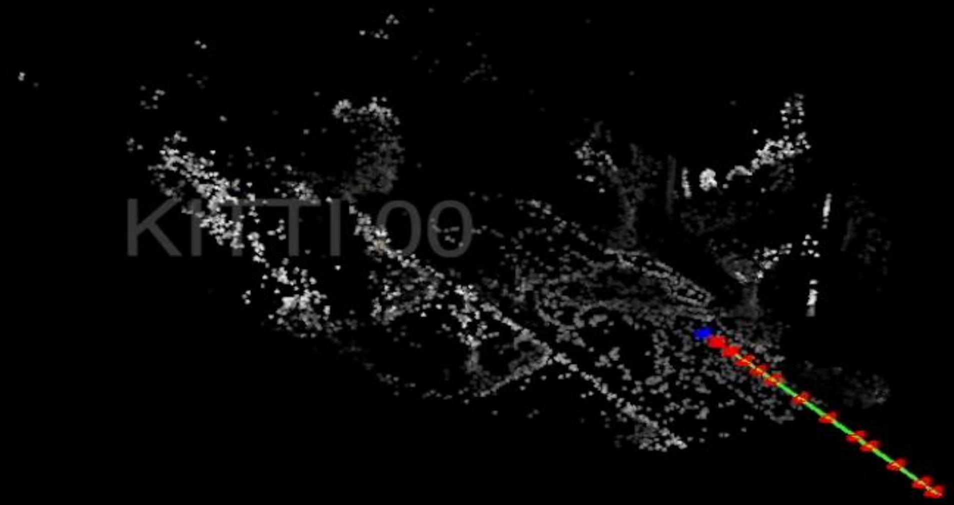


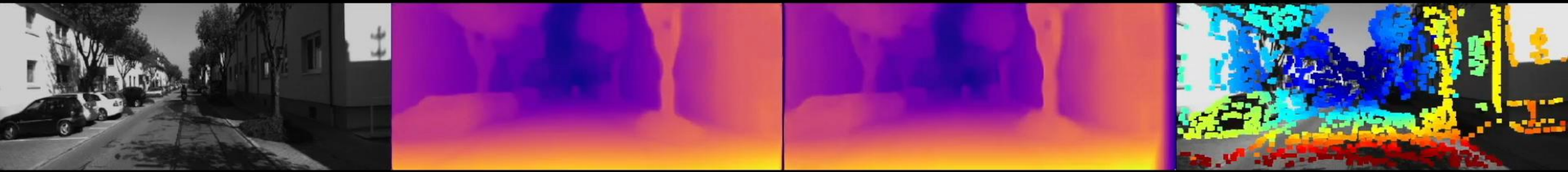
Self-supervised learning:

$$L_{self} = \frac{r(a_t^{t'} I_t + b_t^{t'}, I_{t' \to t})}{\Sigma_t} + \log \Sigma_t$$

Yang et al., "D3VO", CVPR '20

Deep Nets for Visual SLAM





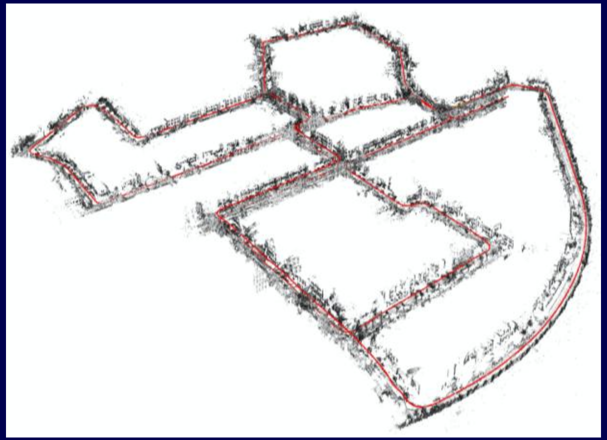
Experiments: Visual Odometry

KITTI: classical methods

		mean
Mono	ORB-SLAM	37.0
Stereo	Stereo DSO	0.89
Mono	D3VO	**0.82**

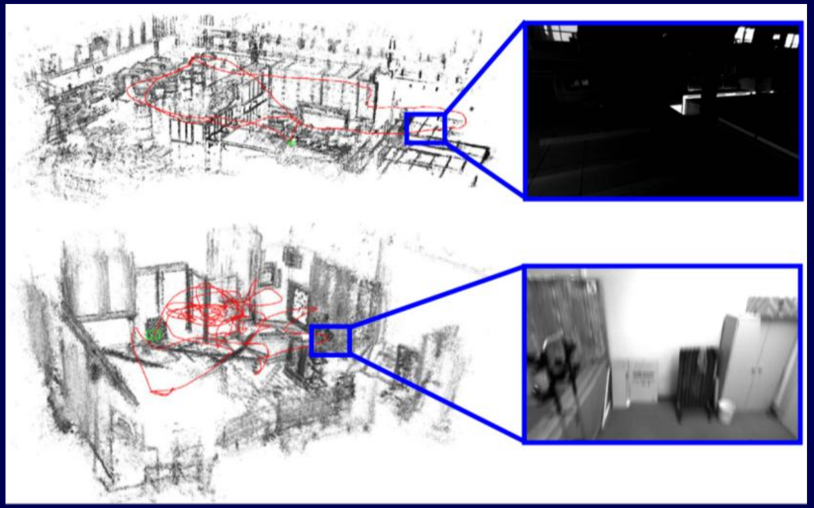
KITTI: deep methods

		Seq. 09	Seq. 10
End-to-end	Gordon et al.	2.7	6.8
Hybrid	Zhan et al.	2.61	2.29
Hybrid	D3VO	**0.78**	**0.62**



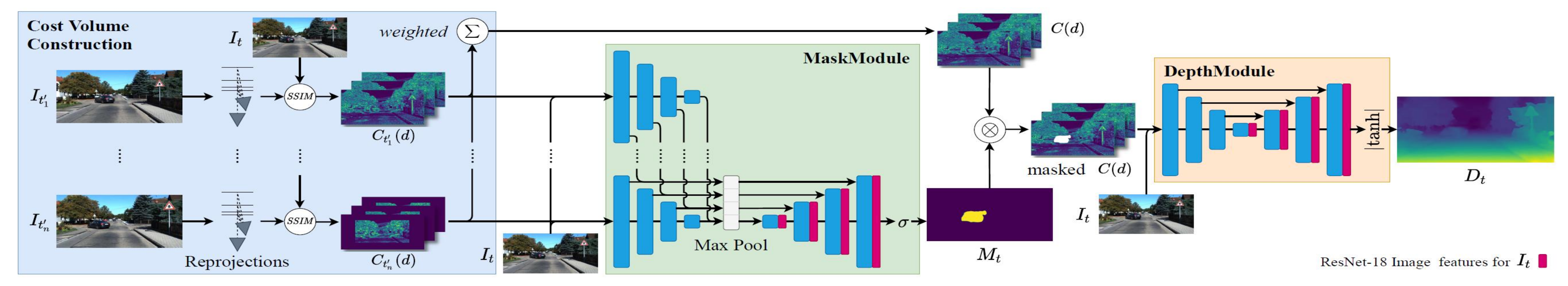
EuRoC: classical methods

		mean
Mono	DSO	0.48
Mono-VIO	VI-DSO	0.11
Stereo-VIO	Basalt	**0.08**
Mono	D3VO	**0.08**



Yang et al., "D3VO", CVPR '20

Dense Reconstructions from a Single Camera



cost volume

mask module

depth module

Aggregate brightness from multiple warped frames.

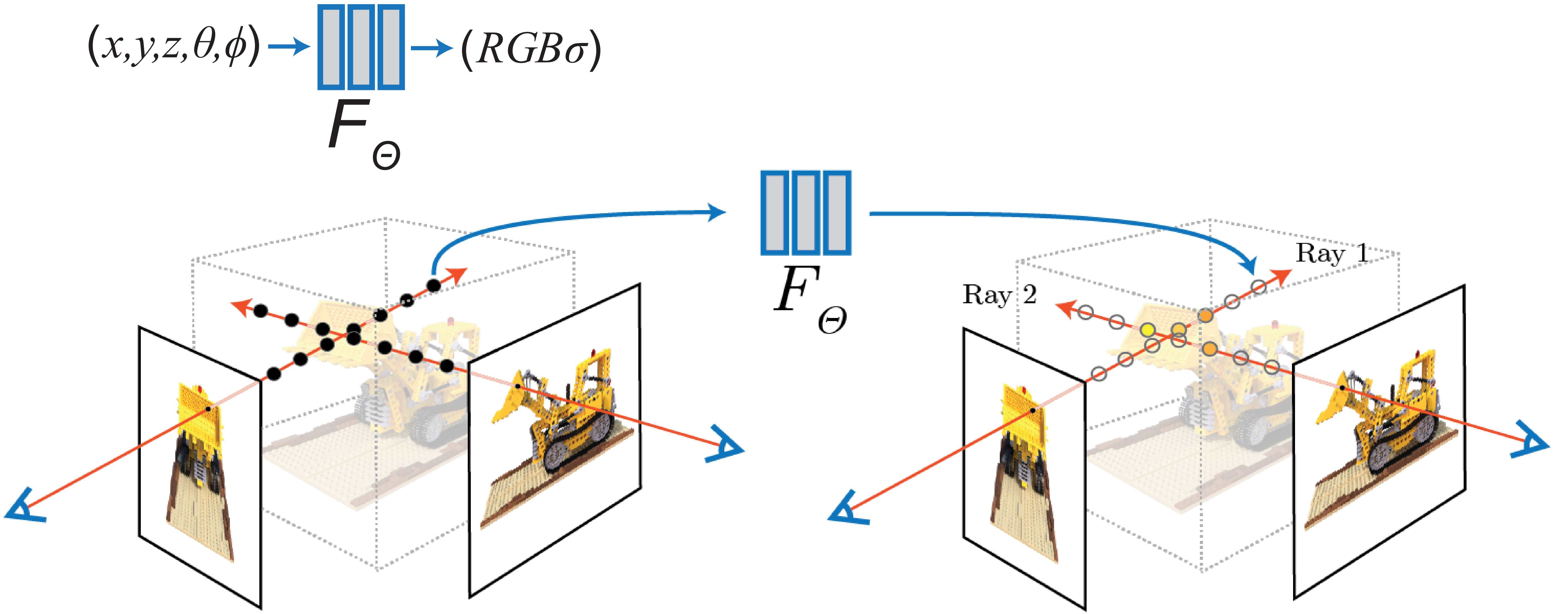
Filter out moving objects based on lack of brightness consistency.

Wimbauer et al., “MonoRec: Monocular Dense Reconstruction”, CVPR ‘21

Dense Reconstructions from a Single Camera

Wimbauer et al., “MonoRec: Monocular Dense Reconstruction”, CVPR ‘21

Neural Radiance Fields for Novel View Synthesis

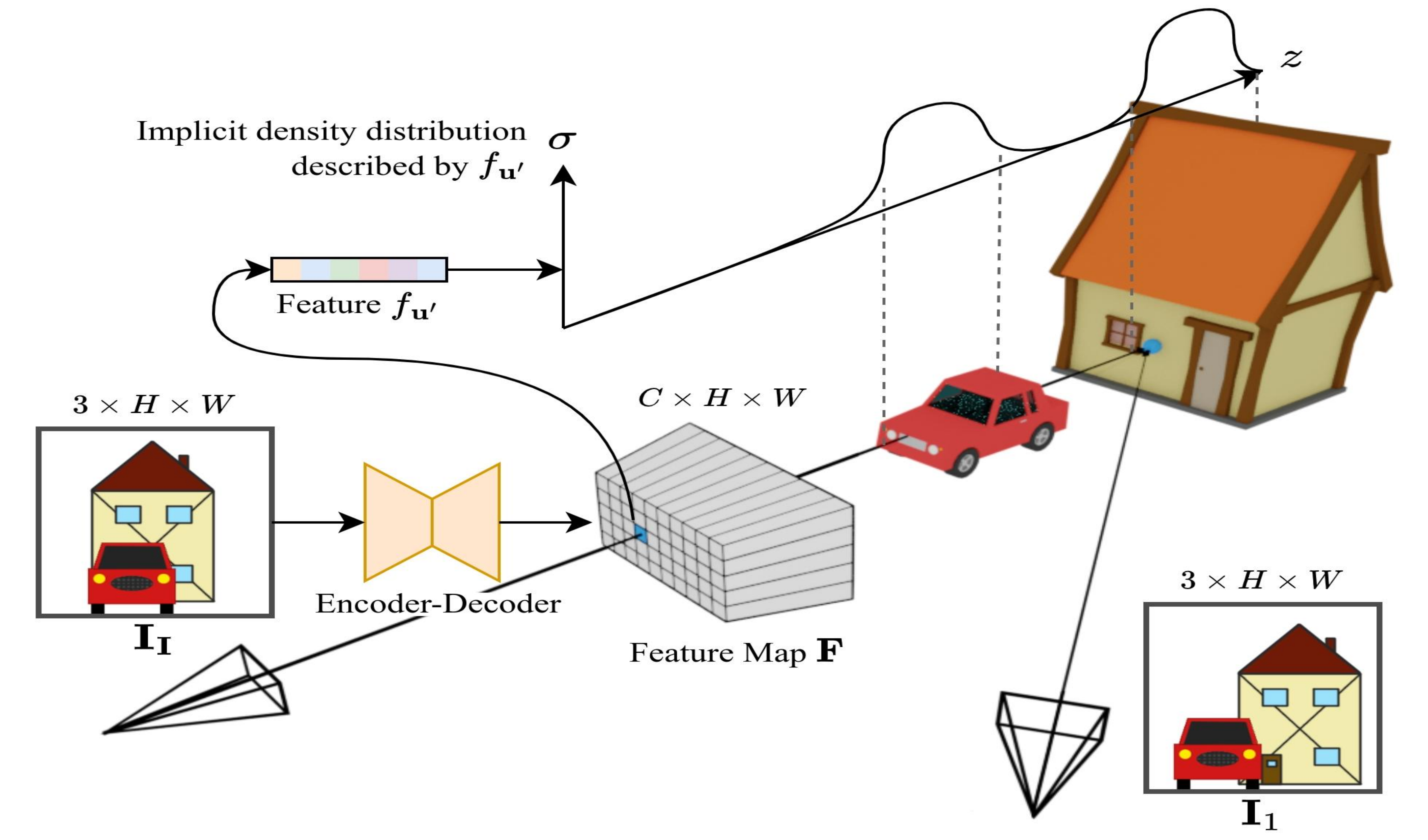


Mildenhall et al., “Neural Radiance Fields”, ECCV 2020

Neural Radiance Fields for Novel View Synthesis

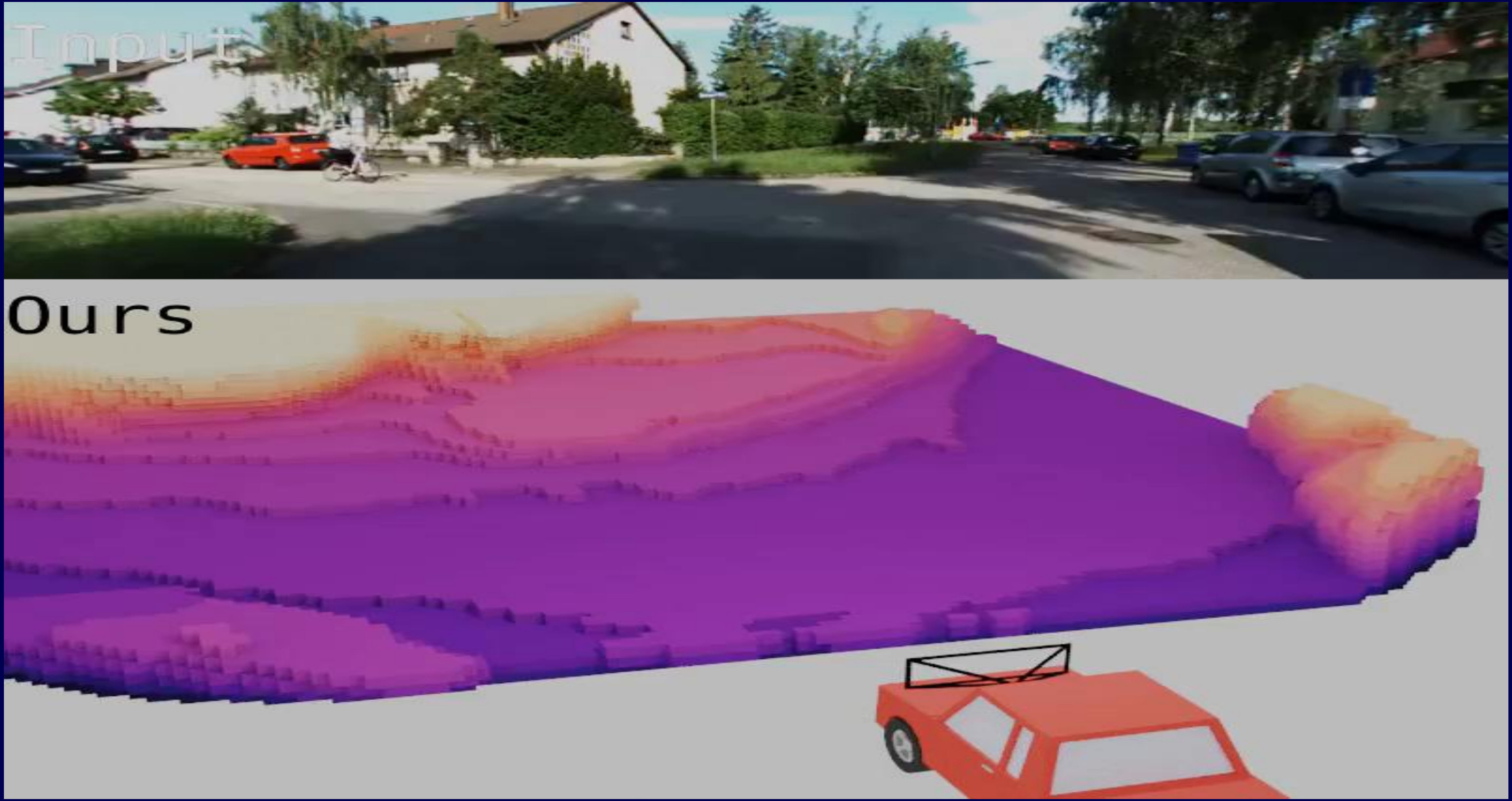
Mildenhall et al., "Neural Radiance Fields", ECCV 2020

Density-Fields for Single View Reconstruction



Wimbauer et al., "Behind the Scenes: Density Fields for Single View Reconstruction", CVPR '23

Density-Fields for Single View Reconstruction



Wimbauer et al., "Behind the Scenes: Density Fields for Single View Reconstruction", CVPR '23

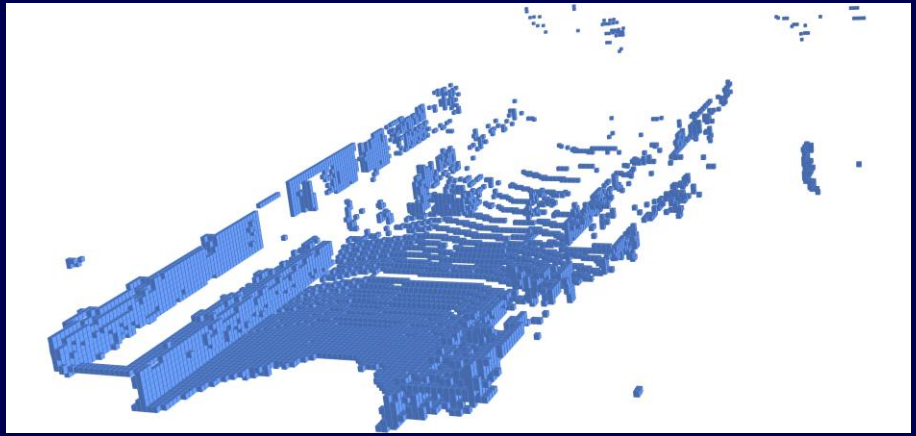
Novel View Synthesis Results

Wimbauer et al., “Behind the Scenes: Density Fields for Single View Reconstruction”, CVPR ‘23

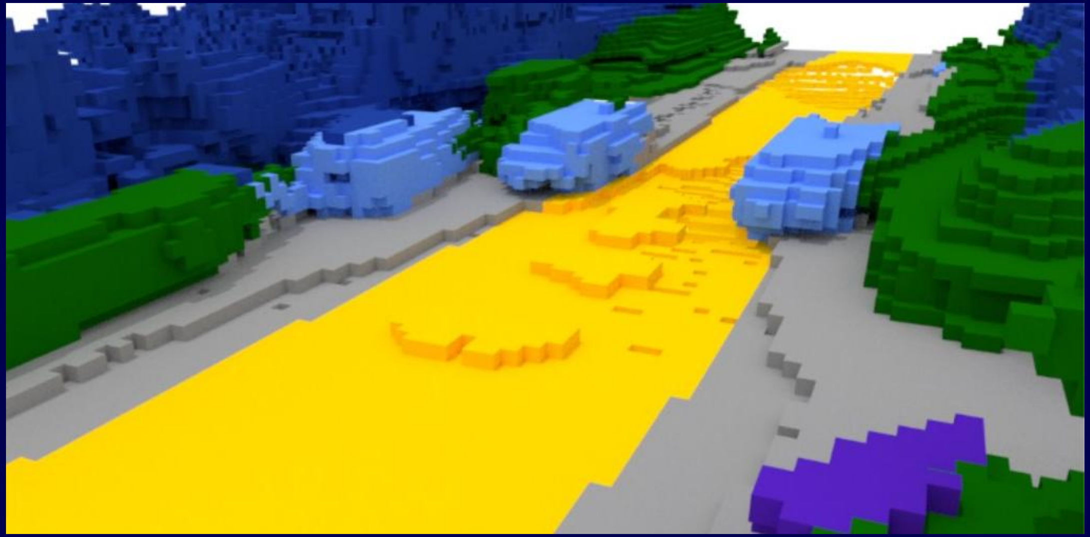
Novel View Synthesis Results

Wimbauer et al., "Behind the Scenes: Density Fields for Single View Reconstruction", CVPR '23

Color image



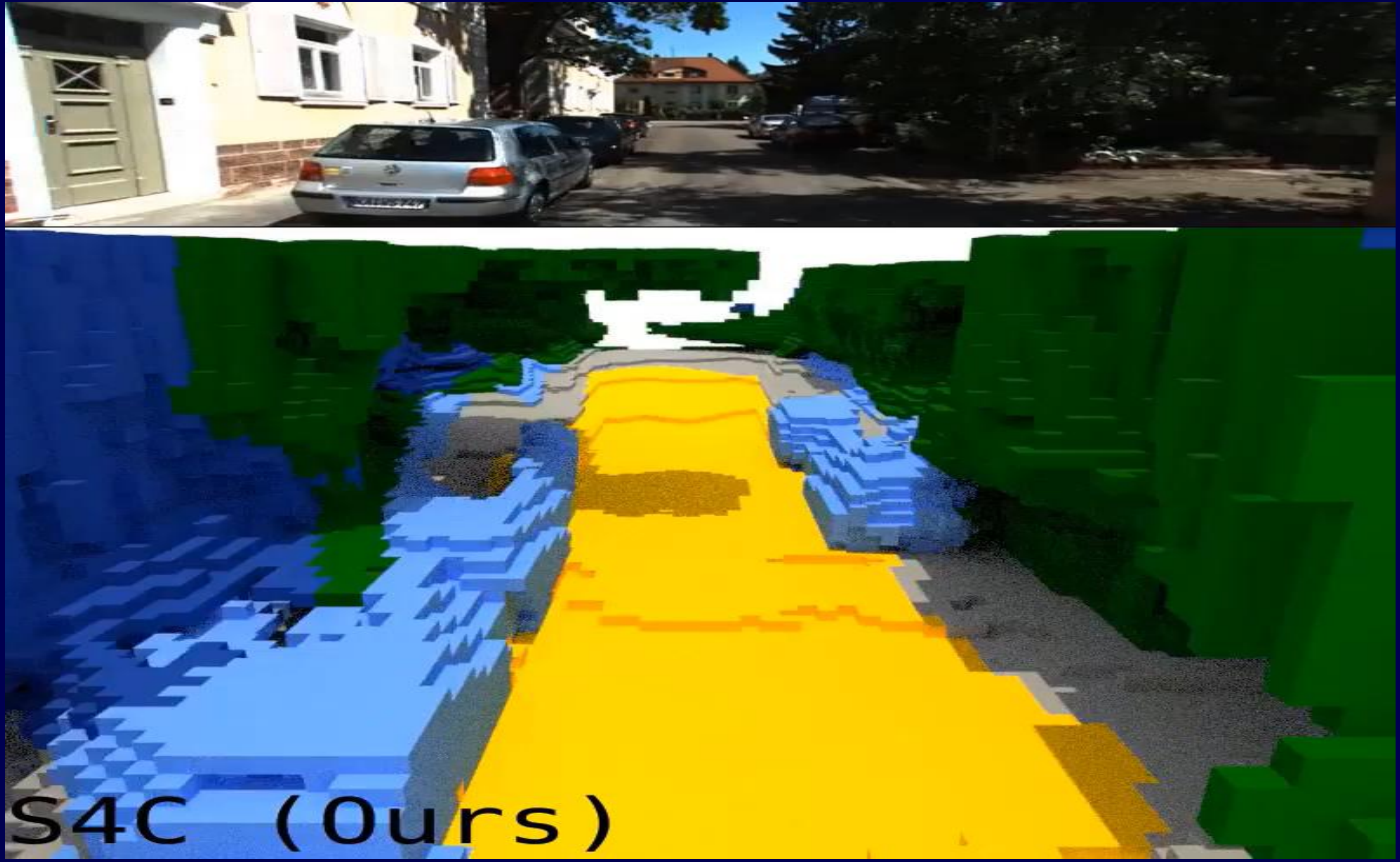
(Depth / Lidar)



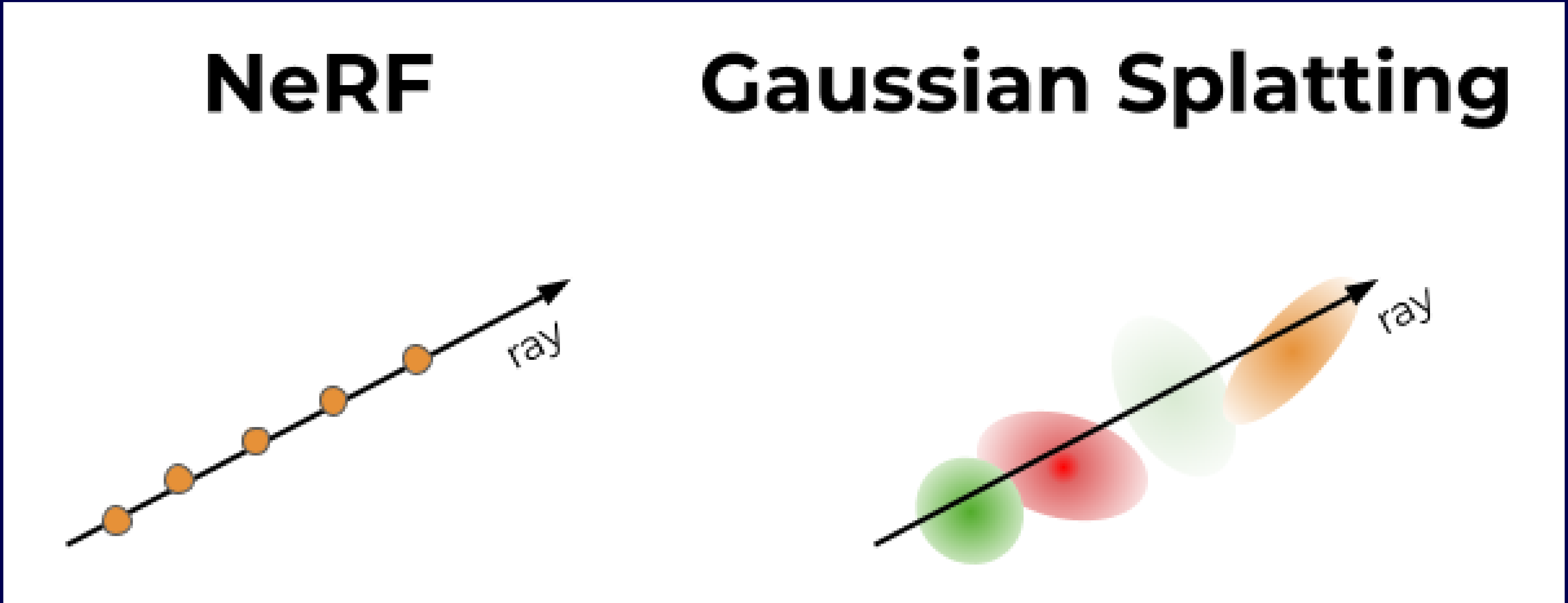
Complete **Geometry** and **Semantics**

Hayler et al., “S4C: Self-Supervised Semantic Scene Completion”, 3DV ‘24

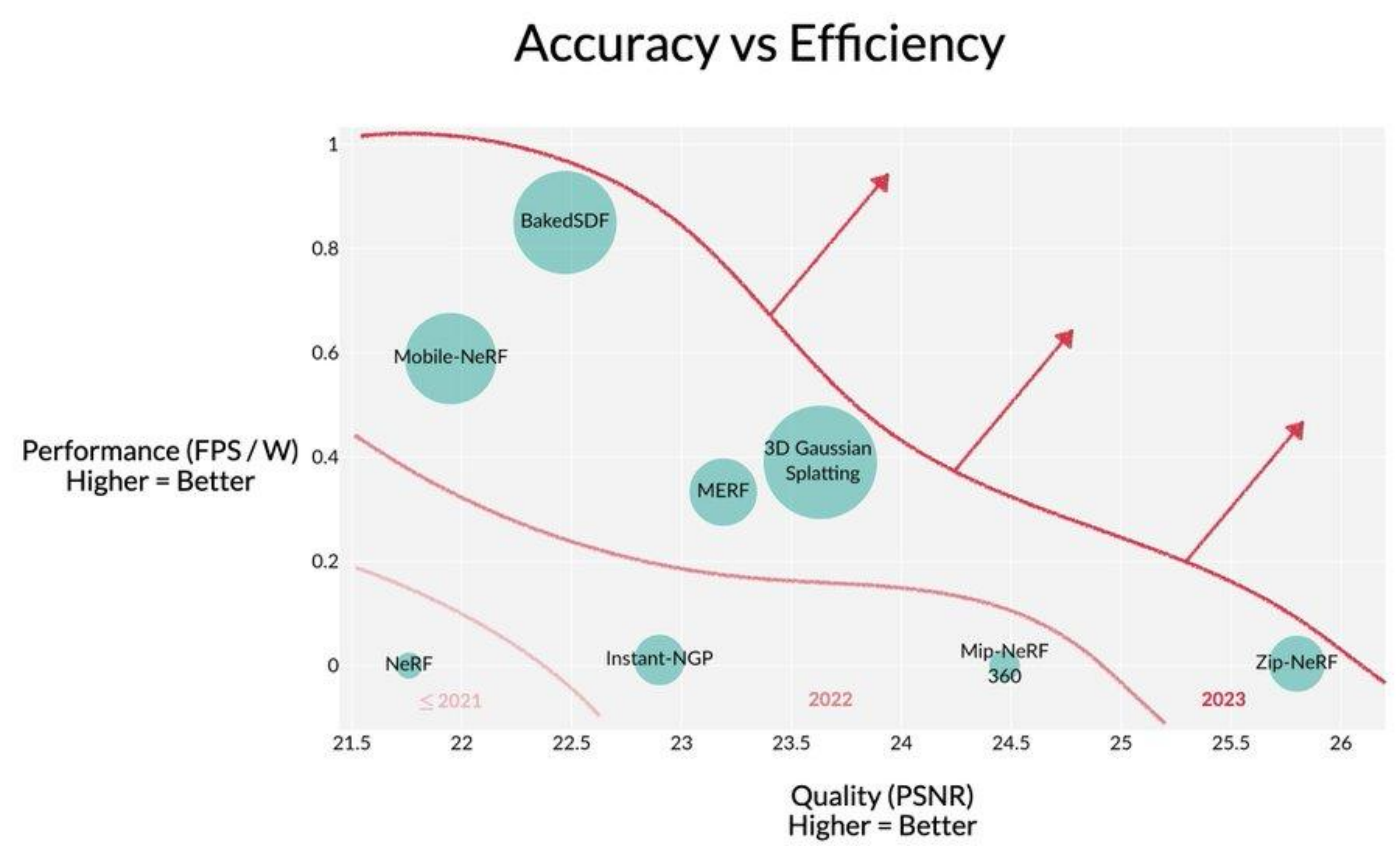
Self-supervised Semantic Scene Completion



Hayler et al., "S4C: Self-Supervised Semantic Scene Completion", 3DV '24



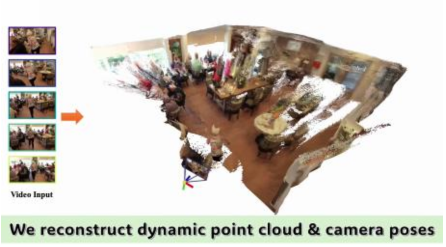
Lee Alan Westover 1991, Kerbl et al., "3D Gaussian Splatting", Siggraph 2023



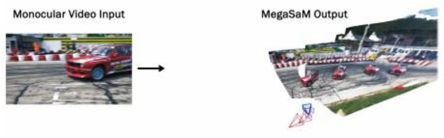
Source: Jon Barron

Reconstruction from Casual Videos

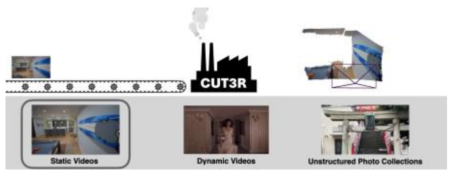
Monst3r (ICLR 2025)



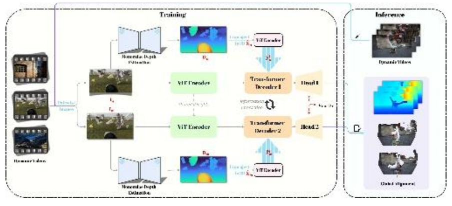
MegaSaM (CVPR 2025)



Cut3r (CVPR 2025)



Align3r (CVPR 2025)



Impressive results

but

Supervised training

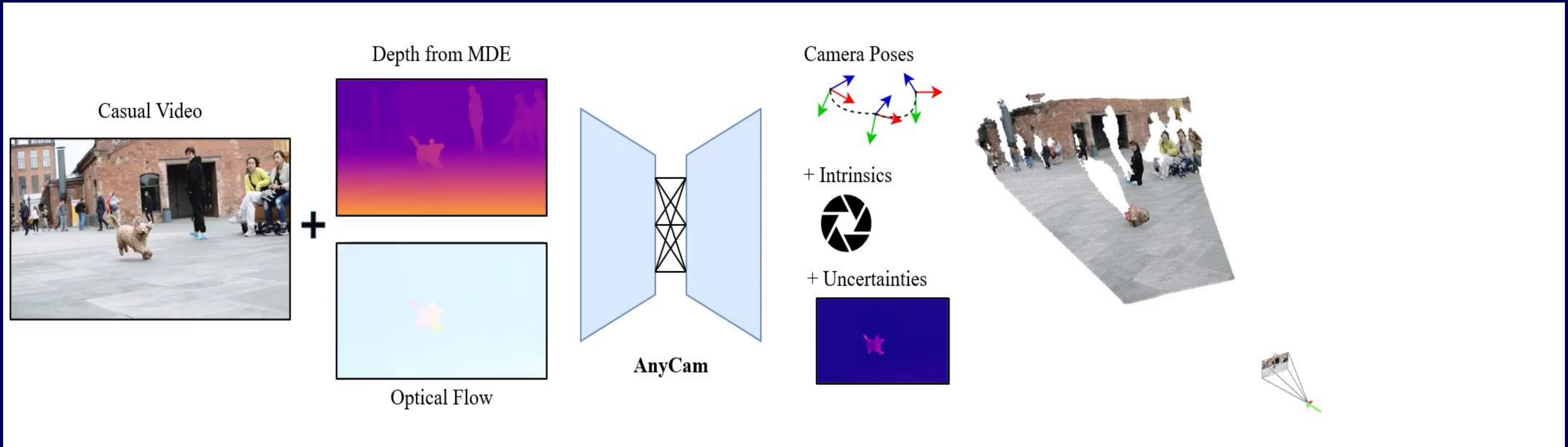
⚠️ Expensive data collection

⚠️ Limited datasets

⚠️ Dataset biases

⚠️ Sim-to-real gap

AnyCam: Reconstruction from Casual Videos



Felix Wimbauer[1,2,3] Weirong Chen[1,2,3] Dominik Muhle[1,2] Christian Rupprecht[3] Daniel Cremers[1,2]

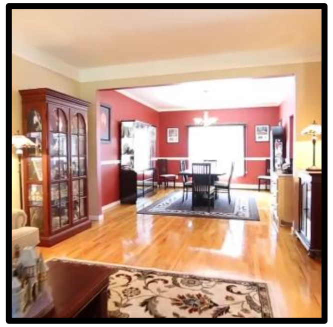
[1]Technical University of Munich [2]MCML [3]University of Oxford

Wimbauer et al., "AnyCam: Learning to Recover Camera Poses and Intrinsics from Casual Videos", CVPR '25

AnyCam: Reconstruction from Casual Videos

AnyCam is self-supervised on casual videos:

RealEstate10K



WalkingTours

OpenDV

YouTube VOS

EpicKitchens

From YouTube

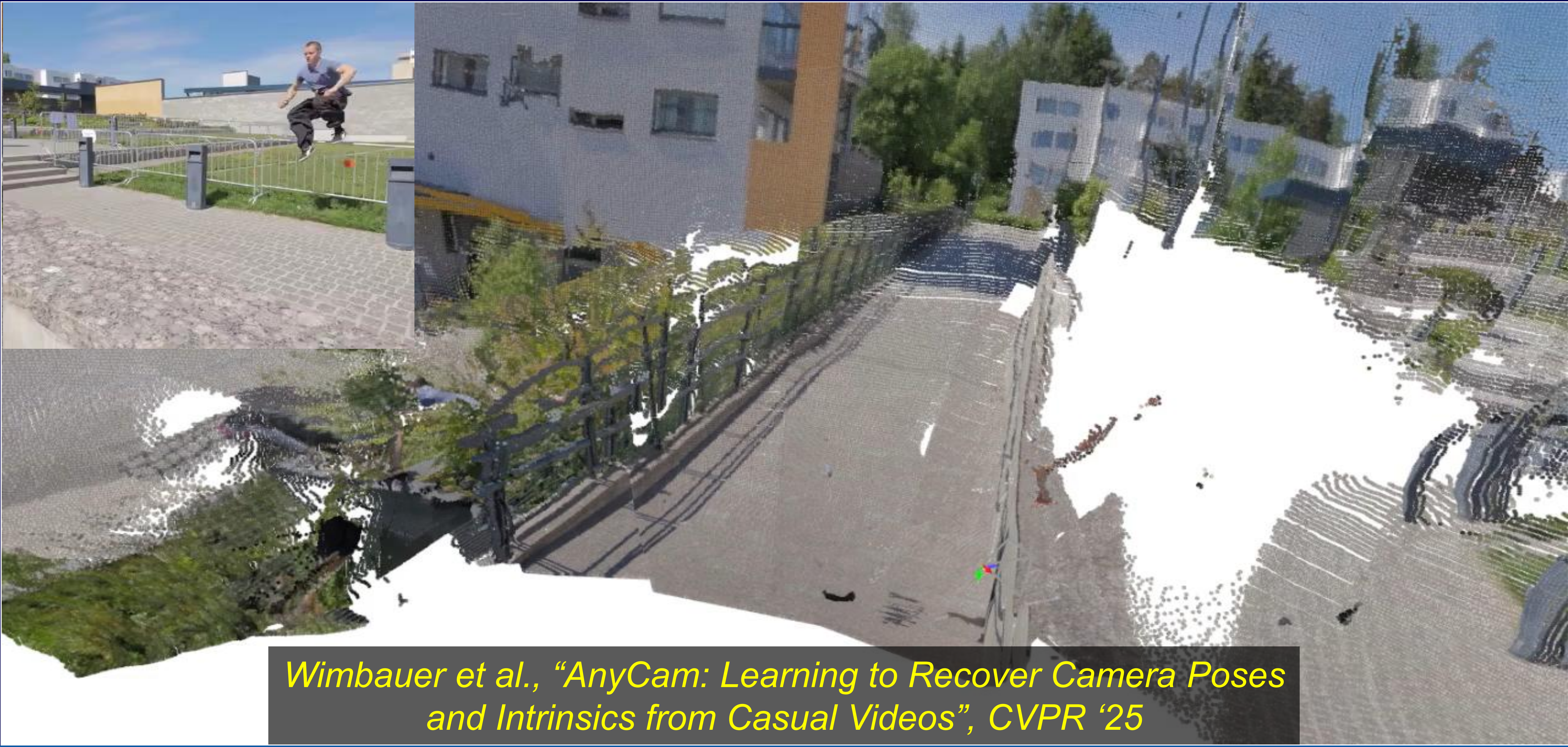
No ground truth data

Wimbauer et al., "AnyCam: Learning to Recover Camera Poses and Intrinsics from Casual Videos", CVPR '25

AnyCam: Reconstruction from Casual Videos

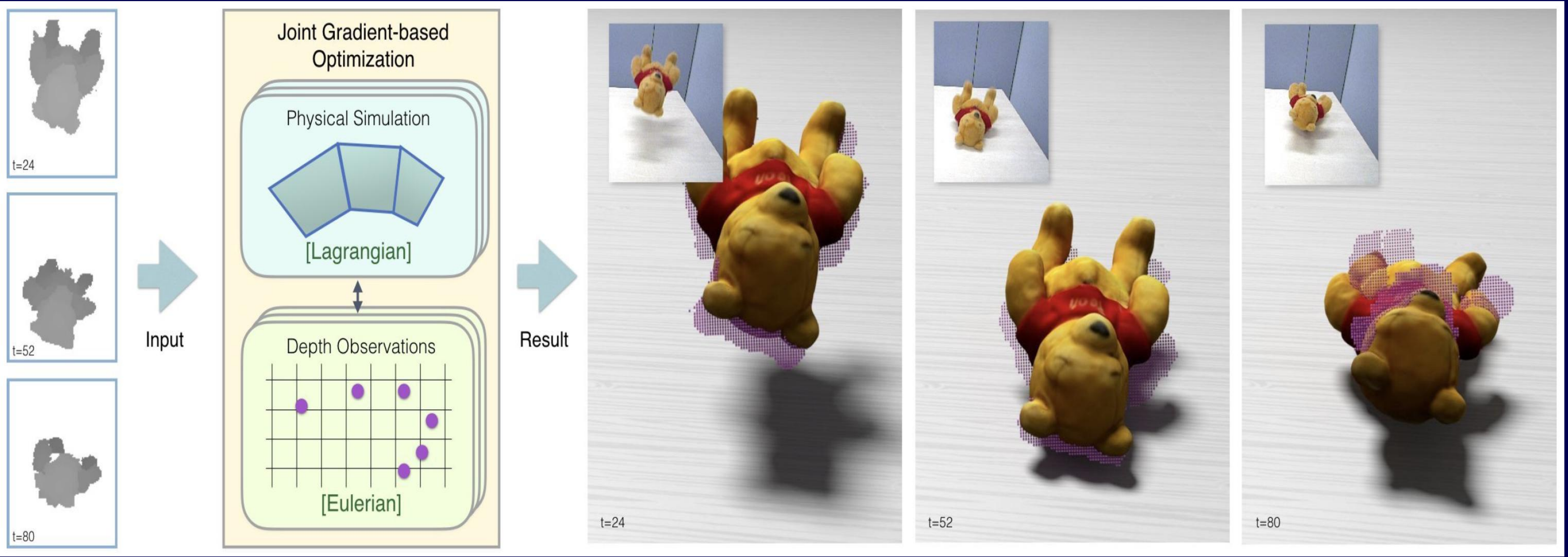
Wimbauer et al., "AnyCam: Learning to Recover Camera Poses and Intrinsics from Casual Videos", CVPR '25

AnyCam: Reconstruction from Casual Videos

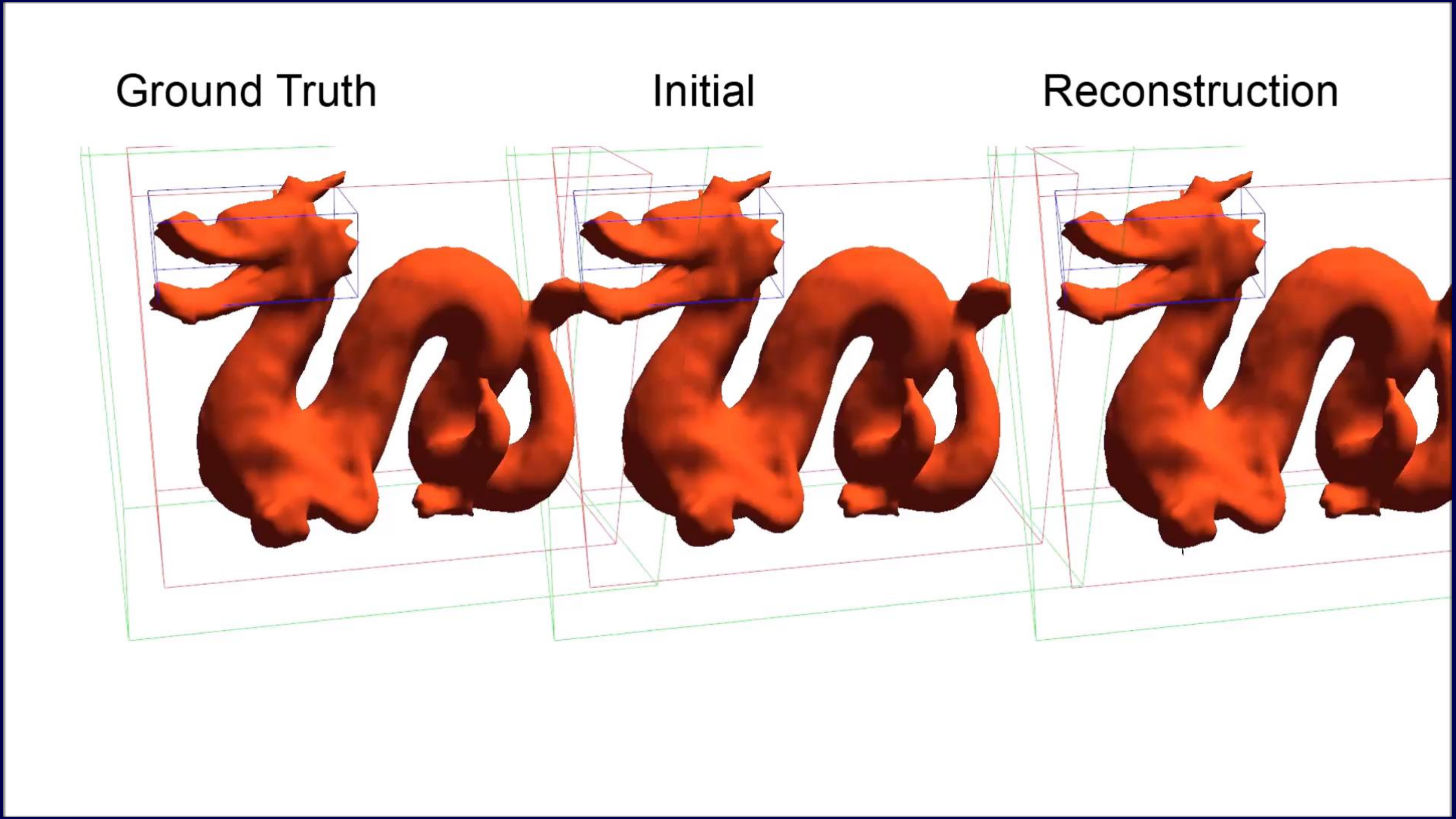


Wimbauer et al., "AnyCam: Learning to Recover Camera Poses and Intrinsics from Casual Videos", CVPR '25

Oswald, Stühmer, Cremers, ECCV '14



Weiss et al., CVPR 2020



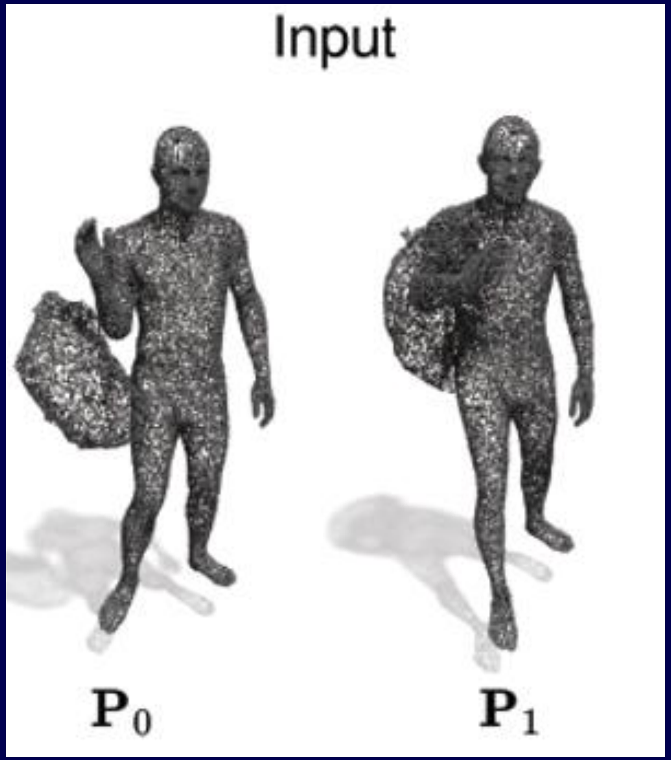
Weiss et al., CVPR 2020

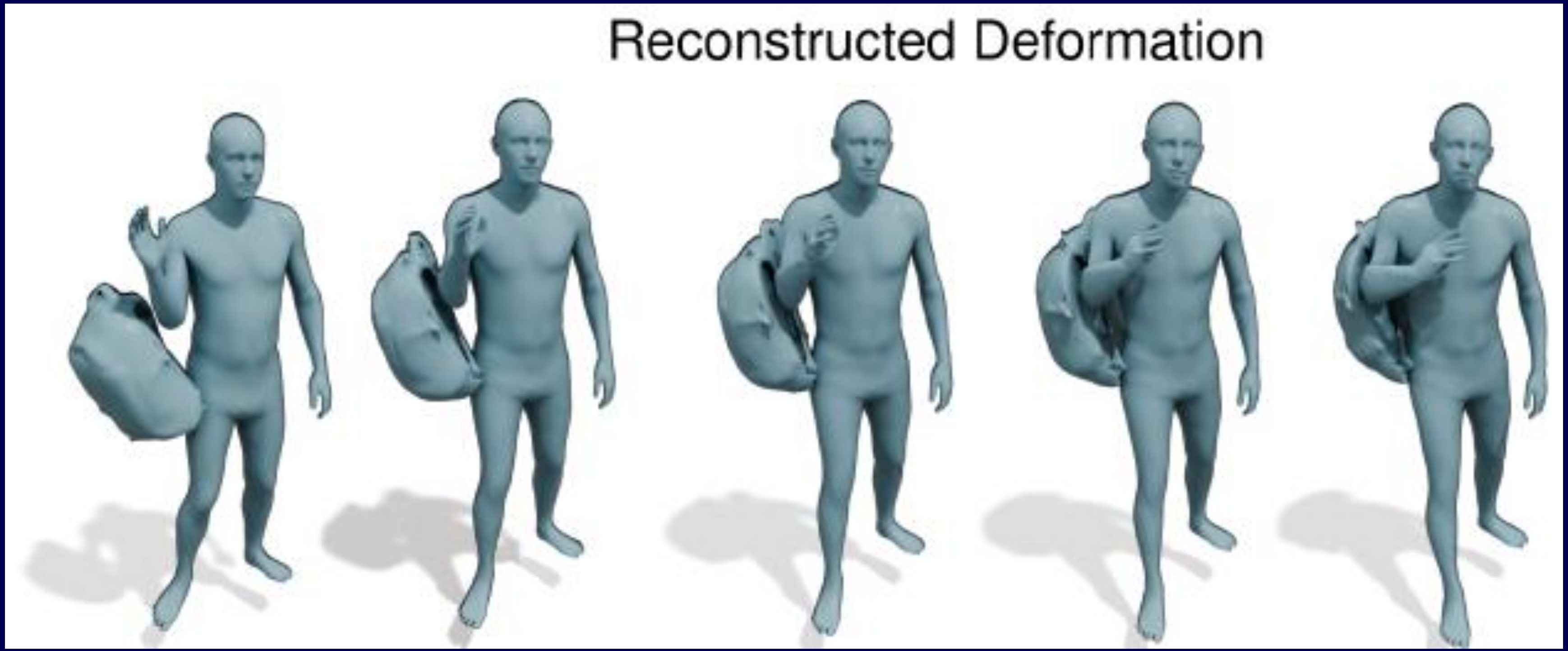
Observation with an RGB-D Camera

Color

Filtered Depth

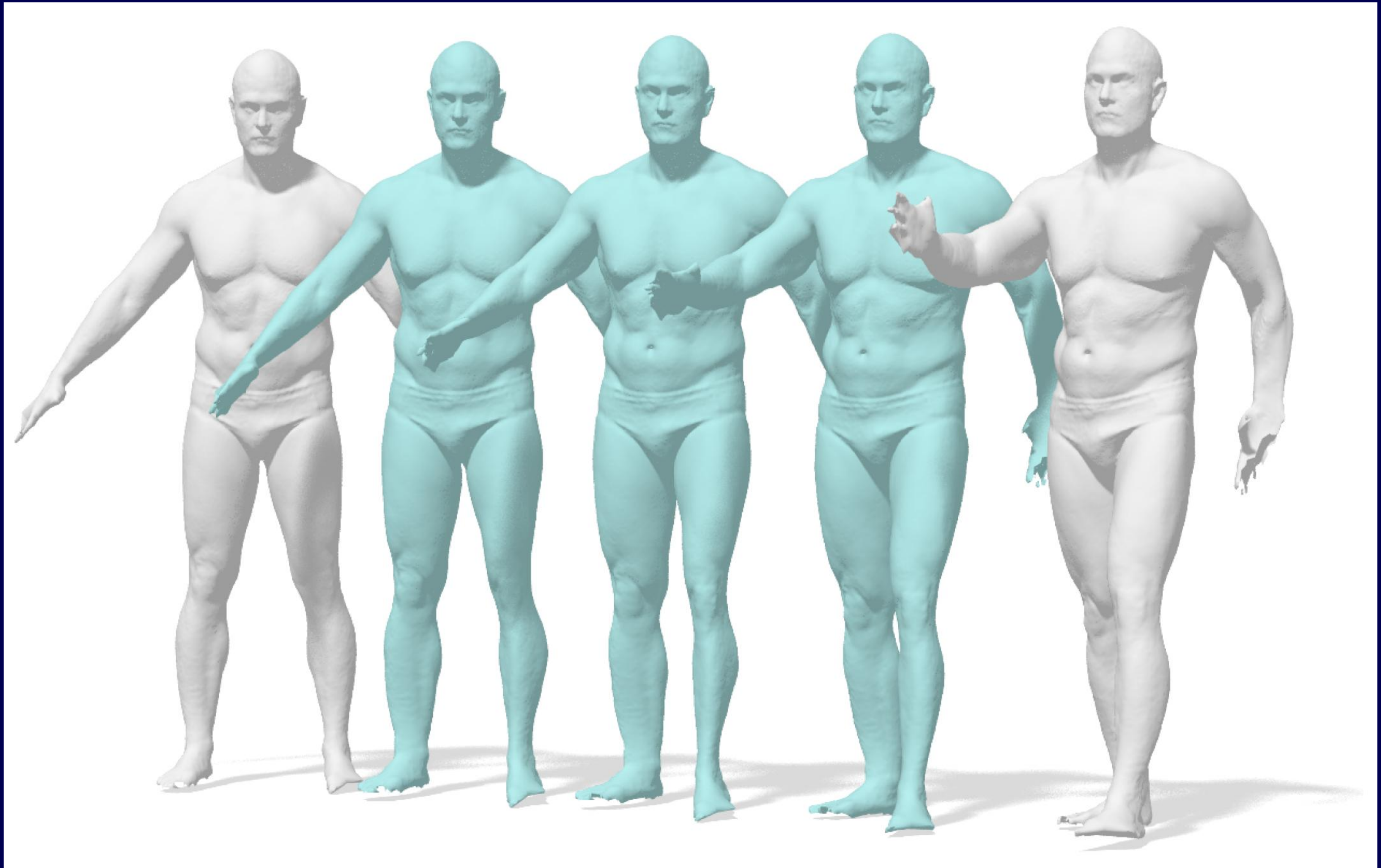
Weiss et al., CVPR 2020





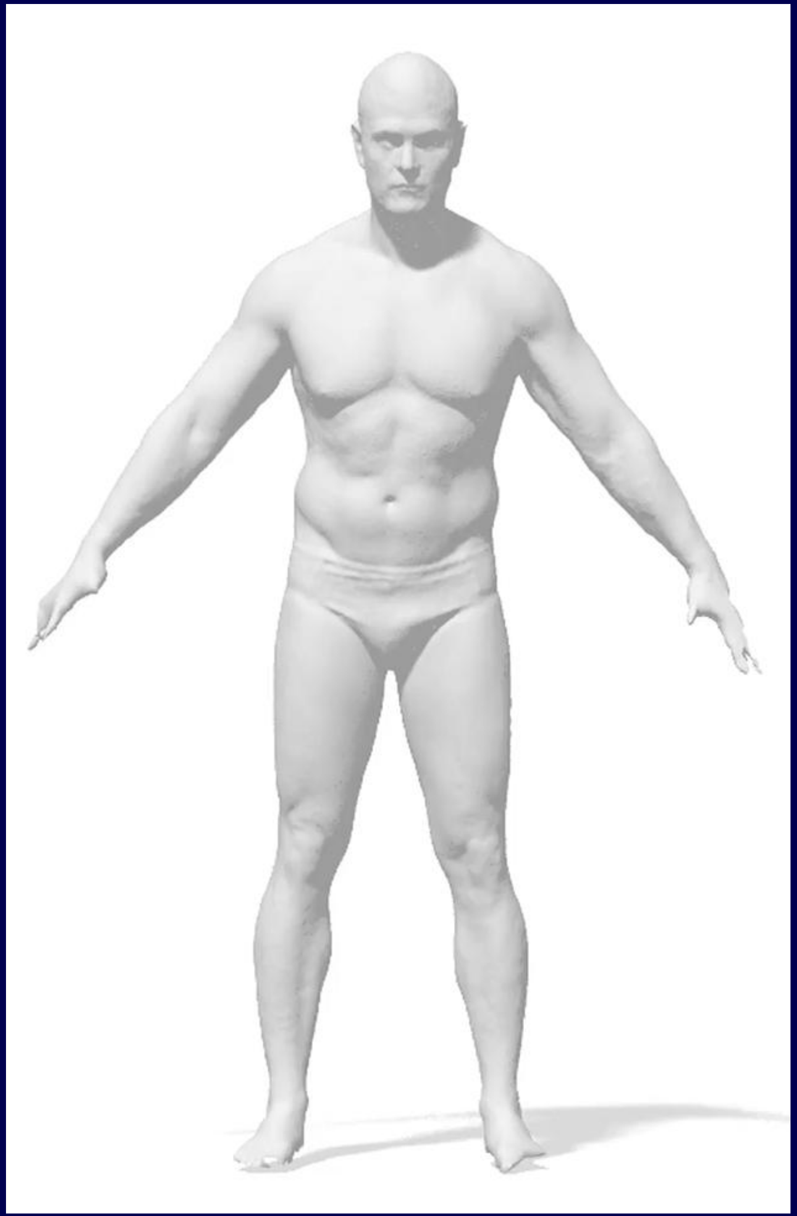
Sang et al., "4Deform: Neural Surface Deformation for Robust Shape Interpolation", CVPR '25

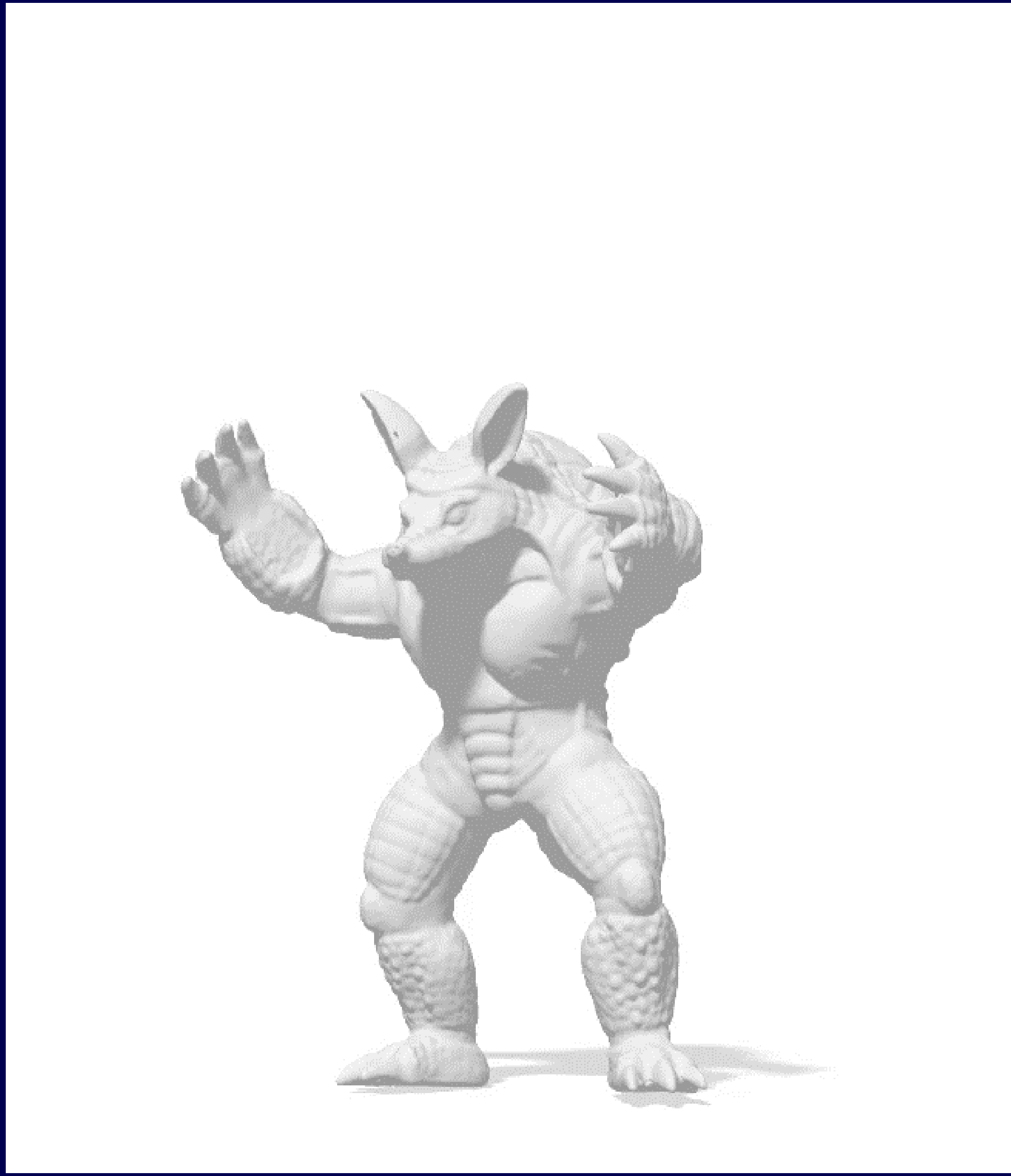
Volume-Preserving Shape Interpolation



Eisenberger, Laehner, Cremers, SGP 2019

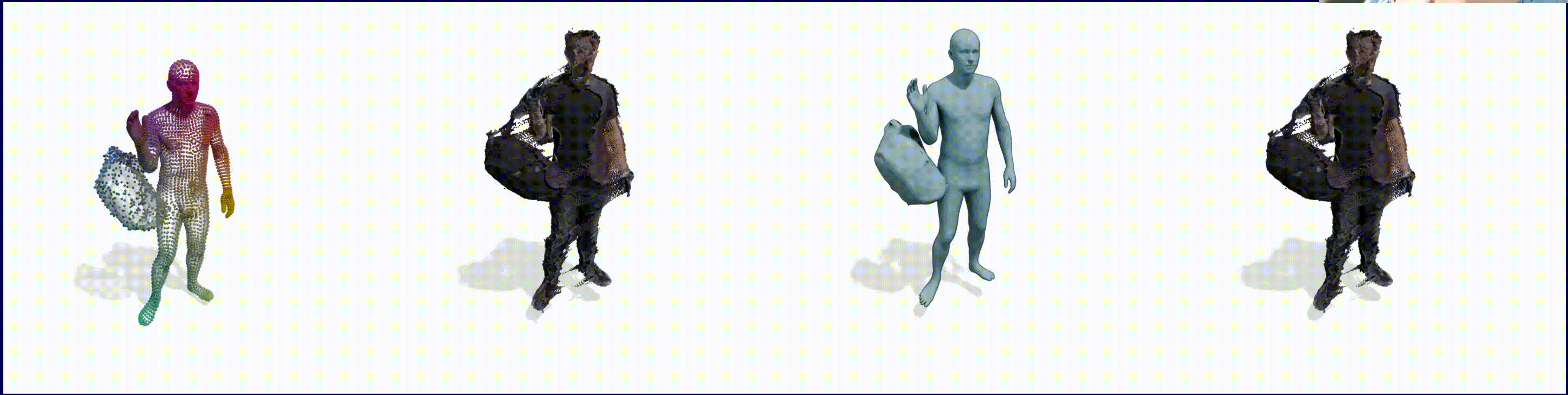
Volume-Preserving Shape Interpolation





Eisenberger, Laehner, Cremers, SGP 2019

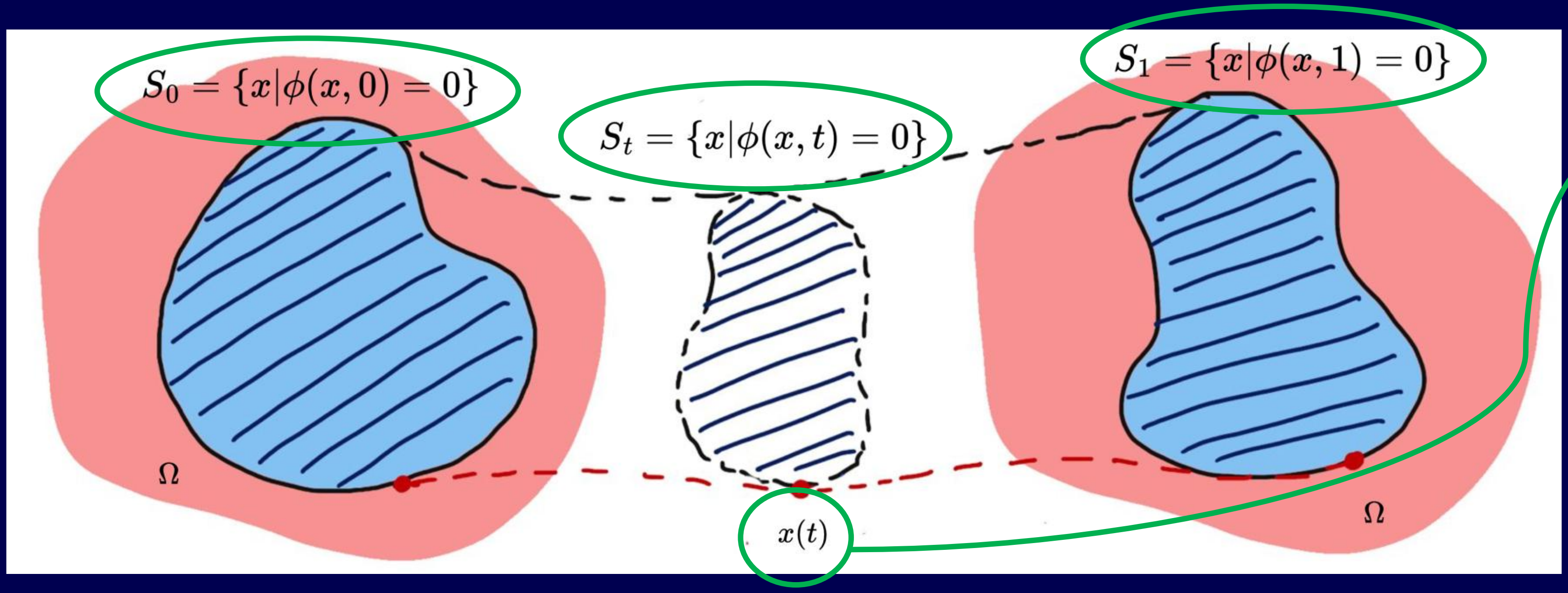
4D Reconstruction from Sparse Observations



keyframe point clouds | keyframe kinect data | generated intermediate frames

Sang et al., "4Deform: Neural Surface Deformation for Robust Shape Interpolation", CVPR '25

4D Reconstruction from Sparse Observations



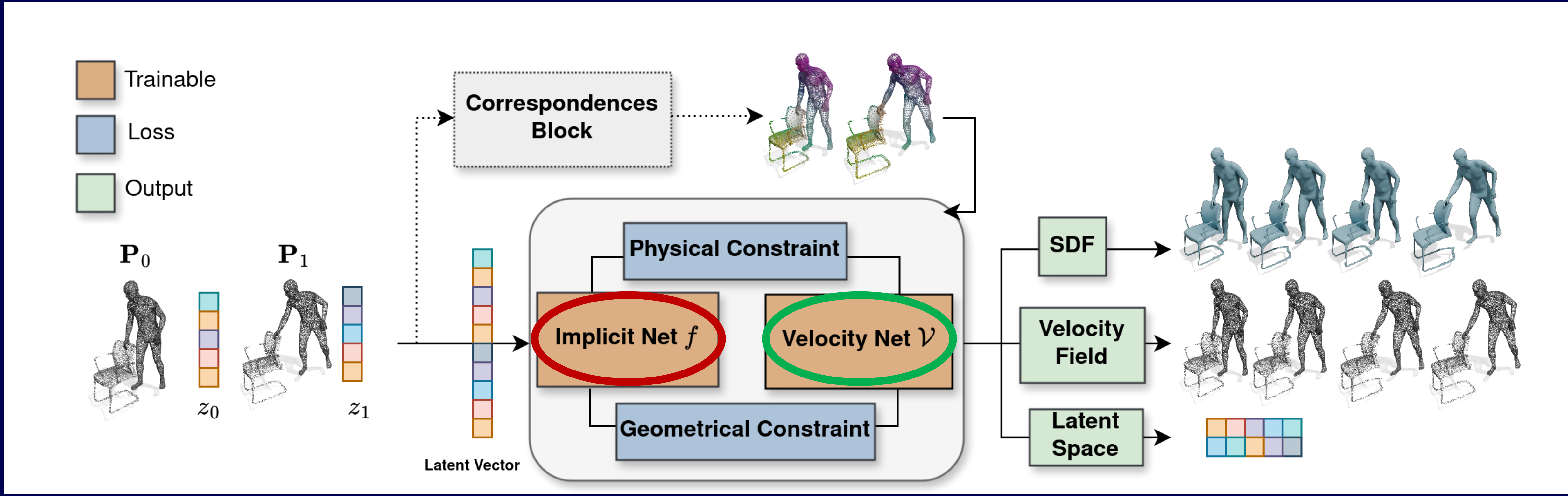
$$\phi(x(t),t) = 0$$

$$\frac{d}{dt}\phi(x,t) = \boxed{\partial_t\phi} + \boxed{V^\top}\boxed{\nabla\phi} = 0$$

$$\frac{d}{dt}x(t)$$

Sang et al., "4Deform: Neural Surface Deformation for Robust Shape Interpolation", CVPR '25

4D Reconstruction from Sparse Observations



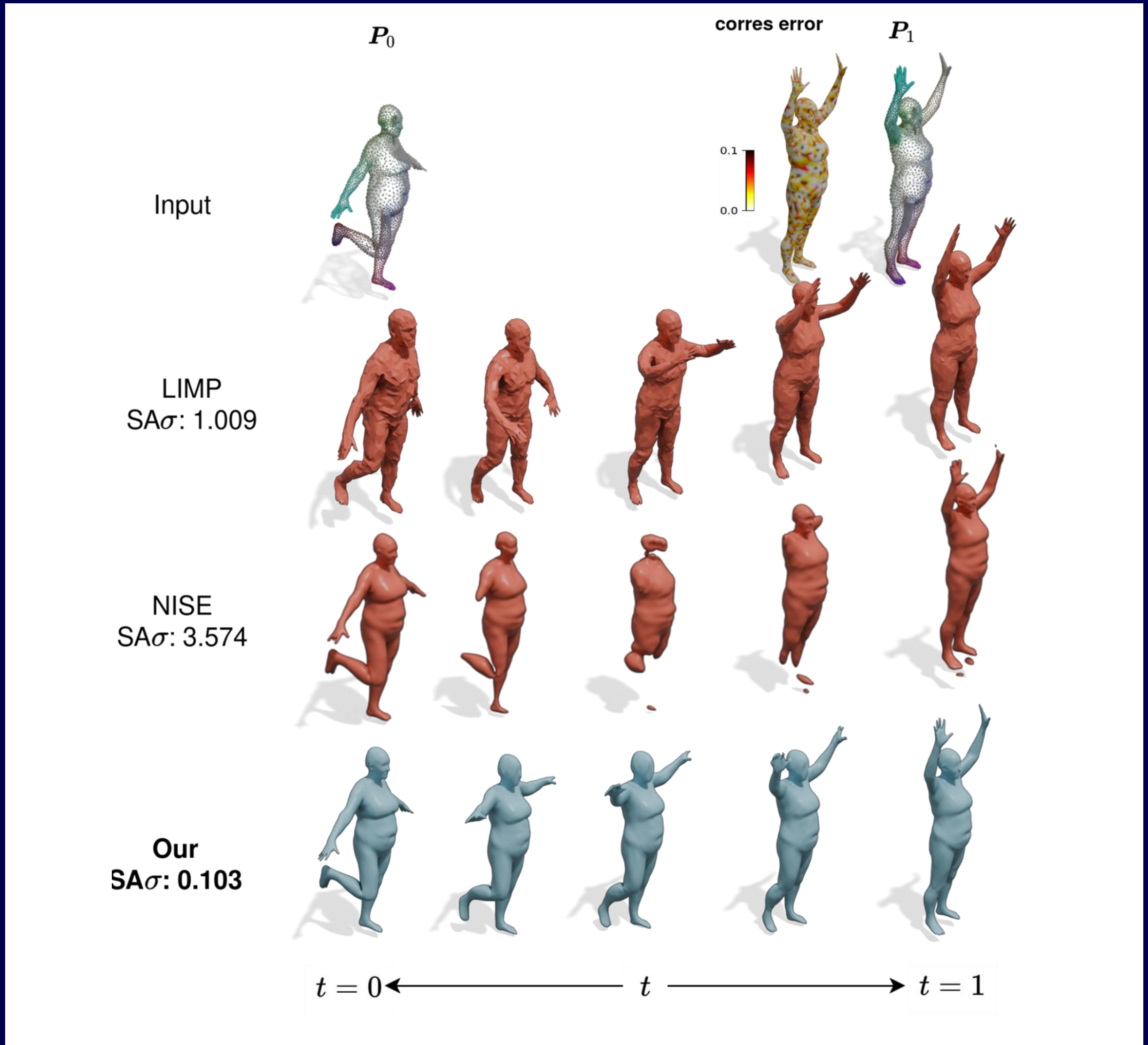
Geometrical Constraints
- Normal deformation constraint
- Level set equation constraint
- Matching loss

Physical Constraints
- Spatial smoothness velocity
- Volume preserving deformation
- Stretching constraint
- Distortion constraint

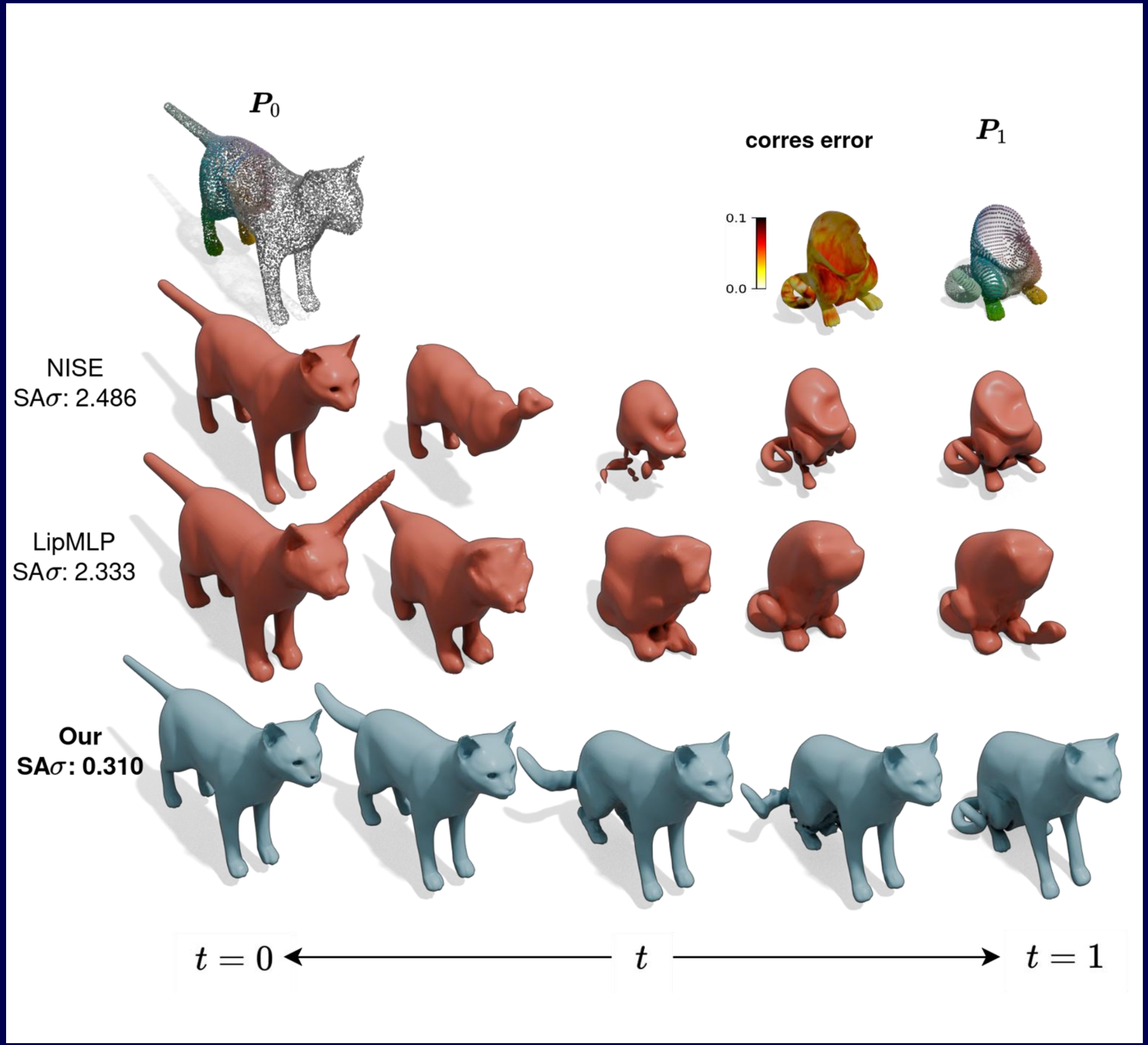
Sang et al., "4Deform: Neural Surface Deformation for Robust Shape Interpolation", CVPR '25

4D Reconstruction from Sparse Observations

Large deformation



Partial shape deformation

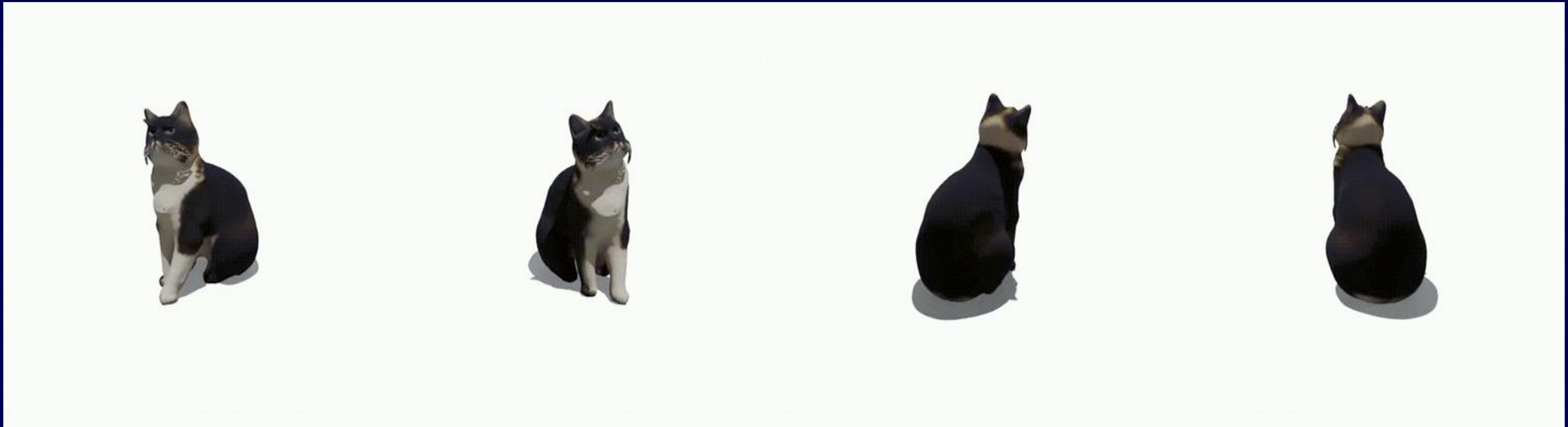


Sang et al., "4Deform: Neural Surface Deformation for Robust Shape Interpolation", CVPR '25

TwoSquared: 4D Generation from 2D Image Pairs

2D input images:

Lu Sang

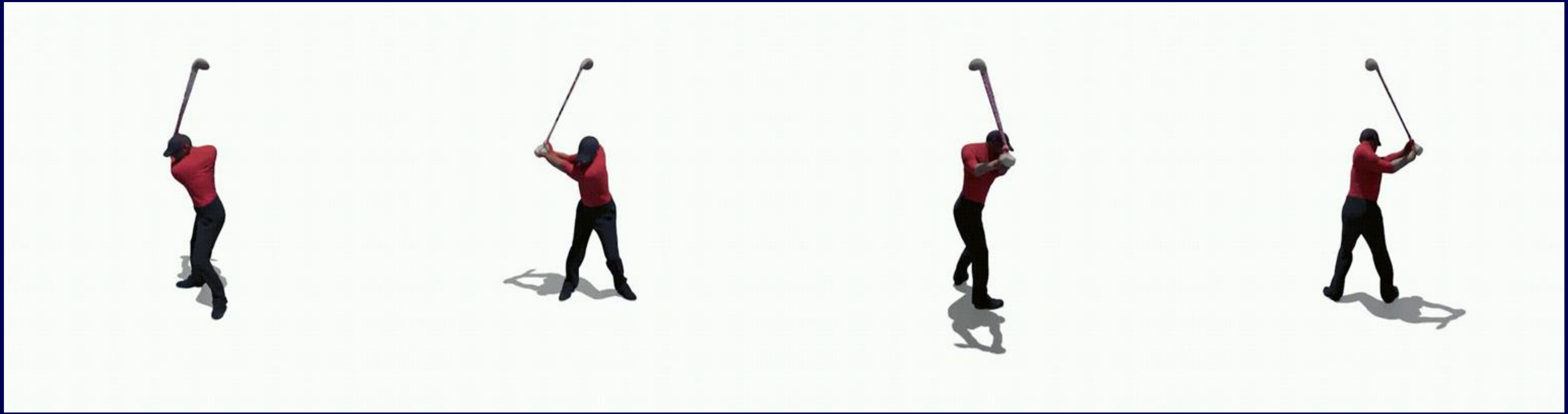


Sang et al., "TwoSquared: 4D Generation from 2D Image Pairs", arxiv '25

TwoSquared: 4D Generation from 2D Image Pairs

Lu Sang

2D input images:



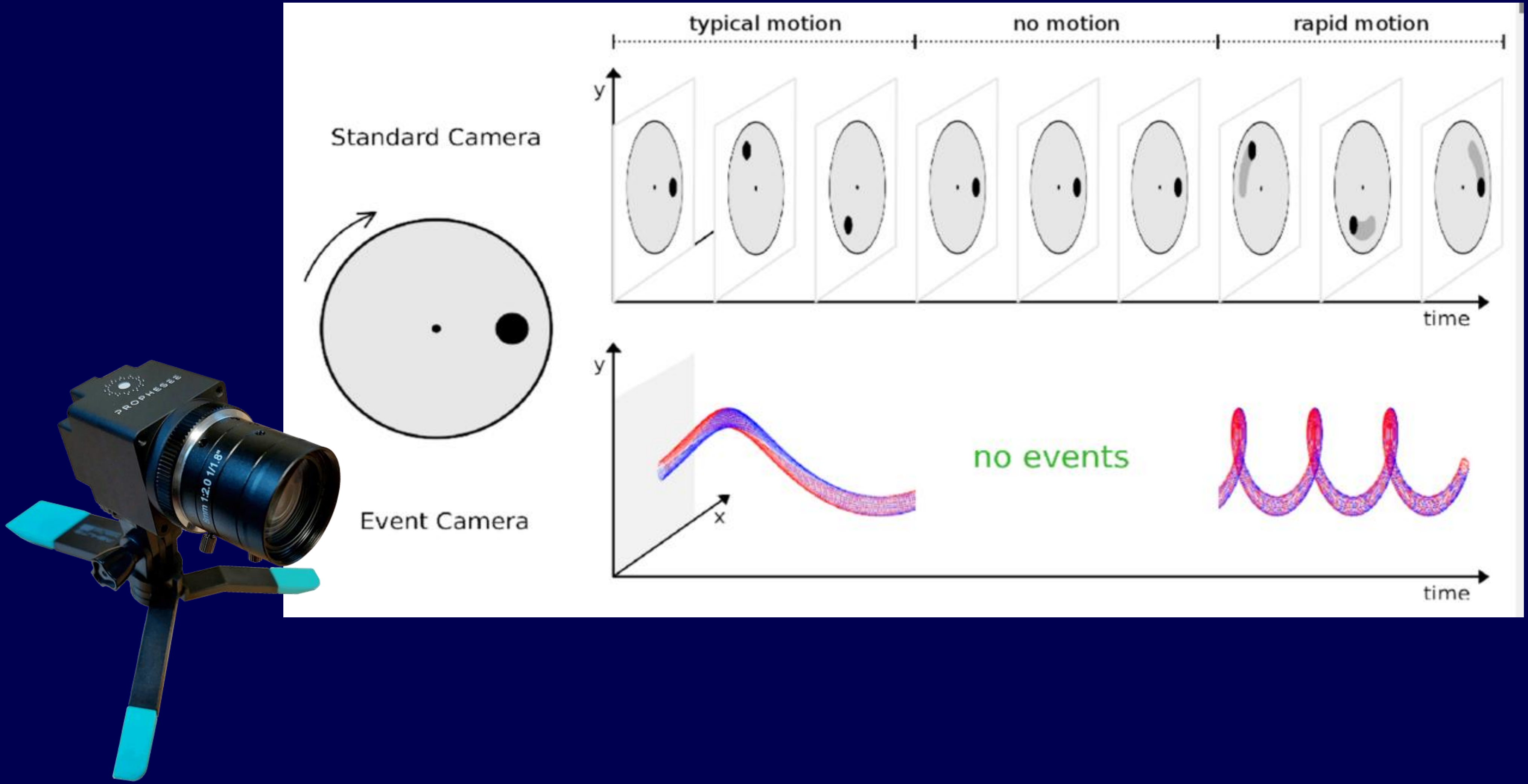
Sang et al., "TwoSquared: 4D Generation from 2D Image Pairs", arxiv '25

TwoSquared: 4D Generation from 2D Image Pairs

Lu Sang

Sang et al., "TwoSquared: 4D Generation from 2D Image Pairs", arxiv '25

Event Cameras



Lichtsteiner, Posch, Dellbrück 2006, IEEE J. Solid-State Circuits 2008

Can we achieve high-precision monocular event-only visual odometry?

D Weikersdorfer, DB Adrian, D Cremers, J Conradt, “Event-based 3D SLAM with a depth-augmented dynamic vision sensor”, ICRA 2014.

H Kim, S Leutenegger, AJ Davison, “Real-Time 3D Reconstruction and 6-DoF Tracking with an Event Camera”, ECCV 2016.

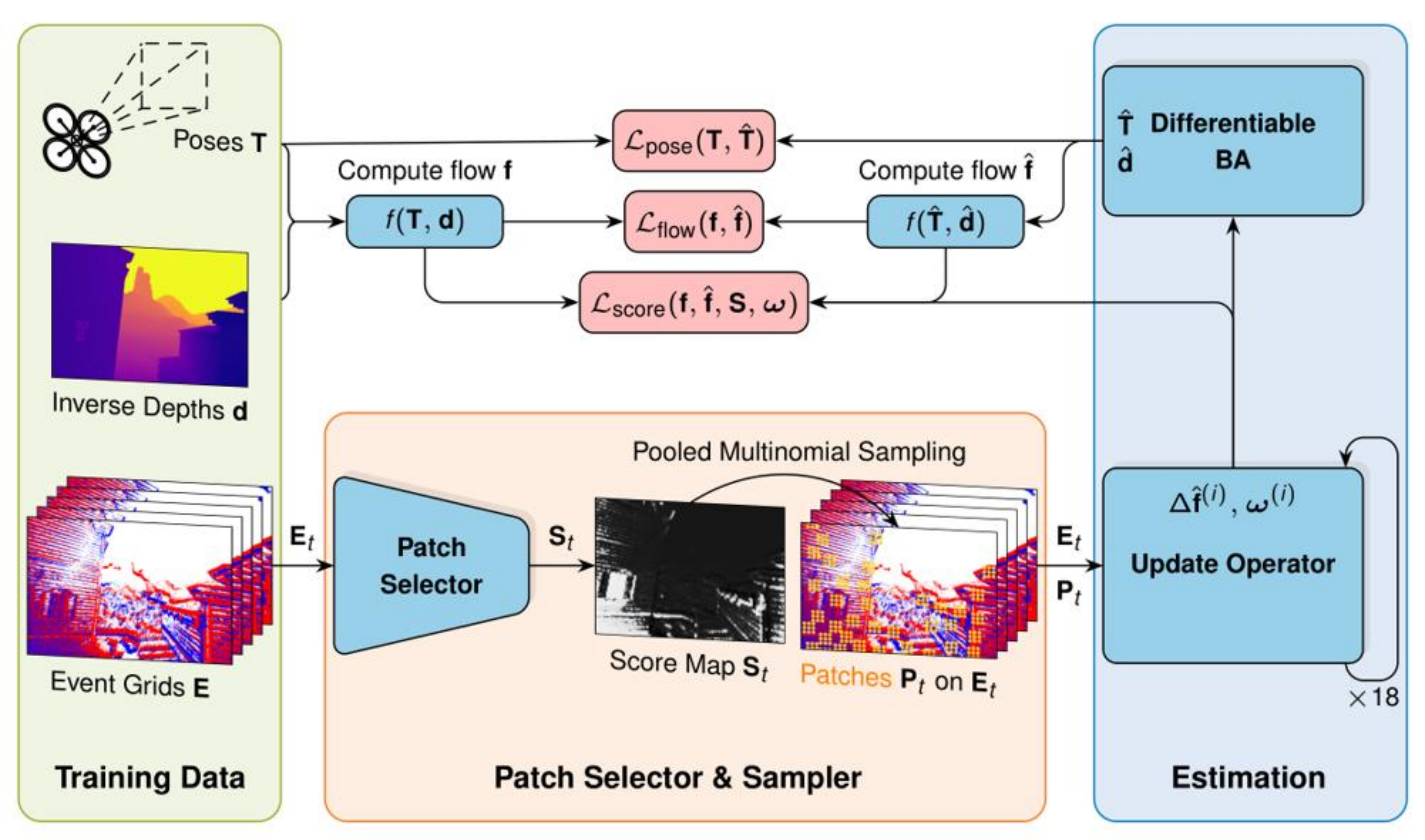
H Rebecq, T Horstschaefer, G Gallego, D Scaramuzza, “EVO: A Geometric Approach to Event-based 6-DOF Parallel Tracking and Mapping in Real-time”, RAL 2017.

AR Vidal, H Rebecq, T Horstschaefer, D Scaramuzza, “Ultimate SLAM? Combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios”, RAL 2018.

W Guan, P Chen, Y Xie, P Lu, “PI-evio: Robust monocular event-based visual inertial odometry with point and line features”, IEEE Aut. Sci. Eng. 2022.

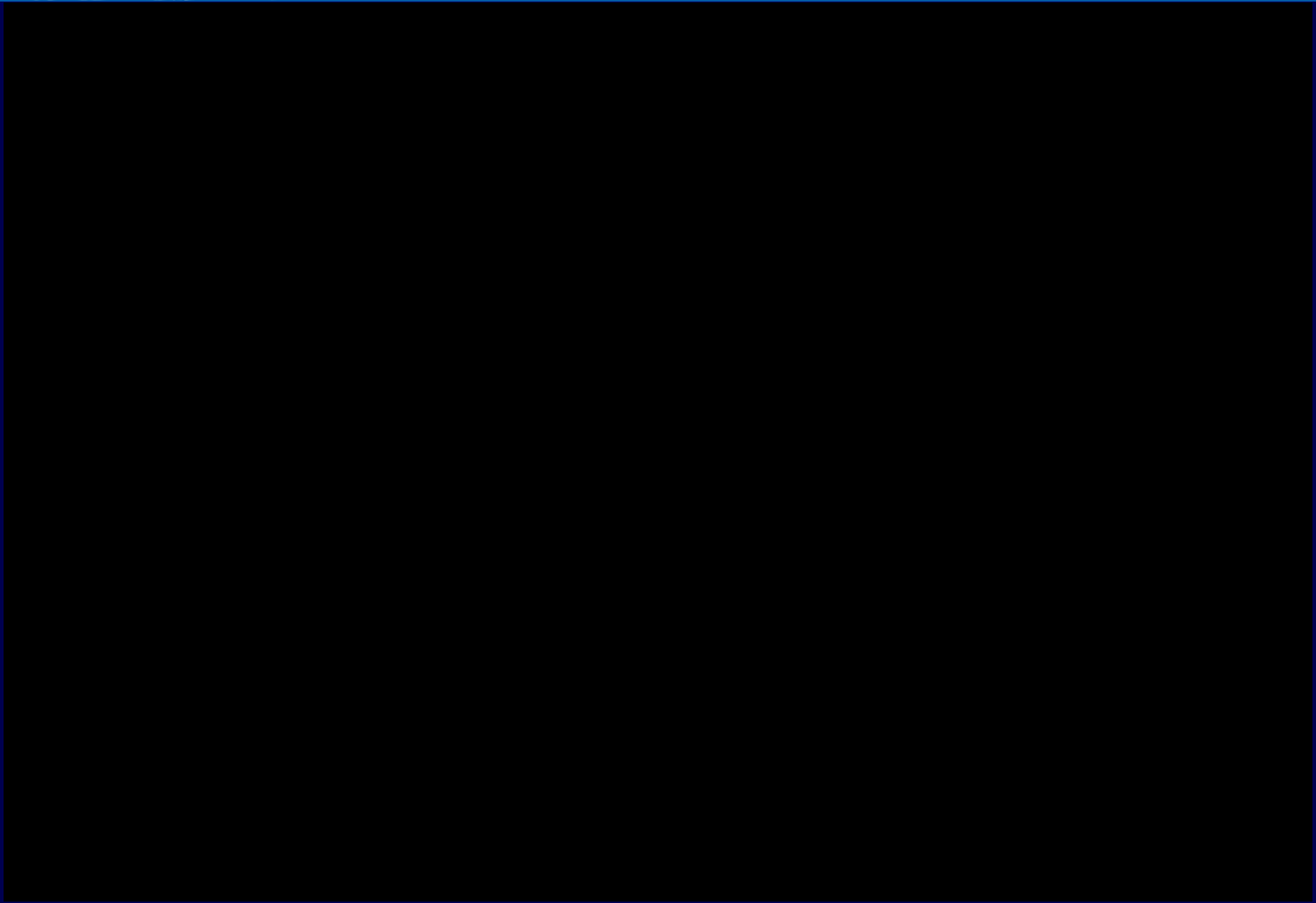
P Chen, W Guan, P Lu, “Esvio: Event-based stereo visual inertial odometry”, IEEE Aut. Sci. Eng. 2023.

DEVO: Deep Event Visual Odometry



Klenk, Motzet, Koestler, Cremers, "Deep Event Visual Odometry", 3DV 2024

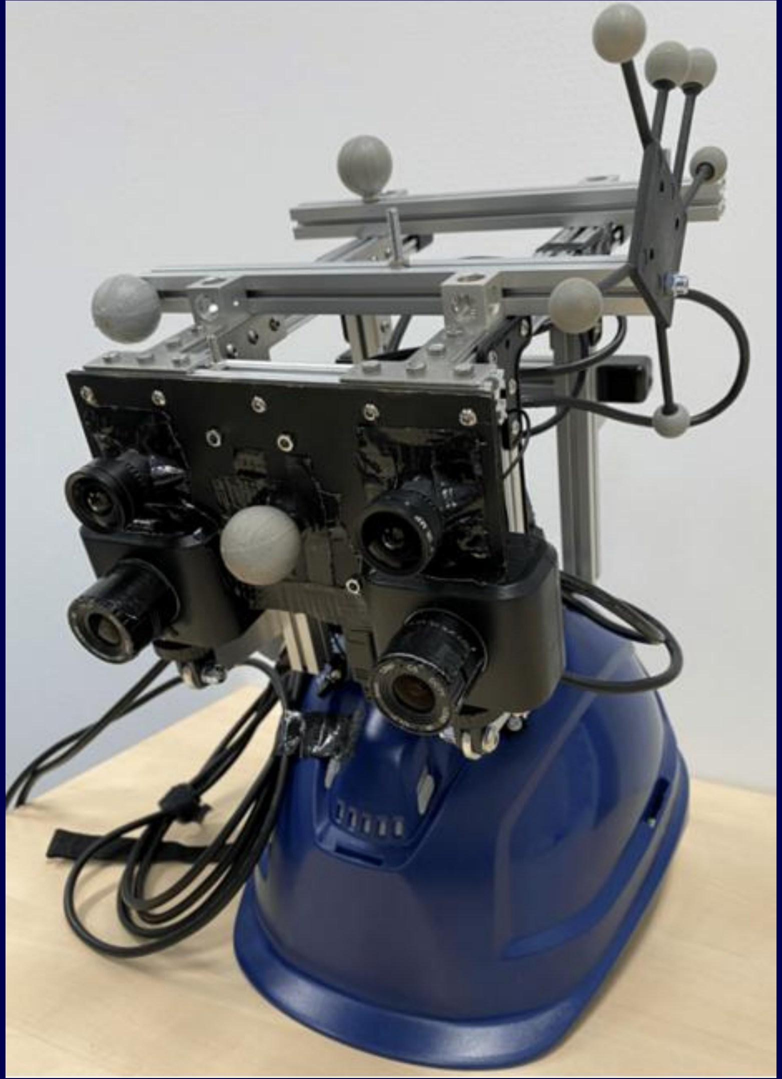
DEVO: Qualitative Evaluation

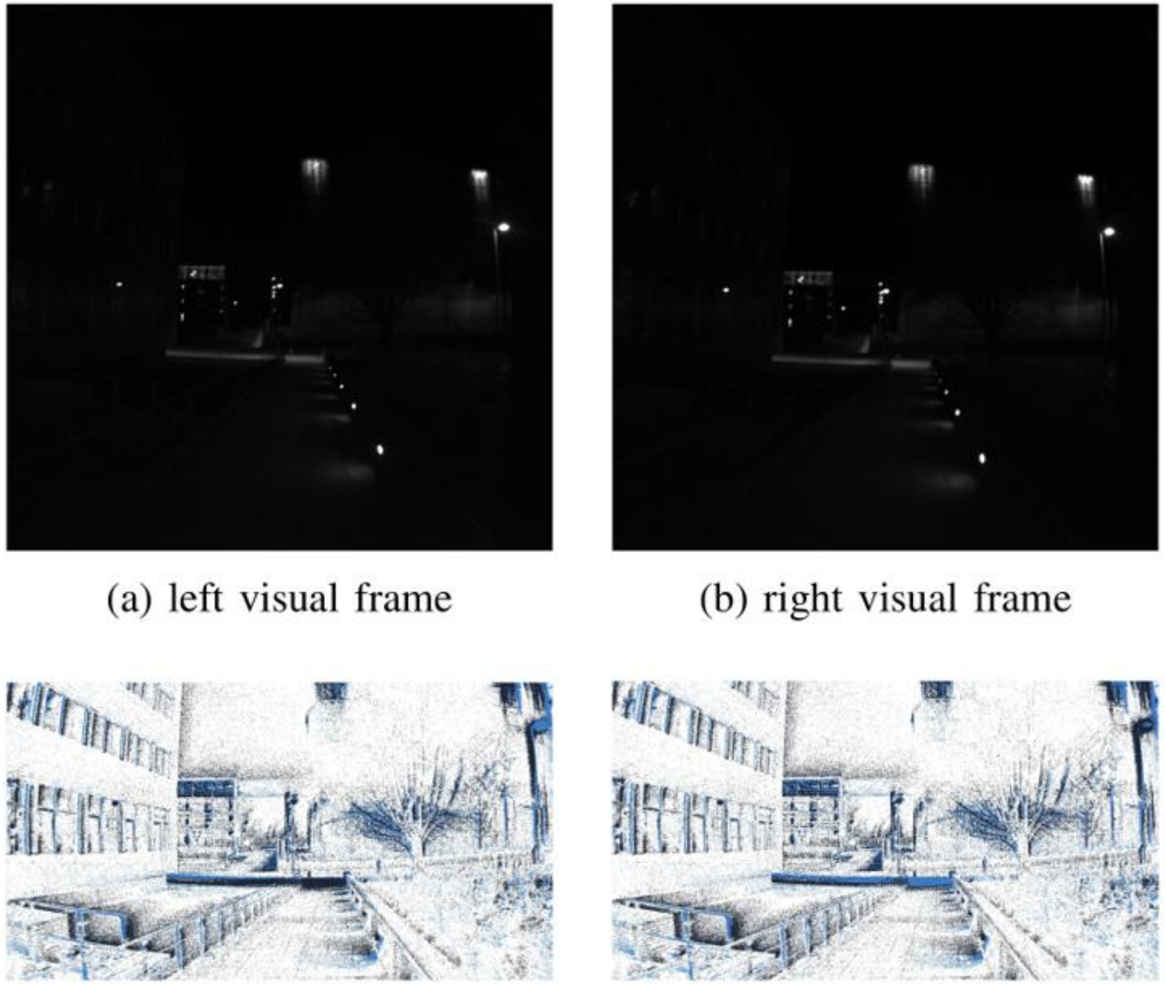


Klenk, Motzet, Koestler, Cremers, "Deep Event Visual Odometry", 3DV 2024

DEVO: Qualitative Evaluation

Klenk, Motzet, Koestler, Cremers, “Deep Event Visual Odometry”, 3DV 2024





(a) left visual frame (b) right visual frame

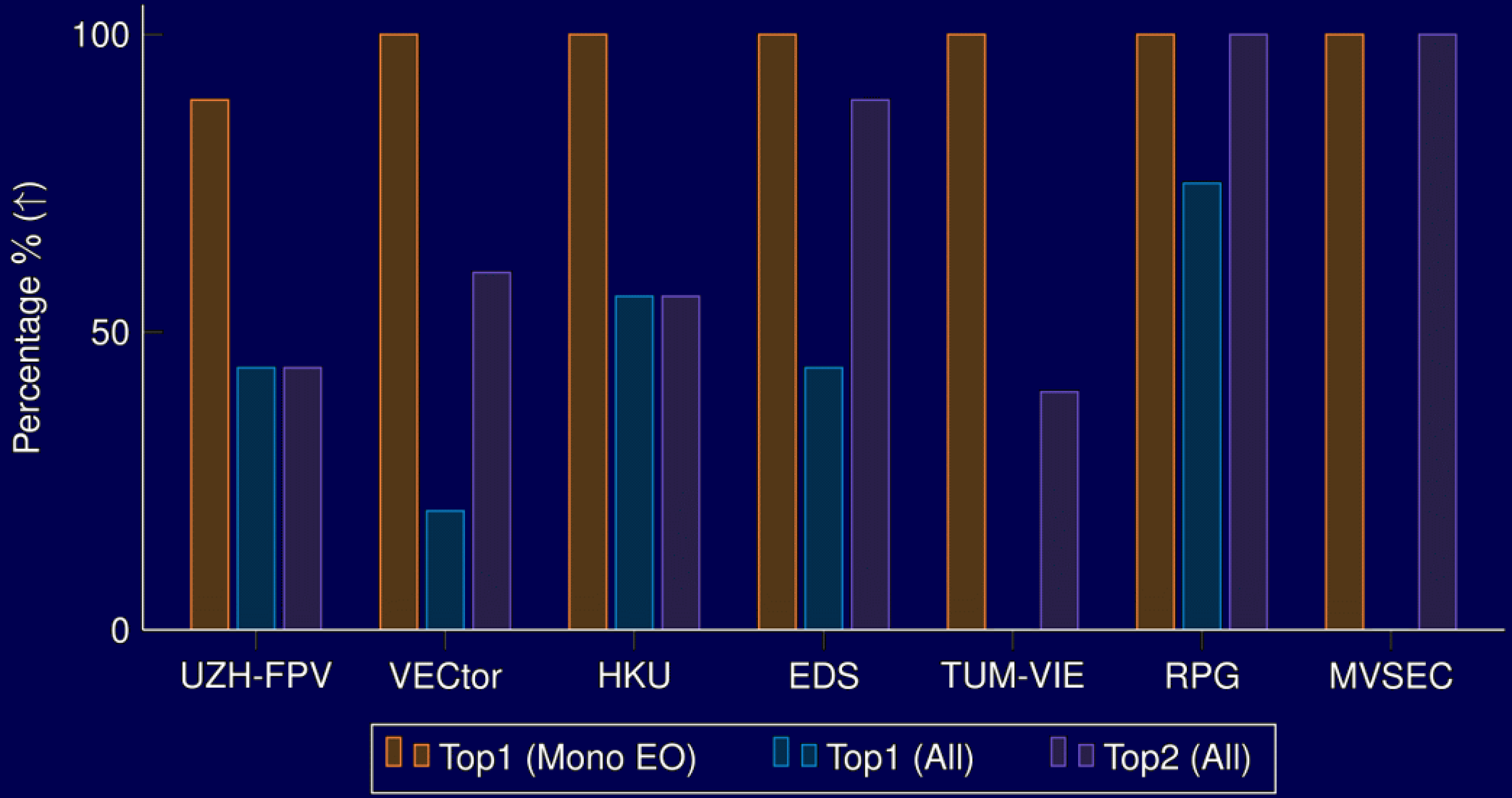
(c) left event frame (d) right event frame

Klenk, Chui, Demmel, Cremers, “TUM-VIE: The TUM stereo visual-inertial event dataset”, IROS 2021

The TUM VIE Dataset

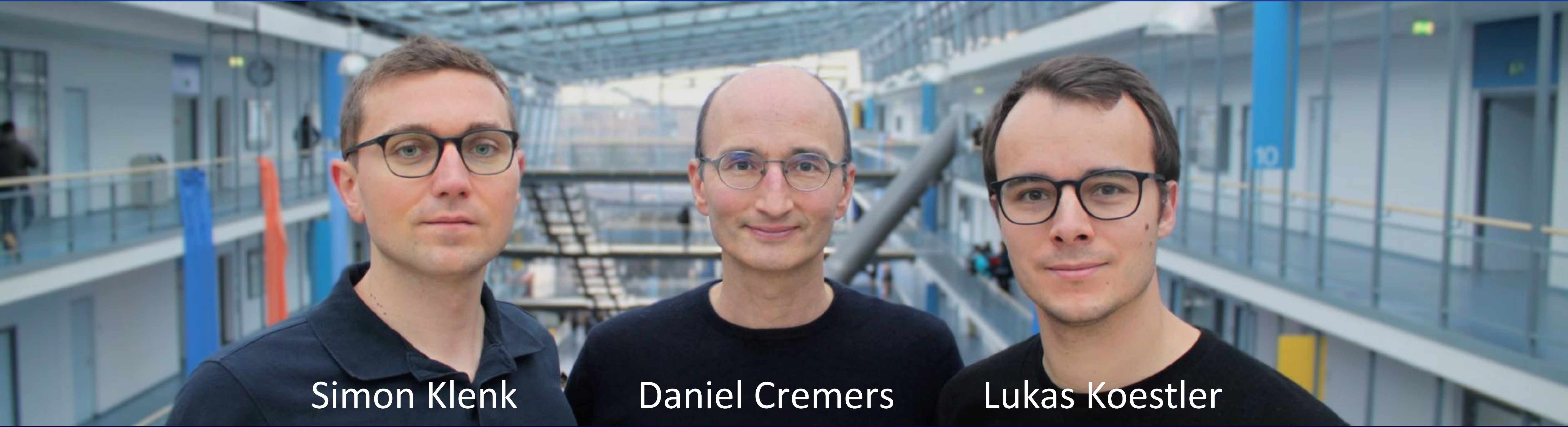
Klenk, Chui, Demmel, Cremers, "TUM-VIE: The TUM stereo visual-inertial event dataset", IROS 2021.

Performance across 7 Datasets



Klenk, Motzet, Koestler, Cremers, "Deep Event Visual Odometry", 3DV 2024

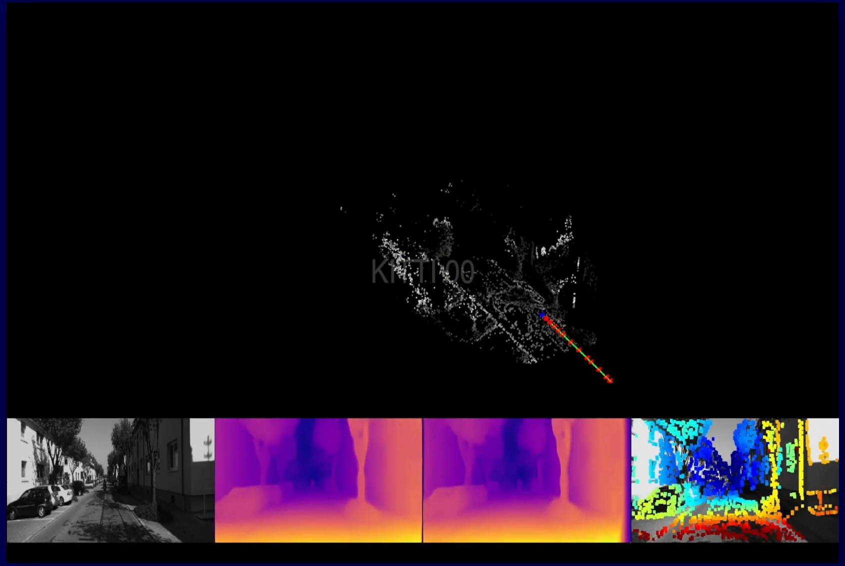
SE3 Labs: Startup on Spatial AI



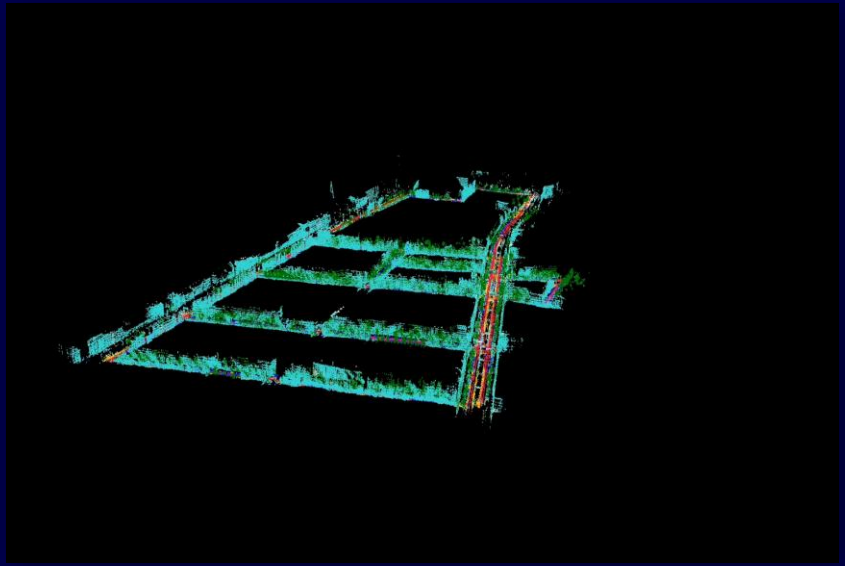
Interested in joining us?

Summary

DM-VIO for visual-inertial odometry

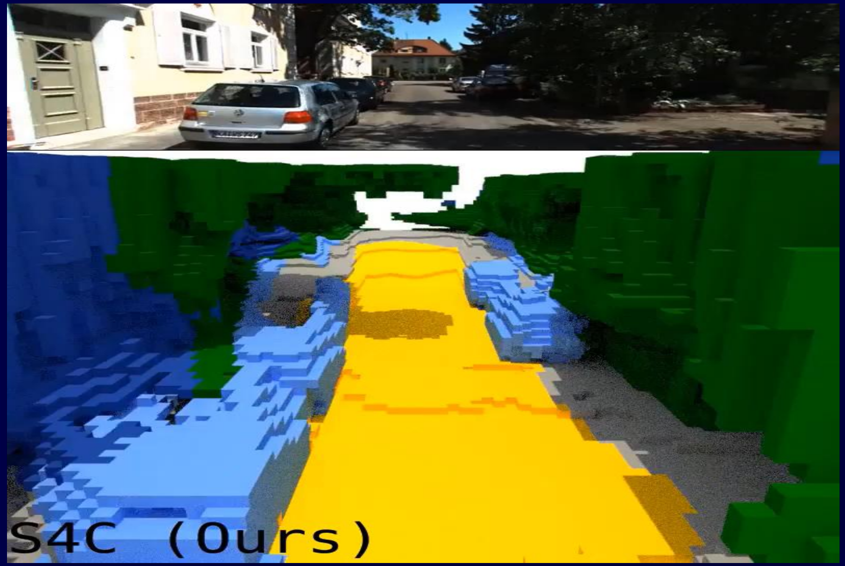


Deep & direct visual SLAM



3D semantic mapping

Monoocular dense mapping



Semantic scene completion

4D reconstruction from sparse data