SAMP: Shape and Motion Priors for 4D Vehicle Reconstruction

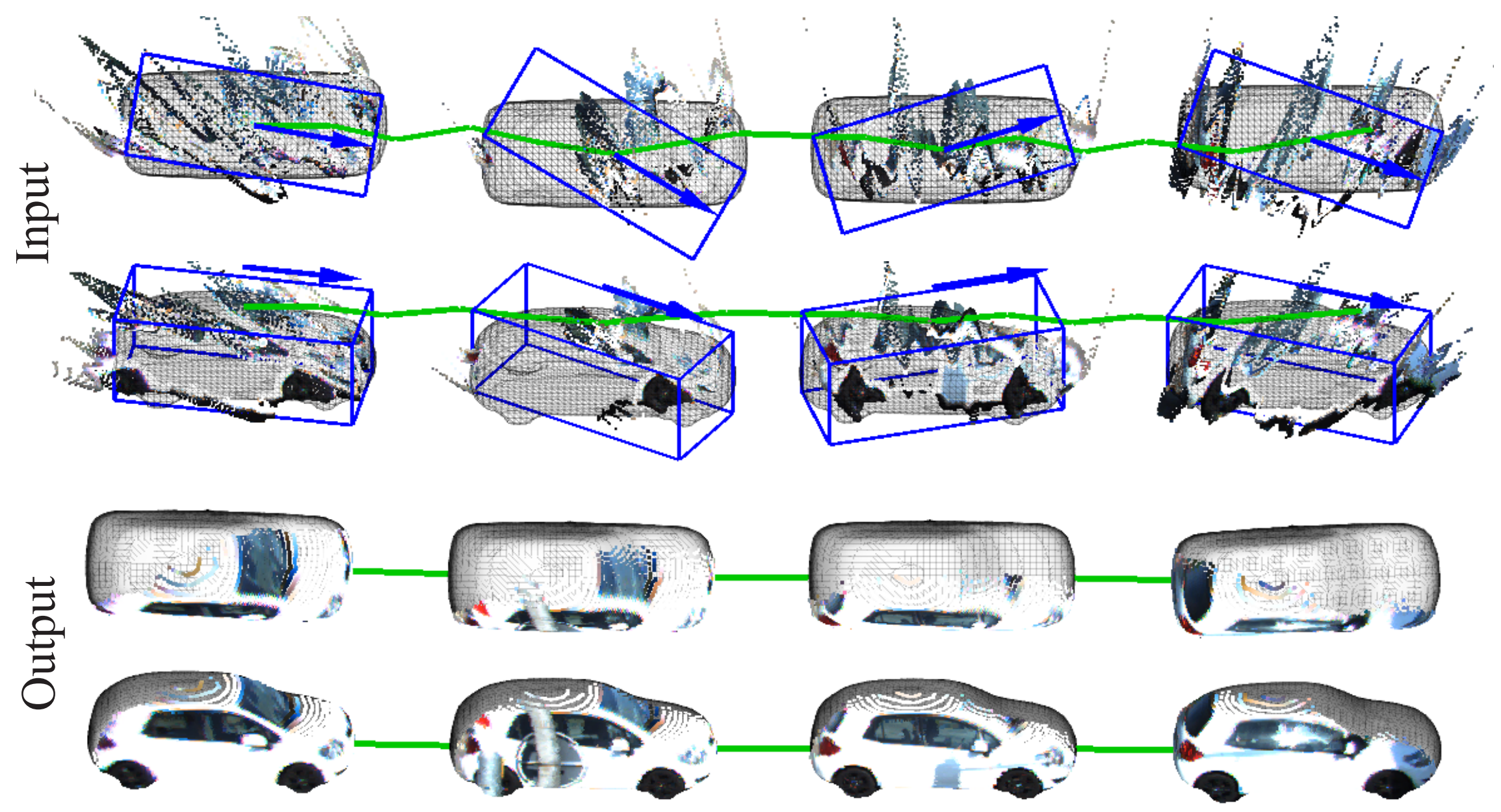
Francis Engelmann, Jörg Stückler, Bastian Leibe

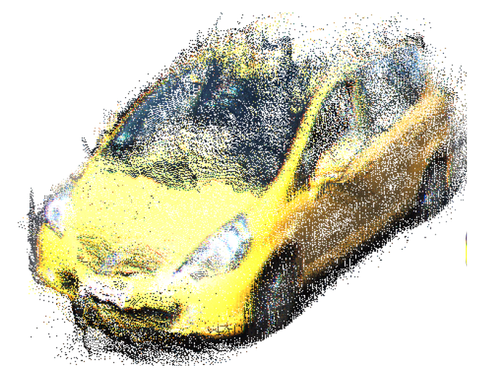
Computer Vision Group, Visual Computing Institute, RWTH Aachen University

Abstract

Inferring the pose and shape of vehicles in 3D from a movable platform still remains a challenging task due to the projective sensing principle of cameras, difficult surface properties such as reflections or transparency, and illumination changes between images. In this work, we propose to use 3D shape and motion priors to regularize the estimation of the trajectory and the shape of vehicles in sequences of stereo images. We represent shapes by 3D signed distance functions and embed them in a low-dimensional manifold. Our optimization method allows for imposing a common shape across all image observations along an object track. We employ a motion model to regularize the trajectory to plausible object motions.

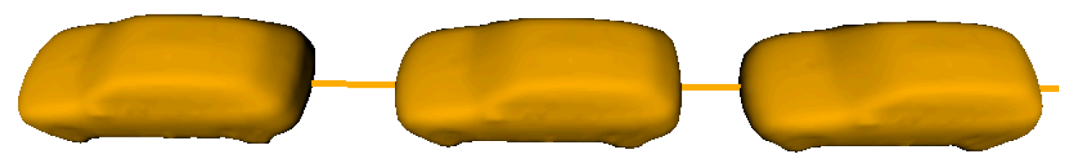
Problem Statement



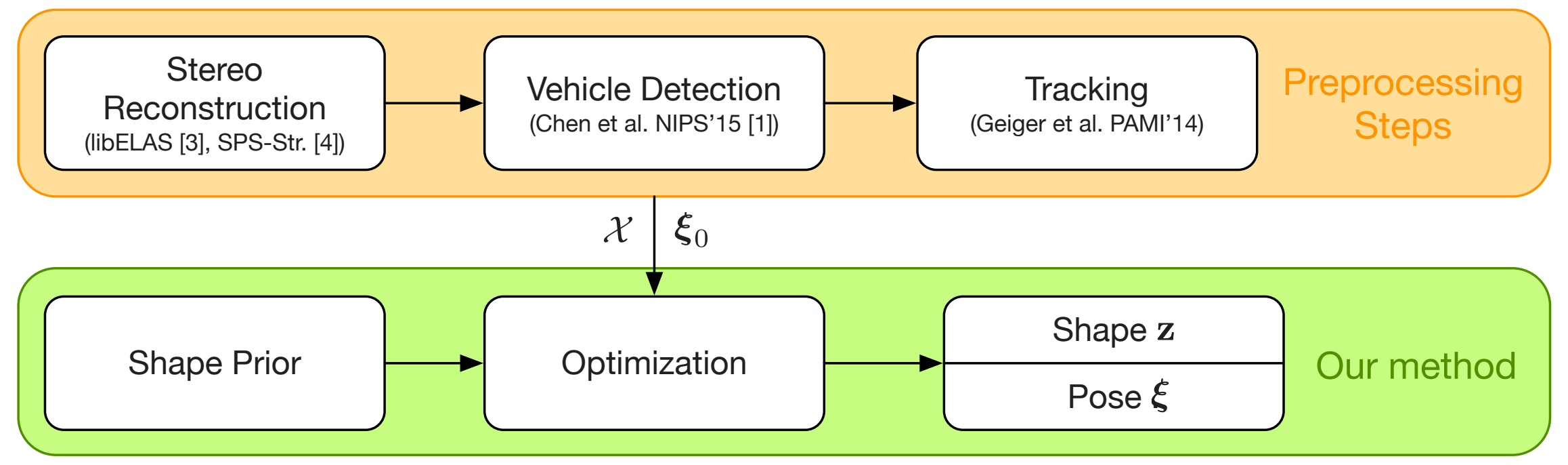


Input. Required prepocessing steps:
- Depth estimation from stereo pairs.
- Detections and associations (tracking).
- Egomotion of observing camera.

Output. Our method then estimates:
- Complete shape for tracked vehicle.
- Precise pose in 3D of each detection.
- Improved tracking results.



Our Model



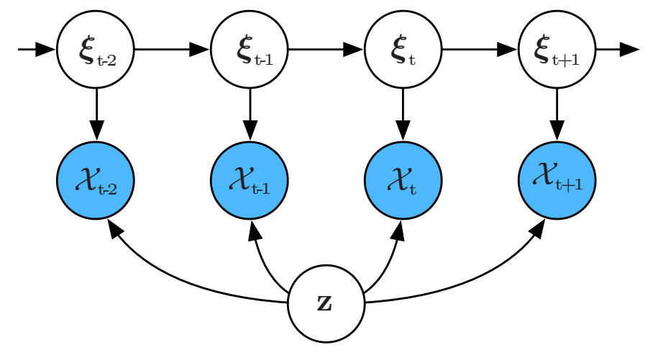
Stereo Reconstruction (libELAS [3], SPS-Str. [4]) → Vehicle Detection (Chen et al. NIPS'15 [1]) → Tracking (Geiger et al. PAMI'14) — Preprocessing Steps

$\mathcal{X}$ | ξ_0

Shape Prior → Optimization → Shape $\mathbf{z}$ / Pose ξ — Our method

Probabilistic Interpretation

- Given: a set of detections associated over time into tracks.
- Find the common shape $\mathbf{z}$ of the tracked vehicle and its poses $\boldsymbol{\xi} = \{\boldsymbol{\xi}_t\}$ given the 3D depth observations $\mathcal{X}_t$ of a detection at time t:

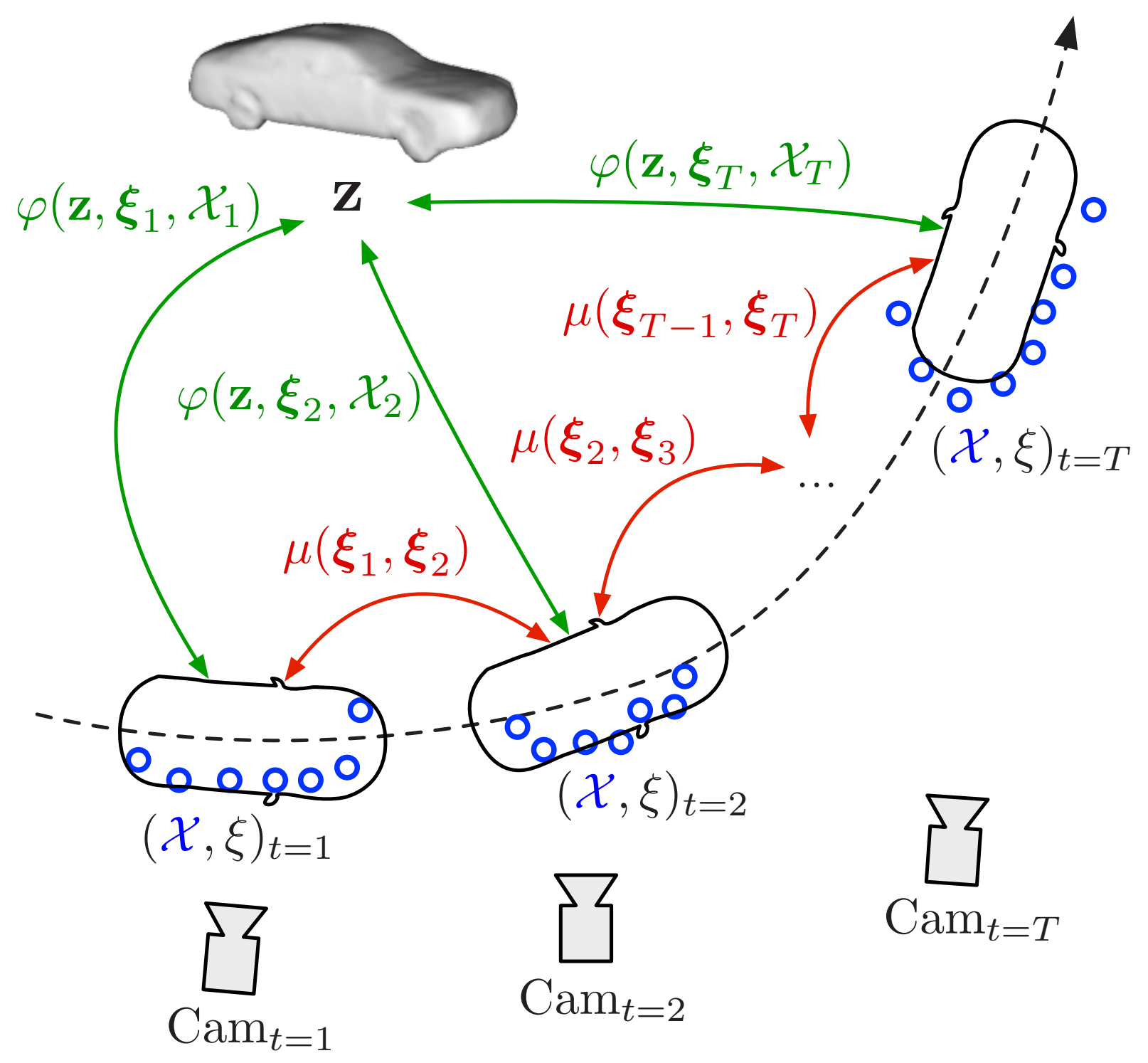


$$p(\mathbf{z}, \boldsymbol{\xi} \mid \mathcal{X}) = \eta\, p(\mathcal{X} \mid \mathbf{z}, \boldsymbol{\xi})\, p(\mathbf{z})\, p(\boldsymbol{\xi})$$

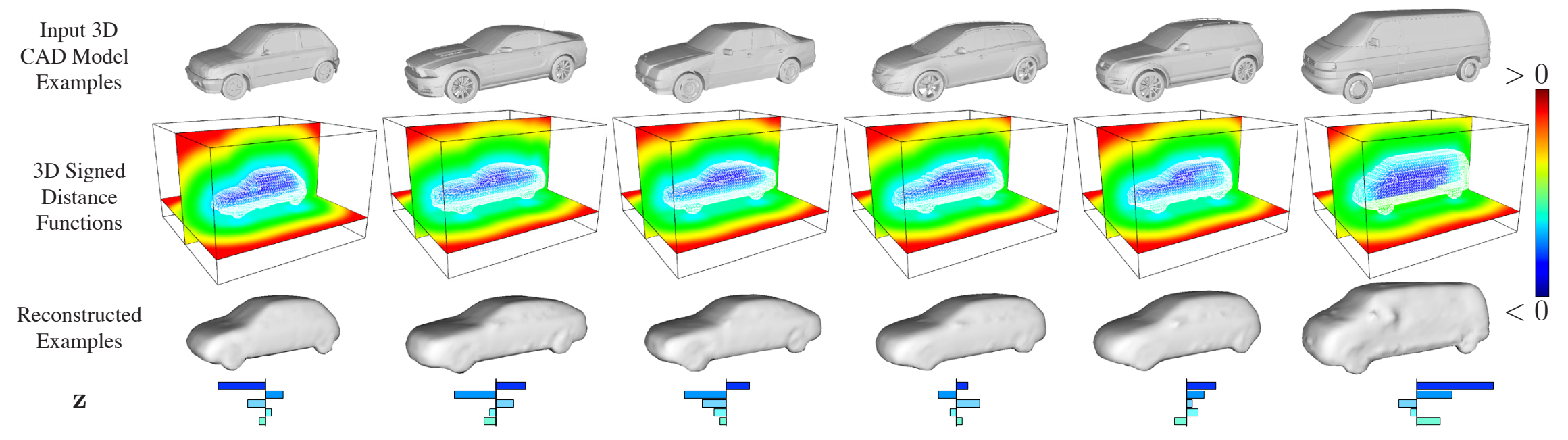
$$\{\mathbf{z}, \boldsymbol{\xi}\}^* = \underset{\mathbf{z}, \boldsymbol{\xi}}{\operatorname{argmax}}\, p(\mathbf{z}, \boldsymbol{\xi} \mid \mathcal{X}).$$

We maximimze the posterior $p(\mathbf{z}, \xi | \mathcal{X})$ by minimizing the energy function obtained by applying the negative logarithm:

$$\mathrm{E}(\mathbf{z}, \boldsymbol{\xi}) = \frac{1}{T} \sum_t [\underbrace{\varphi(\mathbf{z}, \boldsymbol{\xi}_t, \mathcal{X}_t)}_{\text{data term}} + \underbrace{\mu(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1})}_{\text{motion term}}] + \underbrace{\kappa(\mathbf{z})}_{\text{shape regul.}}$$


$\varphi(\mathbf{z}, \boldsymbol{\xi}_1, \mathcal{X}_1)$
$\mathbf{z}$
$\varphi(\mathbf{z}, \boldsymbol{\xi}_T, \mathcal{X}_T)$
$\mu(\boldsymbol{\xi}_{T-1}, \boldsymbol{\xi}_T)$
$\varphi(\mathbf{z}, \boldsymbol{\xi}_2, \mathcal{X}_2)$
$\mu(\boldsymbol{\xi}_2, \boldsymbol{\xi}_3)$
...
$(\mathcal{X}, \xi)_{t=T}$
$\mu(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$
$(\mathcal{X}, \xi)_{t=1}$
$(\mathcal{X}, \xi)_{t=2}$
$\mathrm{Cam}_{t=1}$
$\mathrm{Cam}_{t=2}$
$\mathrm{Cam}_{t=T}$

Shape Prior



- Shape prior learned form collection of 3D CAD models.
- Convert each model into its discrete signed distance function (SDF) representation. We samples points from the surface of a vehicle and for each voxel we store the distance to the closest point.
- Linear low-dimensional embedding obtained with PCA.

Data Term

The data term modifies the shape and the pose of the vehicle to correspond to the observed depth:

$$\varphi(\mathbf{z}, \boldsymbol{\xi}, \mathcal{X}) = \frac{1}{N} \sum_{\mathbf{x} \in \mathcal{X}} \rho \left(\frac{\phi_{\mathbf{z}}(\mathbf{T}_{\boldsymbol{\xi}}\, \mathbf{x})}{\sigma_{d_{\mathbf{x}}}} \right)$$

- $\phi_{\mathbf{z}}$: Signed distance function parametrized by $\mathbf{z}$.
- $\mathbf{T}_{\xi}$: Transformation matrix induced by pose $\boldsymbol{\xi}$.
- $\sigma_{d_{\mathbf{x}}}$: Depth uncertainty.
- ρ : Huber loss.

Motion Term

The motion term guarantees consistent poses along the trajectory of an estimated motion model:

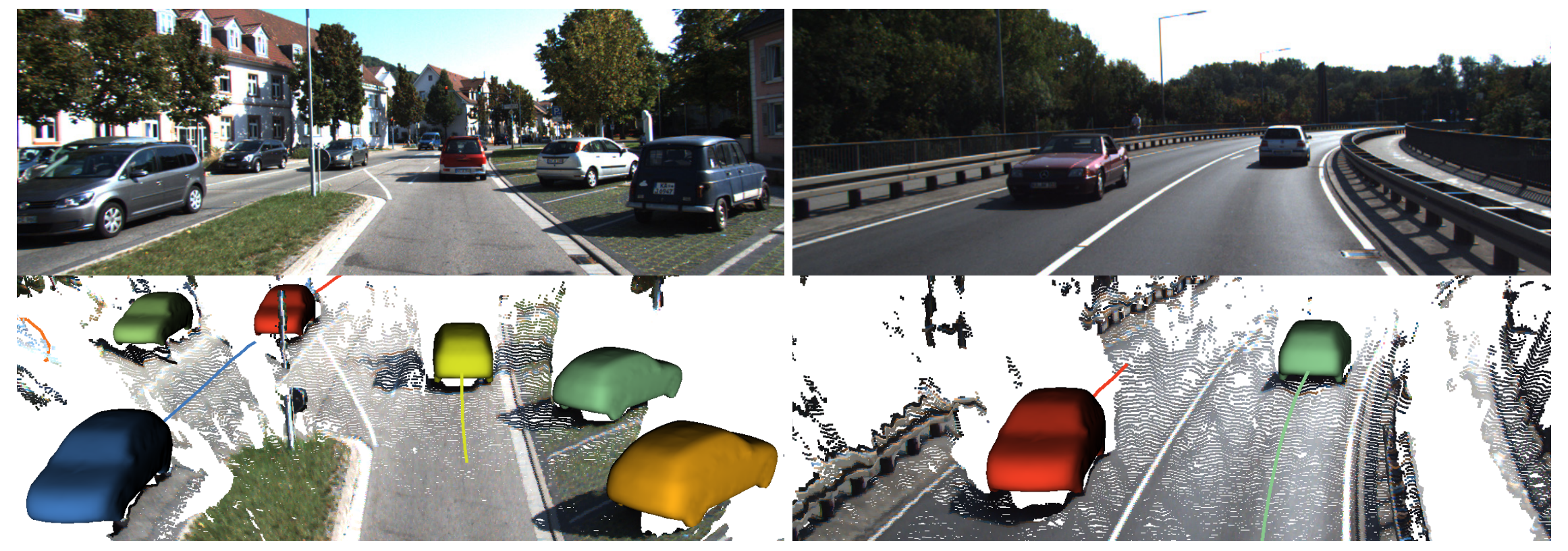
$$\mu(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1}) = ||\boldsymbol{\xi}_t - g(\boldsymbol{\xi}_{t-1})||^2_{\boldsymbol{\Sigma}} + ||\underbrace{t_y - t_{y_{gp}}}_{\text{ground plane prior}}||^2_{\boldsymbol{\Sigma}_{gp}}$$

- g : Motion model depending on vehicle dynamic behavior.

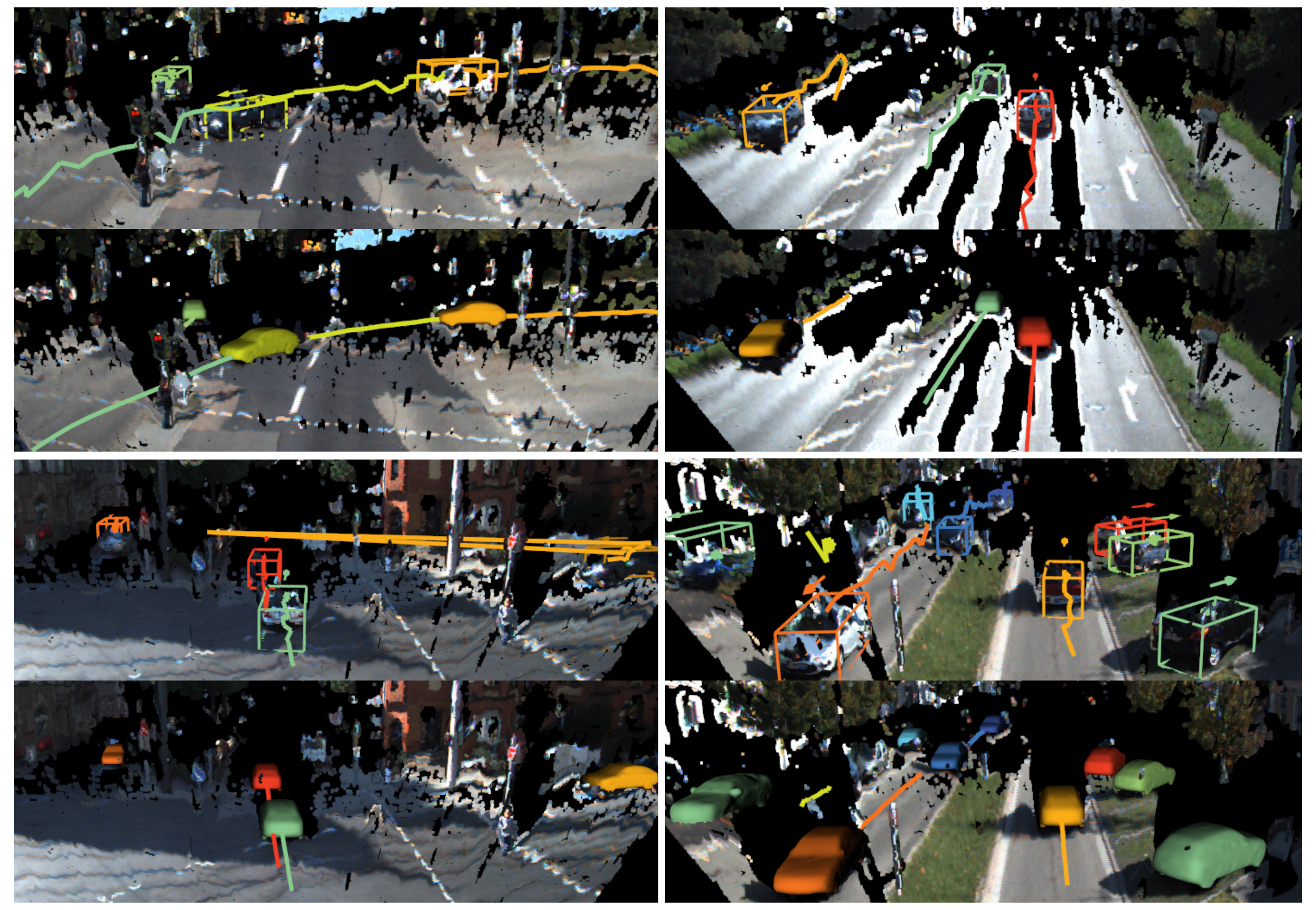
Our approach incorporates three separate motion models:
1. Static: the vehicle is not moving.
2. Line: the vehicle drives forward on a straight line.
3. Turn: the vehicle is currently taking a turn.

The motion model is chosen based on a vehicle's initial track.

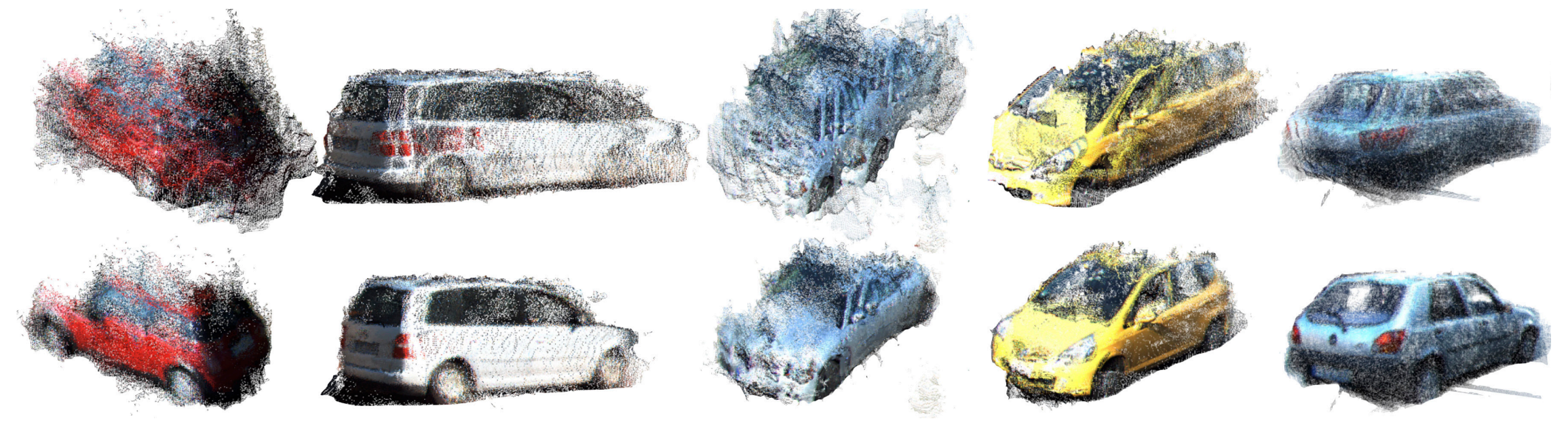
Qualitative Results



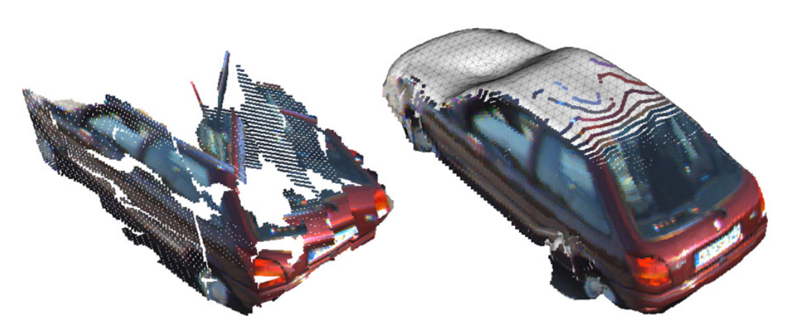
KITTI Stereo 2015. Top: Input stereo image. Bottom: 3D shape and trajectory.



Qualitative Results. Top row: Input consisting of stereo reconstruction and tracked detections. Bottom row: Full shape reconstruction and improved pose estimations.



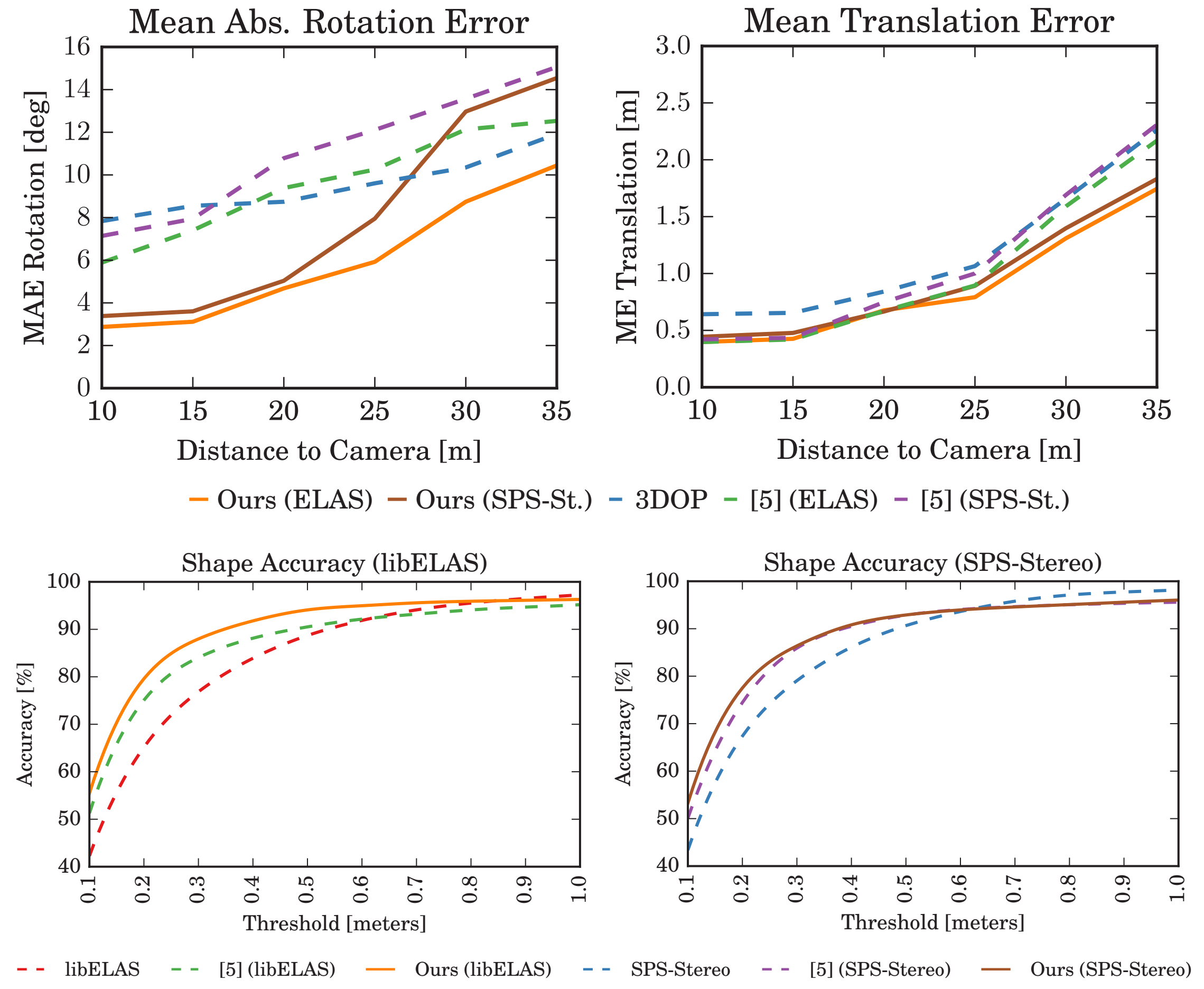
Pose Optimization Results. Top row: All initial poses of a track overlayed Bottom row: all optimized poses of the same track.



Shape Completion. Reconstrucion of previously unobserved surfaces. Left: Initial stereo reconstruction from [4]. Right: Our full shape reconstruction shown as wireframe.

Quantitative Results

Dataset: We evaluate our method on the KITTI Stereo 2015 dataset [3] in terms of shape reconstruction and pose estimation accuracy. The dataset includes dense depth annotations and ground truth pose annotations.



Quantitative results. Top: Mean absolute rotation and translation error. Bottom: Reconstruced shape accuracy comparing different input depths.

References

[1] X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, and R. U r - tasun. 3D object proposals for accurate object class detection. In Proc. of Neural Information Processing Systems (NIPS), 2015.

[2] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proc. of the IEEE Int. Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[3] A. Geiger, M. Roser, and R. Urtasun. Efficient large-scale stereo matching. In Proc. of the Asian Conference on Computer Vision (ACCV), 2010.

[4] K. Yamaguchi, D. McAllester, and R. Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In Proc. of the European Conference on Computer Vision (ECCV), 2014.

[5] F. Engelmann, J. Stückler, and B. Leibe. Joint object pose estimation and shape reconstruction in urban street scenes using 3D shape priors. In Proc. of the German Conference on Pattern Recognition (GCPR), 2016.

M.Sc. Francis Engelmann
Phone: +49 241 80 20760
E-Mail: engelmann@vision.rwth-aachen.de
Website: http://www.vision.rwth-aachen.de/publication/00146/

This project has received funding from the European Research Council (ERC) CV-SUPER (ERC-2012-StG-307432).

Computer Vision Group, Visual Computing Institute
RWTH Aachen University
Mies-van-der-Rohe Str. 15, D-52074 Aachen, Germany
http://www.vision.rwth-aachen.de

Visual Computing Institute

RWTH AACHEN UNIVERSITY