Semantic 3D Reconstruction and localization

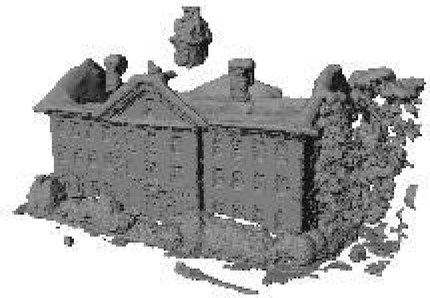
Marc Pollefeys

ETH Zürich & Microsoft

joint work with Christian Haene, Nikolay Savinov, Ian Cherabier, Lubor Ladicky, Martin Oswald, Christopher Zach, Andrea Cohen, Johannes Schoenberger. Andreas Geiger

Classical 3D reconstruction from images

→



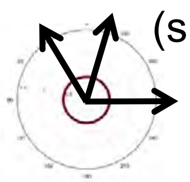
- Classical 3D from image approach
  - Relative pose between images (structure-from-motion)
  - Per pixel depth estimation (multi-view stereo matching)
  - Surface reconstruction (TSDF, poisson, graph energy minimization)

$$E(x) = \sum_{S \in \Omega} \rho_S x_S + \phi\left(\nabla x_S\right)$$

unary depth evidence term ρ_S

isotropic shape prior ϕ

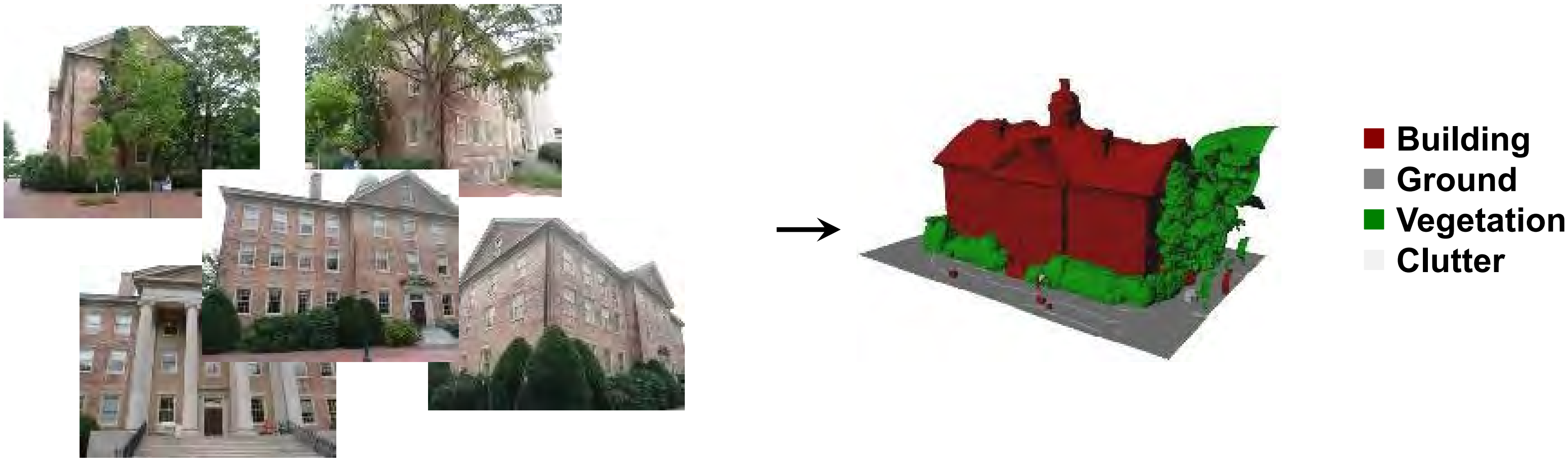
(spherical)



ETH Zürich

Computer Vision and Geometry Lab

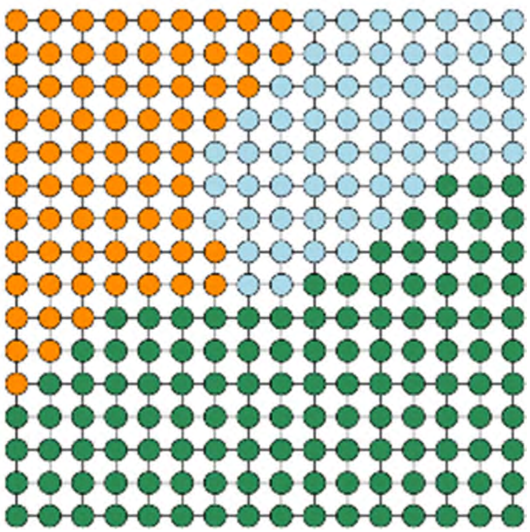
Semantic 3D reconstruction



- Joint 3D reconstruction and class segmentation
  - Obtain separate surface for each class of object
  - Corresponds to **multi-label volumetric segmentation problem**

Discrete and Continuous Formulations

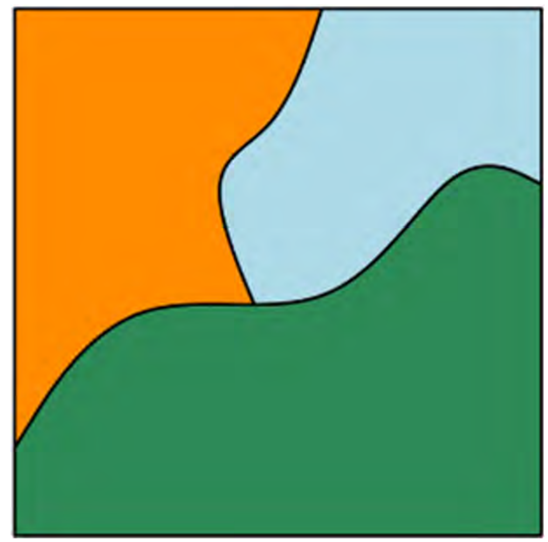
Discrete Domain



Smoothness:
transitions along edges

Formulation: Linear Program
[Schlesinger 1976]
Arbitrary smoothness cost allowed

Continuous Domain



Smoothness:
(anisotropic) boundary length

Formulation: Convex Program
[Chambolle, Cremers, Pock 2008]
Smoothness needs to form a metric

Convex, Continuous Multi-Label Formulation
[Zach, Häne, Pollefeys, CVPR 2012, TPAMI 2014]

- Metric smoothness fulfills triangle inequality
  - **Truncated quadratic** smoothness **non-metric**
- Our continuously inspired formulation
  - Takes best from both worlds
  - **Non-metric** and **anisotropic** boundary length cost

$$E(x) = \sum_{s \in \Omega} \left(\sum_i \rho_s^i x_s^i + \sum_{i,j:i<j} \phi_s^{ij}(x_s^{ij} - x_s^{ji}) \right)$$

$$\text{subject to } \; x_s^i = \sum_j (x_s^{ij})_k, \quad x_s^i = \sum_j (x_{s-e_k}^{ji})_k$$

$$x_s^i \geq 0, \quad \sum_i x_s^i = 1, \quad x_s^{ij} \geq 0$$

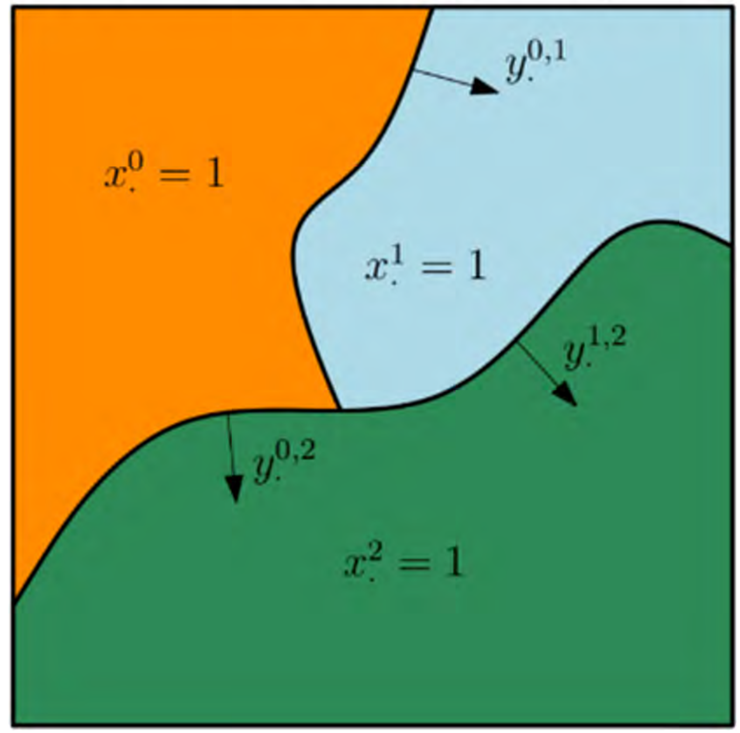
Energy Formulation

$$E(x) = \sum_{s\in\Omega} \left(\sum_i \rho_s^i x_s^i + \sum_{i,j:i<j} \phi_s^{ij}(x_s^{ij} - x_s^{ji}) \right)$$

$$\text{subject to} \quad x_s^i = \sum_j (x_s^{ij})_k, \quad x_s^i = \sum_j (x_{s-e_k}^{ji})_k$$

$$x_s^i \geq 0, \quad \sum_i x_s^i = 1, \quad x_s^{ij} \geq 0$$

Label transition gradients $y_s^{ij} := x_s^{ij} - x_s^{ji}$



Cost for boundary:

$$\phi_s^{ij}(\cdot) : \mathbb{R}^N \to \mathbb{R}_0^+$$

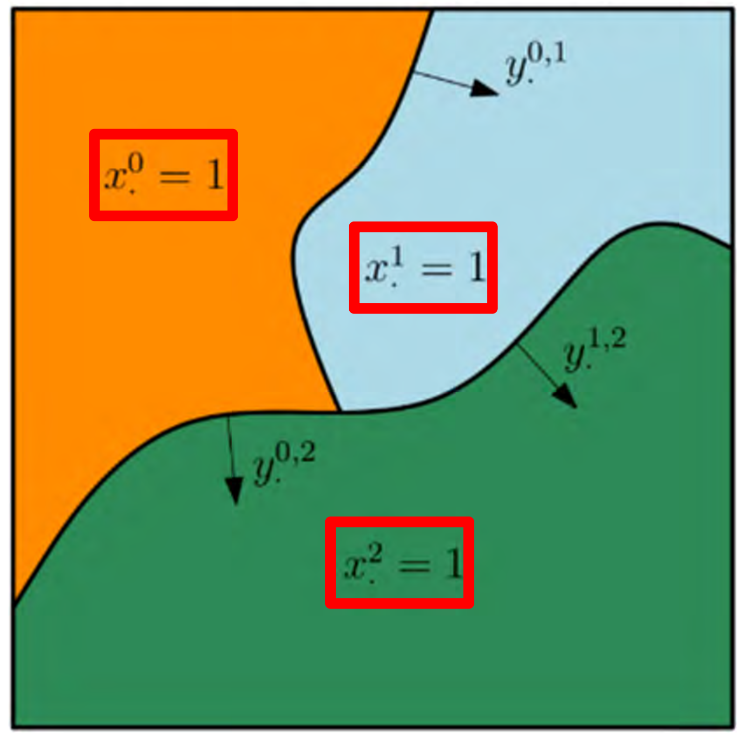
Energy Formulation

$$E(x) = \sum_{s\in\Omega}\left(\sum_i \rho_s^i x_s^i + \sum_{i,j:i<j} \phi_s^{ij}(x_s^{ij} - x_s^{ji})\right)$$

$$\text{subject to}\quad x_s^i = \sum_j (x_s^{ij})_k, \quad x_s^i = \sum_j (x_{s-e_k}^{ji})_k$$

$$x_s^i \geq 0, \quad \sum_i x_s^i = 1, \quad x_s^{ij} \geq 0$$

Label transition gradients $y_s^{ij} := x_s^{ij} - x_s^{ji}$



Cost for boundary:

$$\phi_s^{ij}(\cdot) : \mathbb{R}^N \rightarrow \mathbb{R}_0^+$$

Energy Formulation

$$E(x) = \sum_{s\in\Omega}\left(\sum_i \rho_s^i x_s^i + \sum_{i,j:i<j} \phi_s^{ij}(x_s^{ij} - x_s^{ji})\right)$$

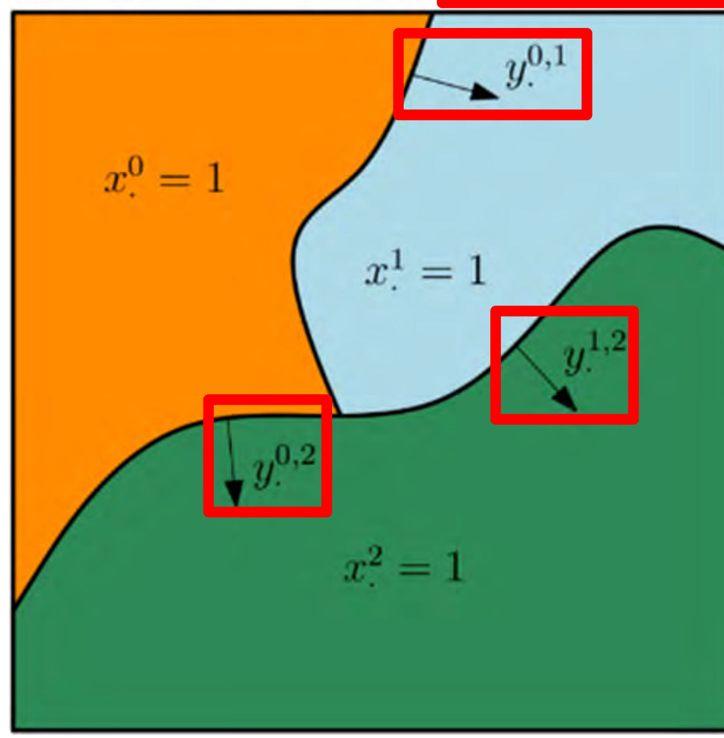
subject to $x_s^i = \sum_j (x_s^{ij})_k, \quad x_s^i = \sum_j (x_{s-e_k}^{ji})_k$

$$x_s^i \geq 0, \quad \sum_i x_s^i = 1, \quad x_s^{ij} \geq 0$$

Cost for boundary:

$$\phi_s^{ij}(\cdot) : \mathbb{R}^N \to \mathbb{R}_0^+$$

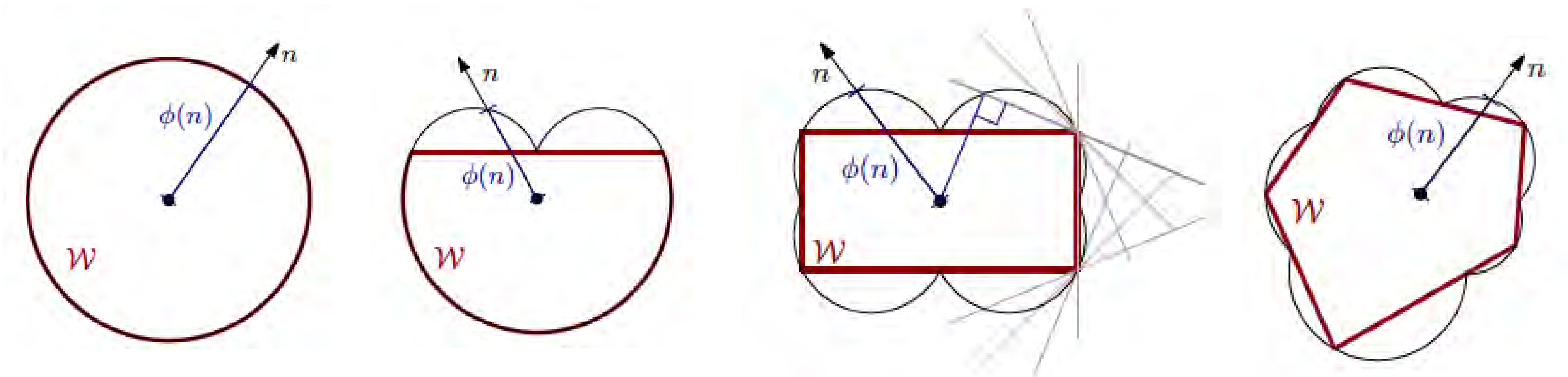
Label transition gradients $y_s^{ij} := x_s^{ij} - x_s^{ji}$



Wulff shapes

The shape of an equilibrium crystal is obtained, according to the Gibbs thermodynamic principle, by minimizing the total surface free energy associated to the crystal-medium interface. http://www.scholarpedia.org/

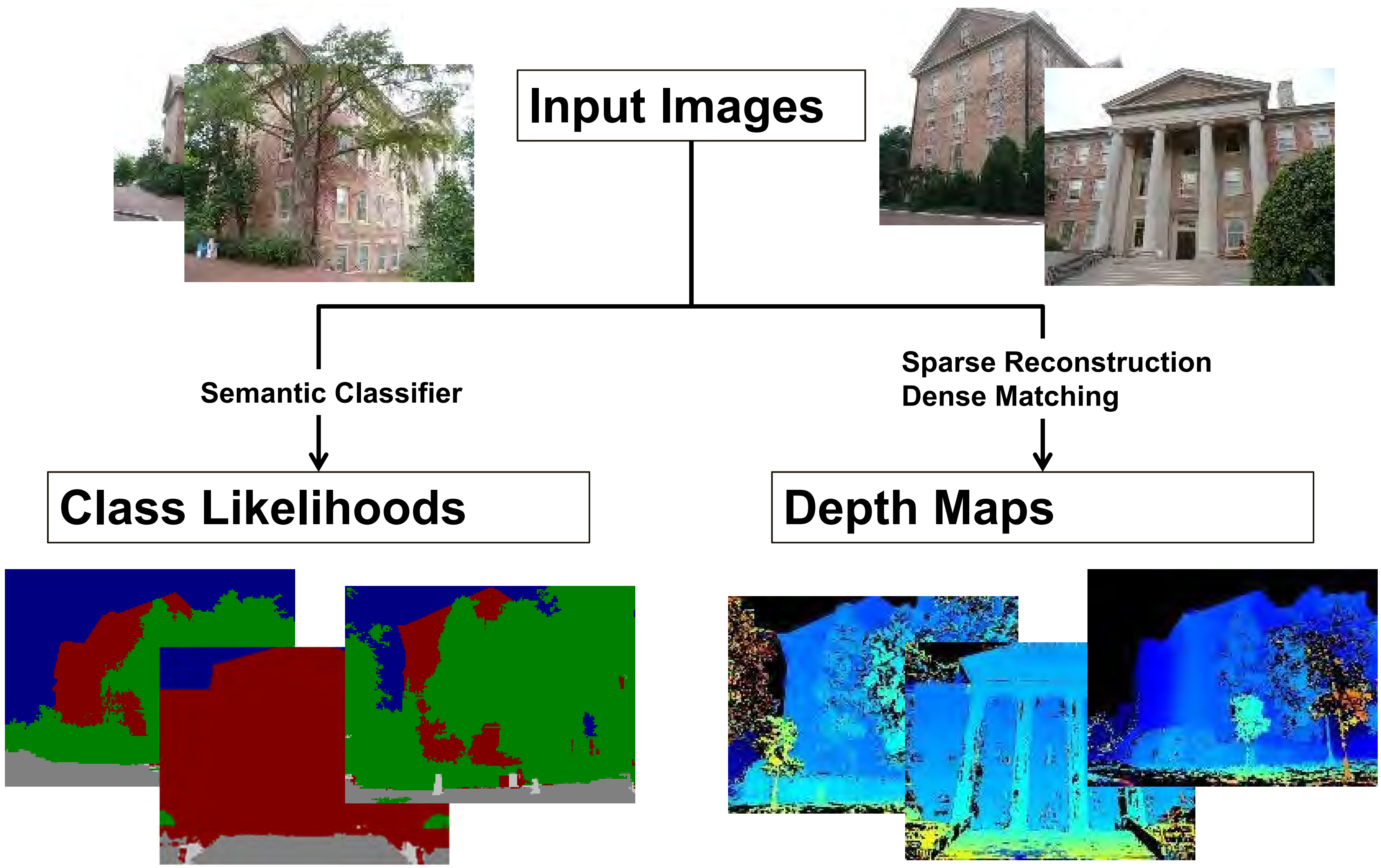
Wulff's theorem: The minimum surface energy for a given volume of a polyhedron will be achieved if the distances of its faces from one given point are proportional to their surface tension



Proposed use for anisotropic regularization
[Esedoglu and Oscher 2004, Zach et al. 2009, Haene et al. 2013/14/15]

Dense Semantic 3D Reconstruction

[Häne, Zach, Cohen, Angst, Pollefeys, CVPR 2013]

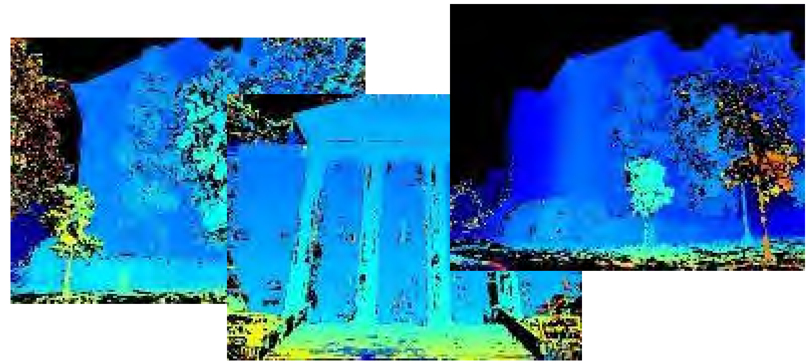


ETH Zürich

Computer Vision and Geometry Lab

Dense Semantic 3D Reconstruction

[Häne, Zach, Cohen, Angst, Pollefeys, CVPR 2013]



Class Likelihoods

Depth Maps

Joint Fusion, Convex Optimization

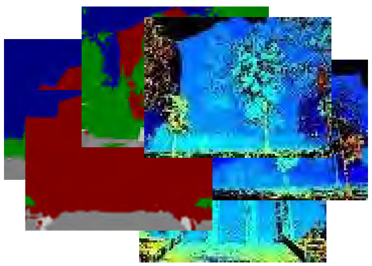
Dense Semantic 3D Model

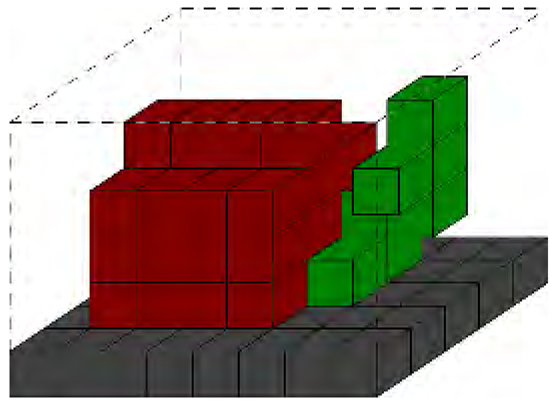
ETH Zürich

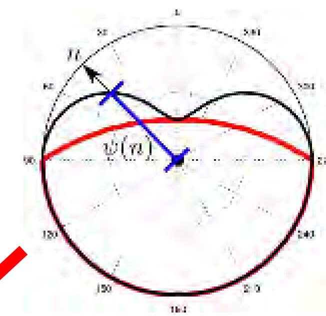
Computer Vision and Geometry Lab

Formulation: Labeling of a Voxel Space

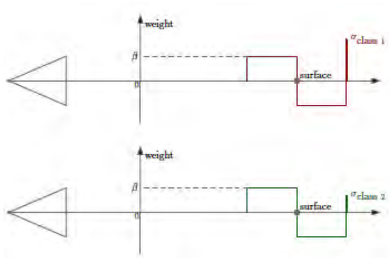
[Häne, Zach, Cohen, Angst, Pollefeys, CVPR 2013]







Data Term: Described as per-voxel unary potentials



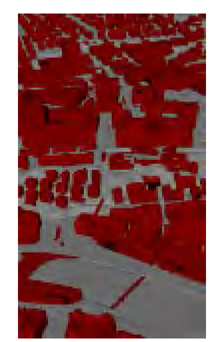
$$E(x) = \sum_{s \in \Omega} \left(\sum_i \rho_s^i x_s^i + \sum_{i,j:i<j} \phi^{ij}(x_s^{ij} - x_s^{ji}) \right)$$

$$\text{subject to} \quad x_s^i = \sum_j (x_s^{ij})_k, \quad x_s^i = \sum_j (x_{s-e_k}^{ji})_k$$

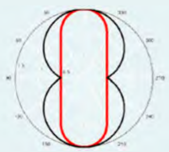
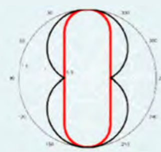
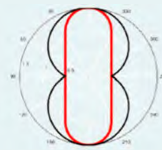
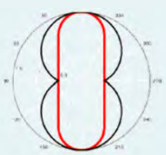
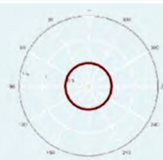
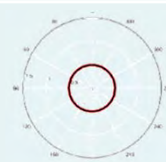
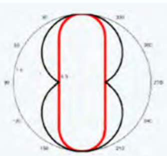
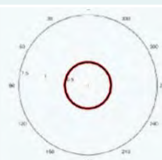
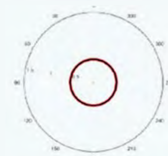
$$x_s^i \geq 0, \quad \sum_i x_s^i = 1, \quad x_s^{ij} \geq 0$$

Regularization Term: Class-specific, direction dependent, surface area penalization

Learned from training data

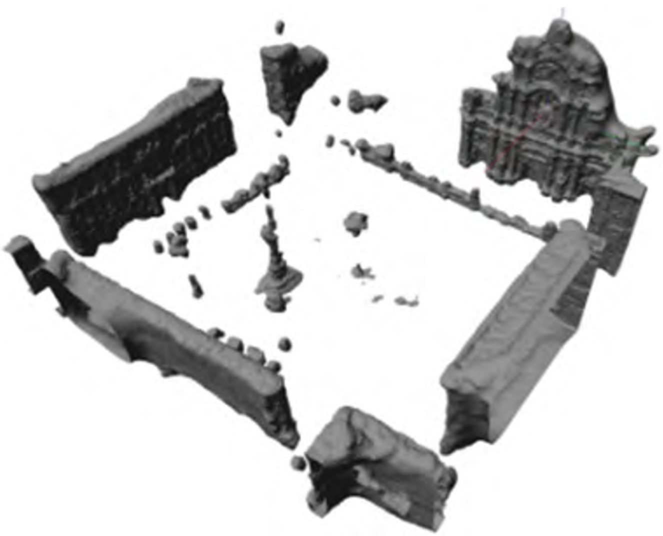


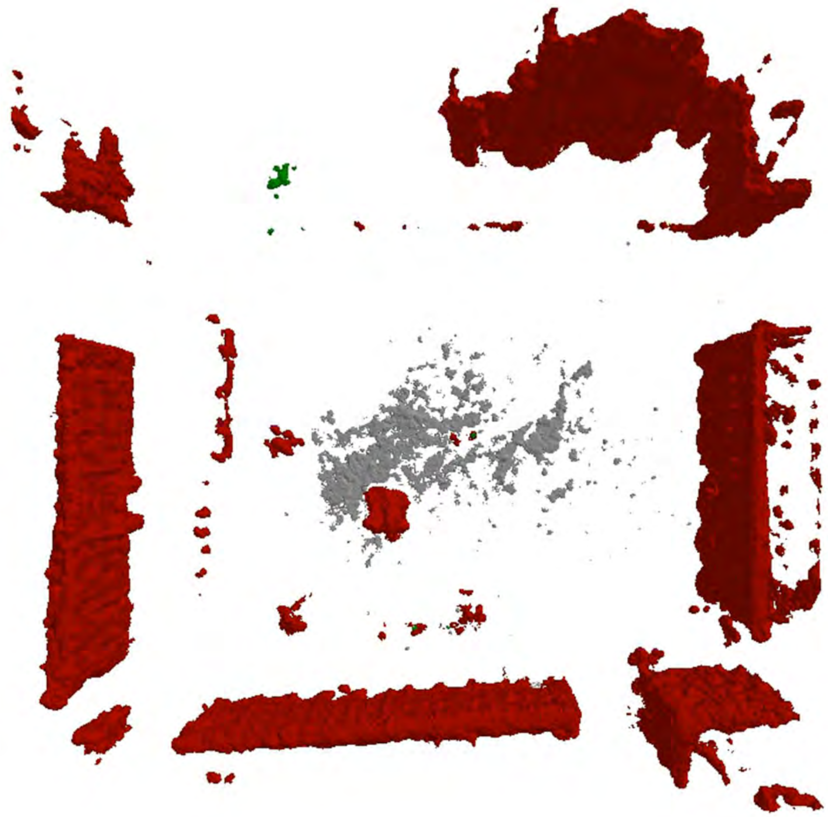
Class specific training/selection of $\phi_s^{ij}(\cdot)$

	building	ground	vegetation	stuff	air
building					
ground					
vegetation					
stuff					

Joint 3D reconstruction and class segmentation

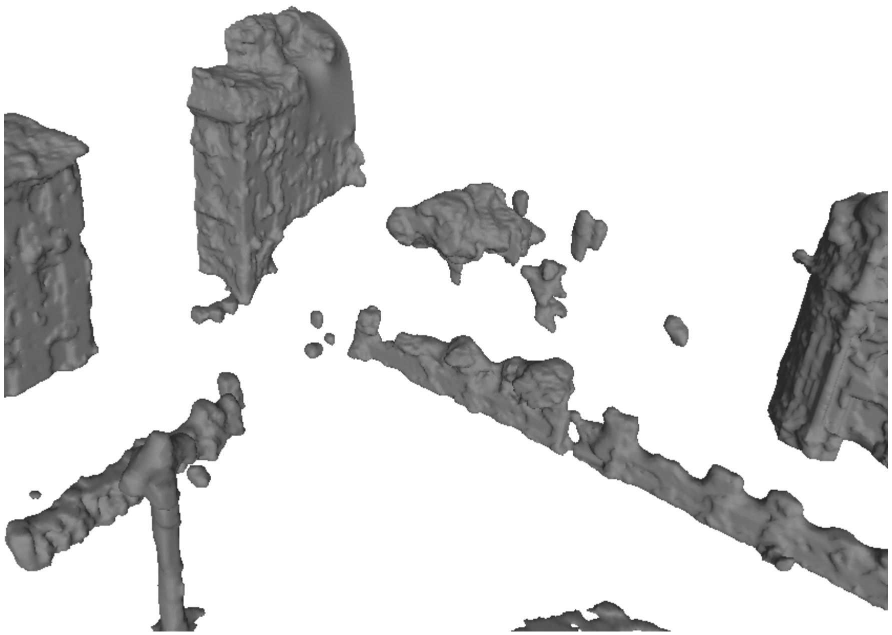
(Haene et al CVPR13)

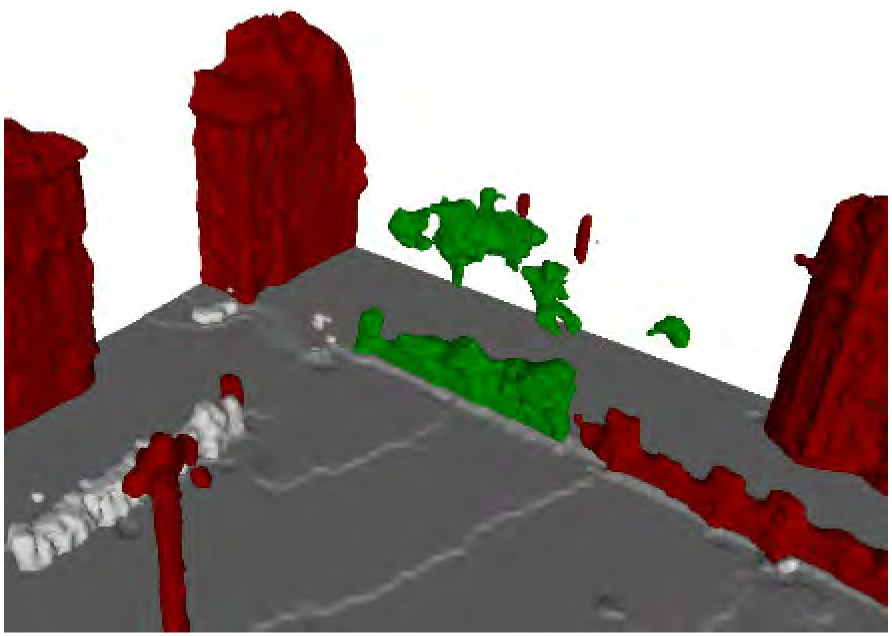




Weakly Observed Structures I

- Buildings standing on the ground





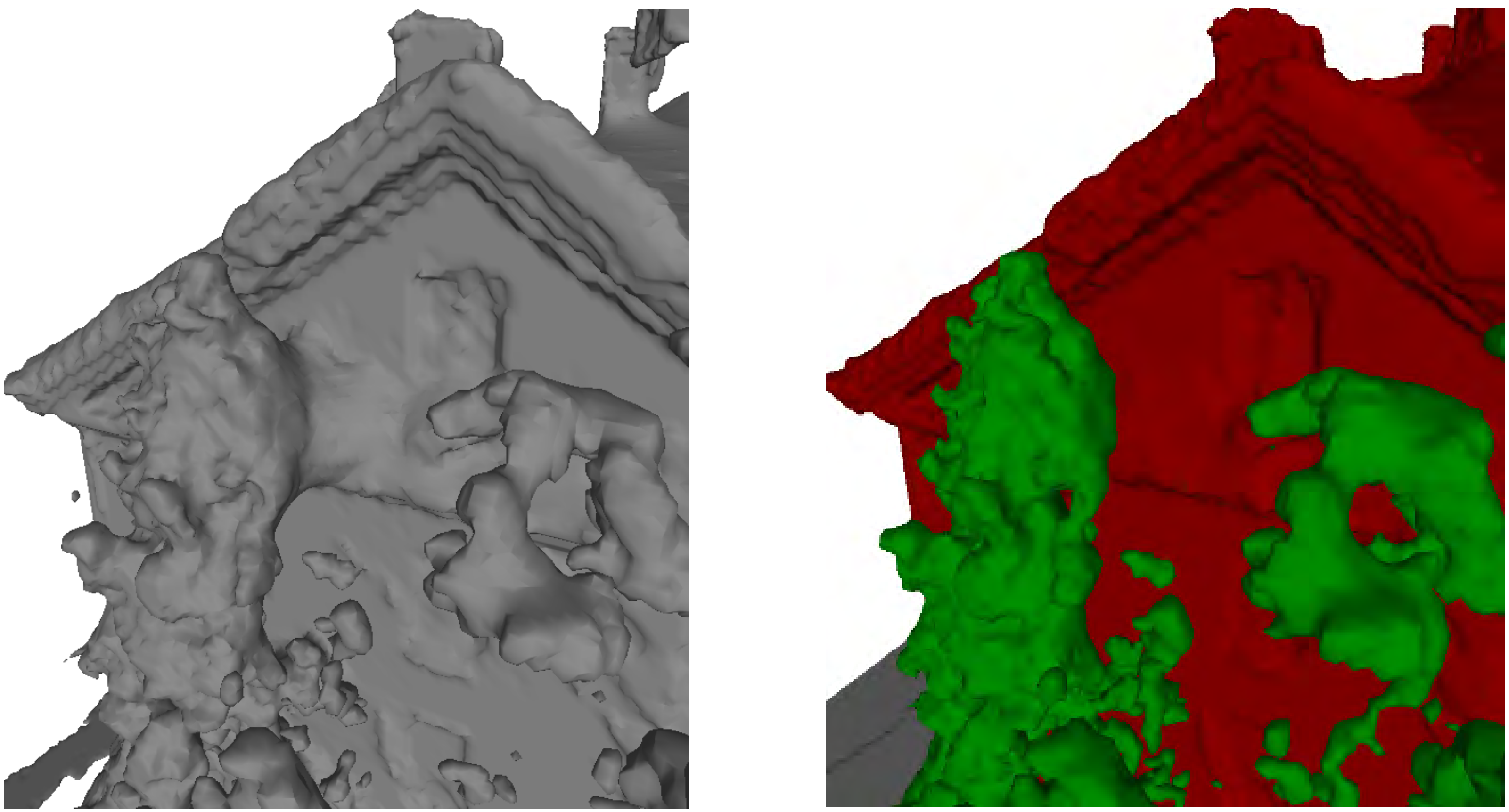
Outline

Semantic 3D reconstruction

- Joint reconstruction, recognition and segmentation
- High-order ray potentials
- Joint classifier
- Modeling objects

Weakly Observed Structures II

- Building separated from vegetation

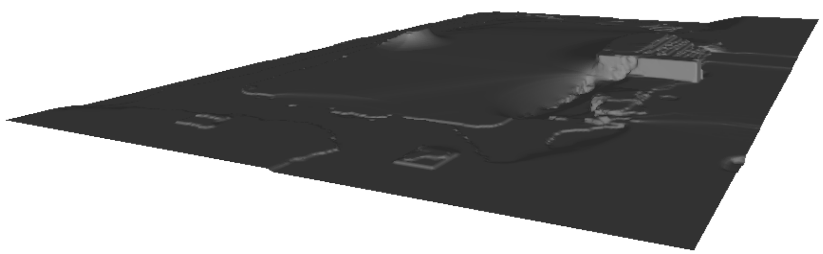


Unobserved Surfaces

- Labels can be separated

Unobserved Surfaces

- Labels can be separated



Higher-order ray potentials to model visibility

(Savinov et al, CVPR15/CVPR16)

- Volumetric formulation

$$E(\mathbf{x}) = \sum_{r\in\mathcal{R}} \psi_r(\mathbf{x}^r) + \sum_{(i,j)\in\mathcal{E}} \psi_p(x_i, x_j)$$

Ray potentials **Pairwise regularizer**

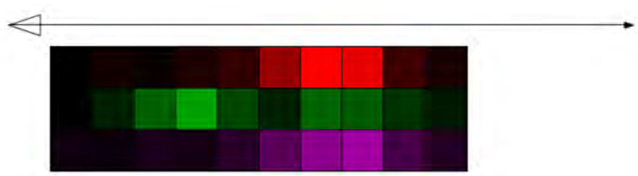
- Cost based on the first occupied voxel along the ray

$$\psi_r(\mathbf{x}^r) = \phi_r(K^r, x^r_{K^r})$$

depth **label**

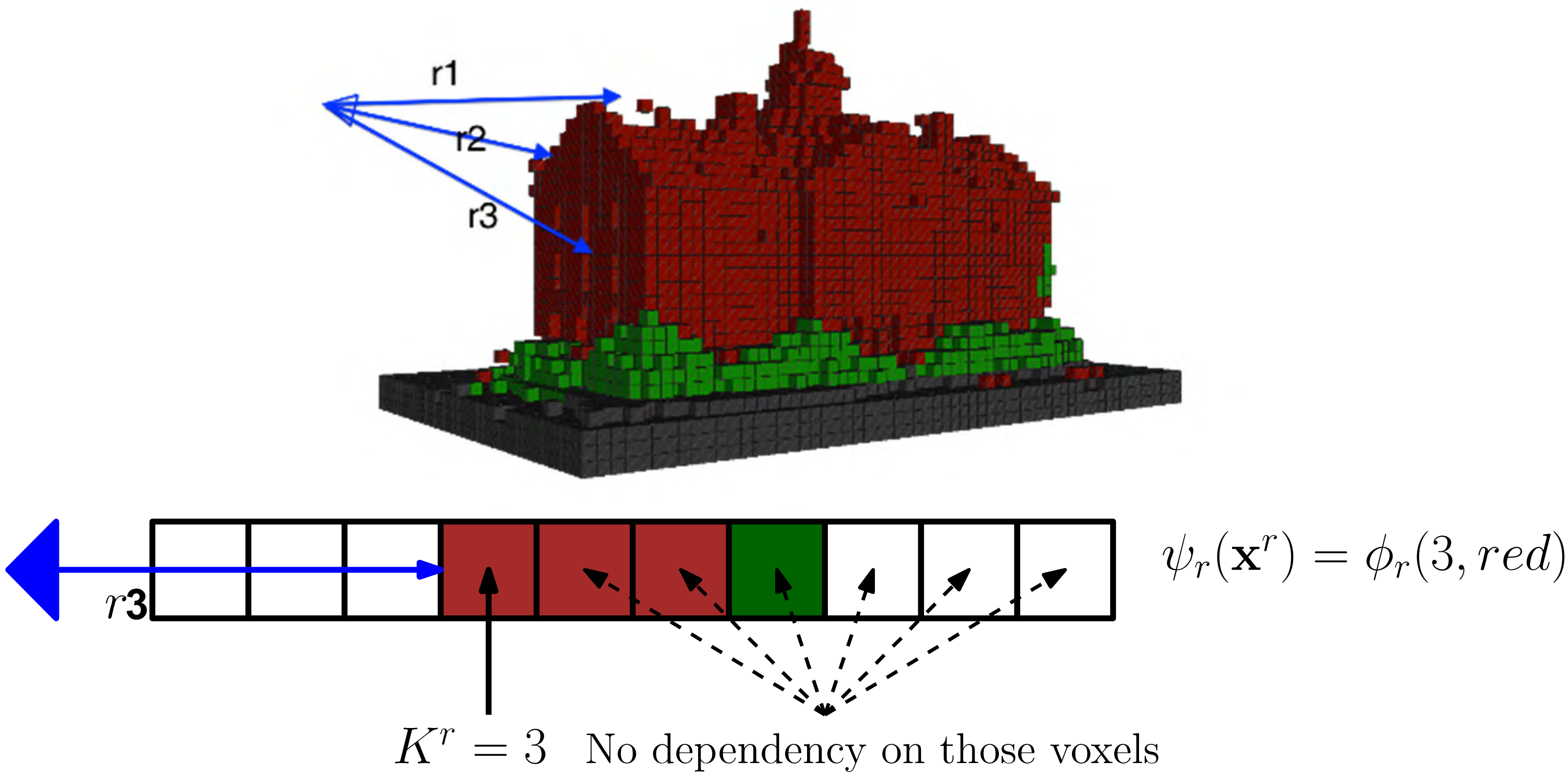
freespace

$$K^r = \begin{cases} \min(i \mid x_i^r \neq l_f) & \text{if } \exists x_i^r \neq l_f \\ N_r & \text{otherwise} \end{cases}$$



Cost based on the first occupied voxel along the ray

(Savinov et al, CVPR15/CVPR16)



Visibility Consistency Constraint

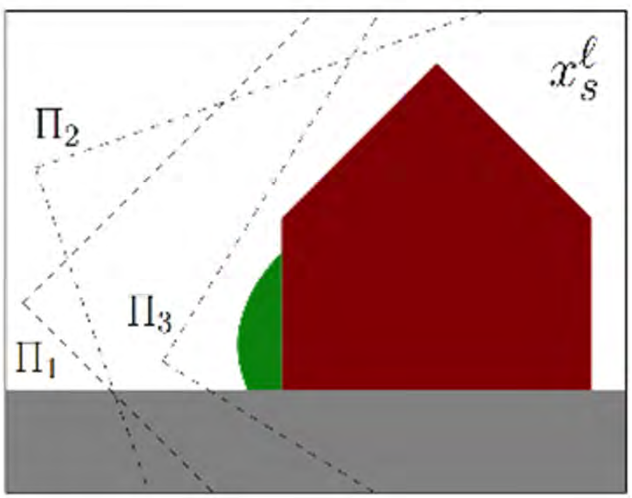
(Savinov et al, CVPR16)

$$\psi_r(\mathbf{x}_r) = \sum_{\ell \in \mathcal{L}} \sum_{i=0}^{N} c_i^\ell y_i^\ell$$

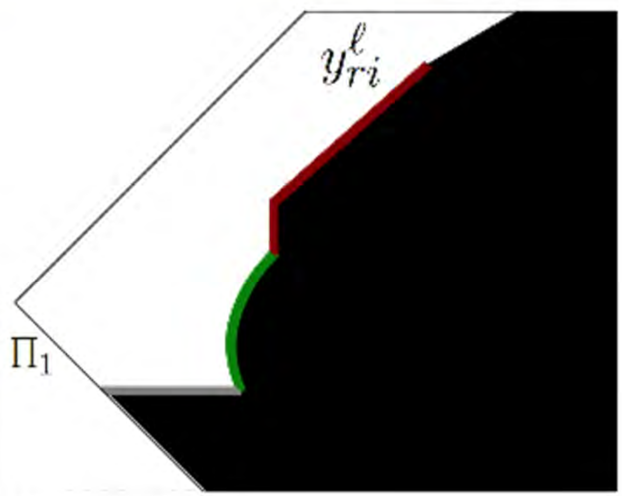
$$\text{s.t.} \quad y_i^\ell \leq y_{i-1}^f, \; y_i^\ell \leq x_{s_i}^\ell, \; y_i^\ell \geq 0 \quad \forall \ell \in \mathcal{L}, \forall i$$

$$\sum_{\ell \in \mathcal{L} \setminus \{f\}} y_i^\ell \leq \max(0, y_{i-1}^f - x_{s_i}^f) \qquad \forall i$$

(non-convex)



Global View

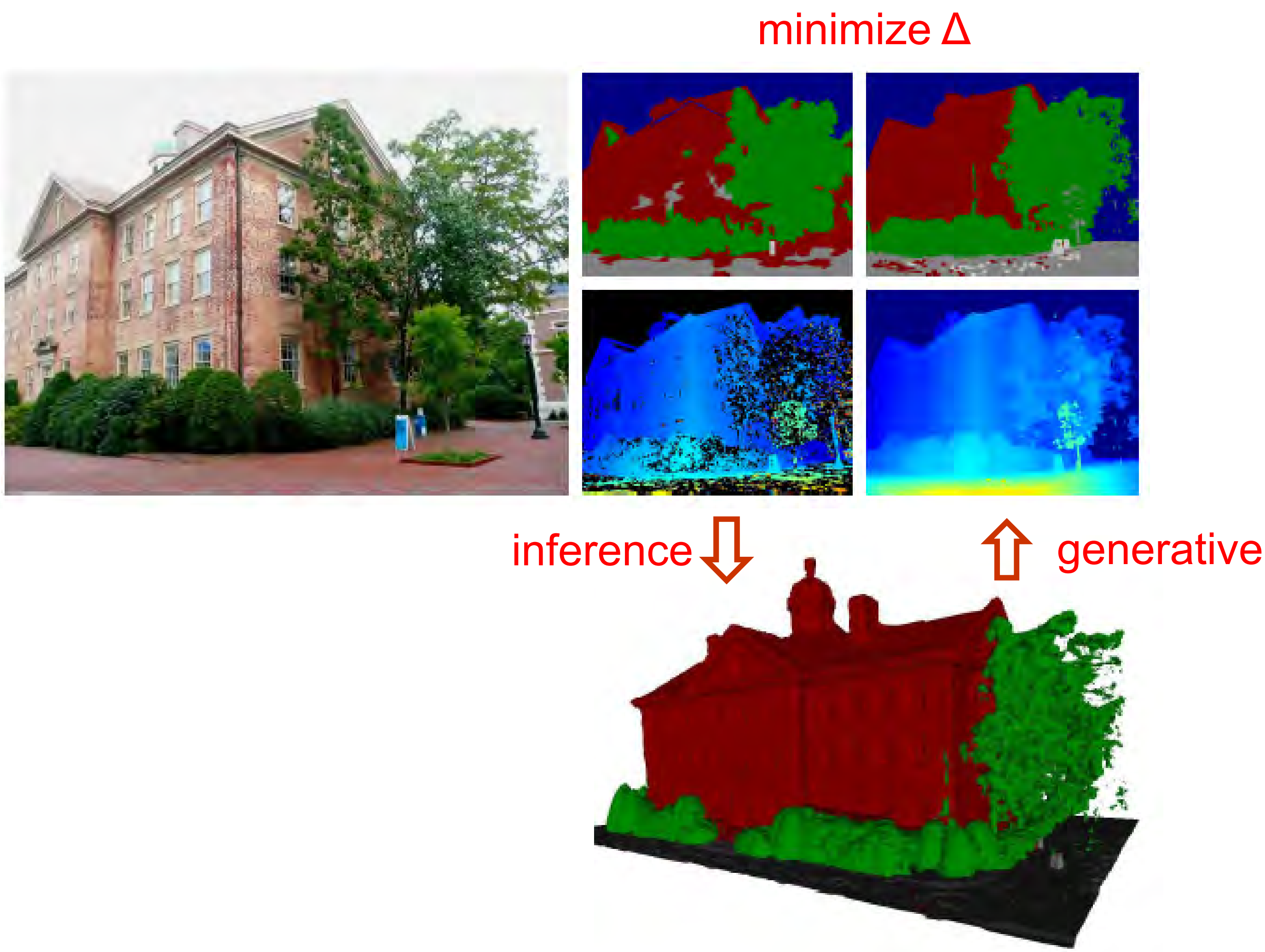


View from Camera Π_1

ETH Zürich

Computer Vision and Geometry Lab

Results



(Savinov et al, CVPR15/16)

2-label V-Constraint: SOTA on Middlebury

Multi-View Stereo | Evaluation · Datasets · Submit · Code

Acc. Threshold: 90% | Comp. Threshold: 1.25 mm

Data in new window | Open Data Window | Data: View 1 and Ground Truth | Image Size: Small

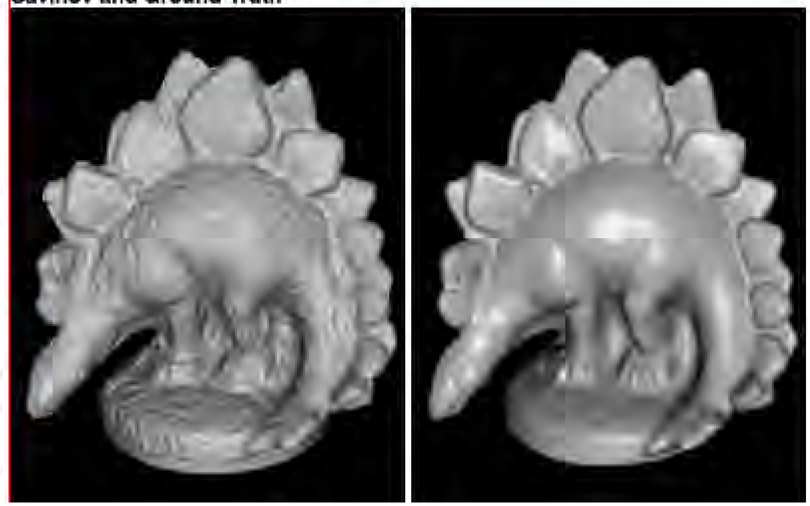
Tip: Mousing over any portion of a method's row will show its reference

Sort By	Temple Full 312 views Acc [mm]	Comp [%]	Temple Ring 47 views Acc [mm]	Comp [%]	Temple Sparse 16 views Acc [mm]	Comp [%]	Dino Full 363 views Acc [mm]	Comp [%]	Dino Ring 48 views Acc [mm]	Comp [%]	Dino Sparse 16 views Acc [mm]	Comp [%]
Savinov	0.41	99.7	0.5	99.5	0.69	97.8	0.26	99.8	0.25	99.9	0.34	98.7
Furukawa 3	0.49	99.6	0.47	99.6	0.63	99.3	0.33	99.8	0.28	99.8	0.37	99.2
DCV			0.73	98.2	0.66	97.3			0.28	100	0.3	100
Galliani	0.39	99.2	0.48	99.1	0.53	97.0	0.31	99.9	0.3	99.4	0.36	98.6
ECCV2016_624	0.37	98.9	0.49	97.6	1.27	39.2	0.26	97.8	0.31	99.5	0.28	98.1
3DV2014_25			0.51	96.4	1.23	90.2			0.32	97.3	0.42	96.7
Furukawa 2	0.54	99.3	0.55	99.1	0.62	99.2	0.32	99.9	0.33	99.6	0.42	99.2
Schroers	0.57	99.1	0.64	96.4	2.12	62.9	0.33	99.7	0.33	99.7	0.54	98.6
Kostrikov			0.57	99.1	0.79	95.8			0.35	99.6	0.37	99.3
Yichao Li	0.46	96.4	0.56	89.6			0.4	94.9	0.37	80.6		
Song			0.61	98.3					0.38	99.4	0.54	95.5
Khuboni			0.67	98.3					0.38	99.5		
Zhu			0.4	99.2	0.45	95.7			0.38	98.3	0.48	95.4

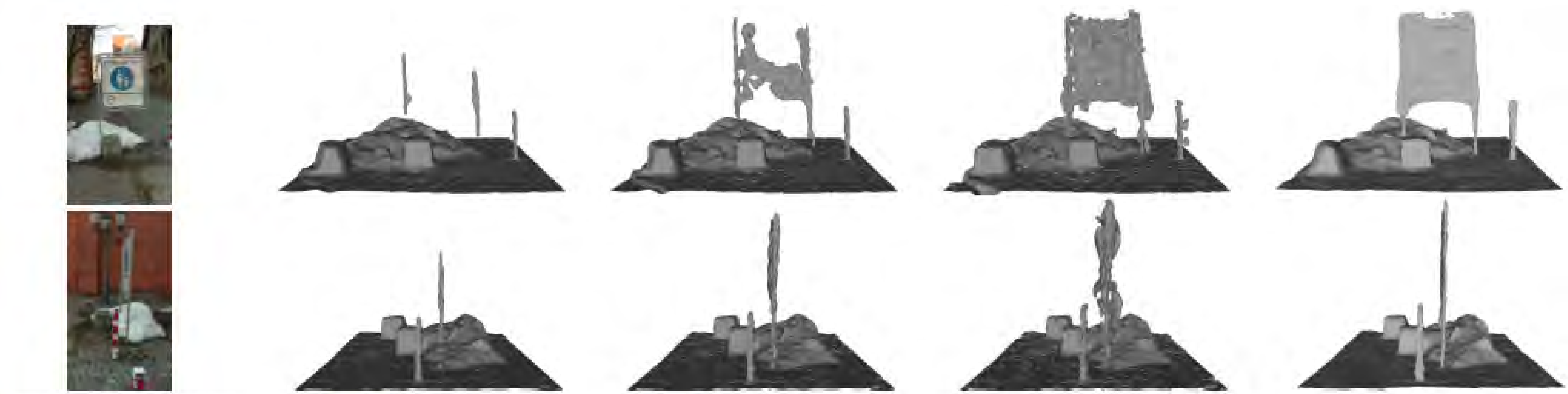
Reference: N. Savinov, C. Haene, L. Ladicky, and M. Pollefeys. Semantic 3D reconstruction with continuous regularization and ray potentials using a visibility consistency constraint. CVPR 2016.

Normalized Time (H:M:S): 43:12:00

Savinov and Ground Truth



V-Constraint: nicely handles thin objects



Example Images | TV-Flux (high) | TV-Flux (medium) | TV-Flux (low) | Our Method

Can our approach also handle objects?

- Extend approach from „Stuff“ to „Things“

$$E(x) = \sum_{\Omega \in s} \left(\sum_i \rho_s^i x_s^i + \sum_{i,j:i<j} \phi_s^{ij} (x_s^{ij} - x_s^{ji}) \right)$$

Introduce <u>location dependent</u> anisotropic smoothness prior

Learning location-dependent anisotropic smoothnes prior

(Haene et al CVPR 2014)

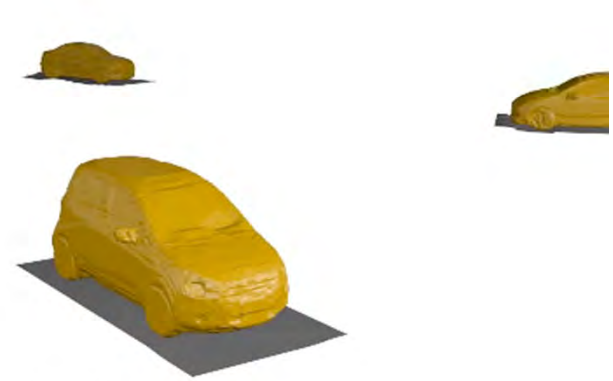
- Download training data from Google 3D warehouse

- At every voxel in 3D bounding box, estimate distribution of observed shape normals $P_s(n)$
- Determine convex Wulff shape which best represents observed statistics

$$d_s^n = -\log\left(P_s(n)\right)$$

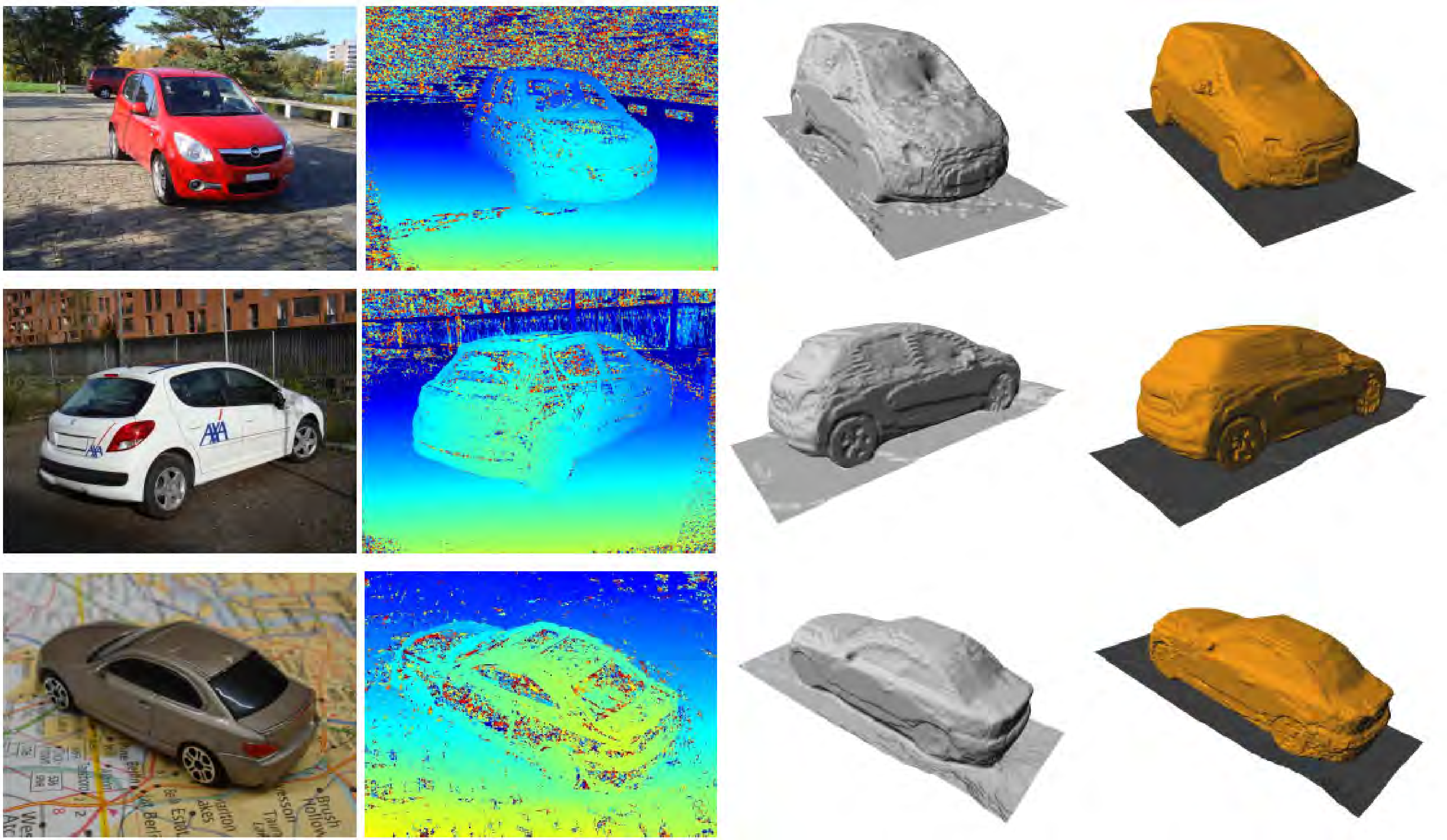
Cars

TV-Flux Fusion

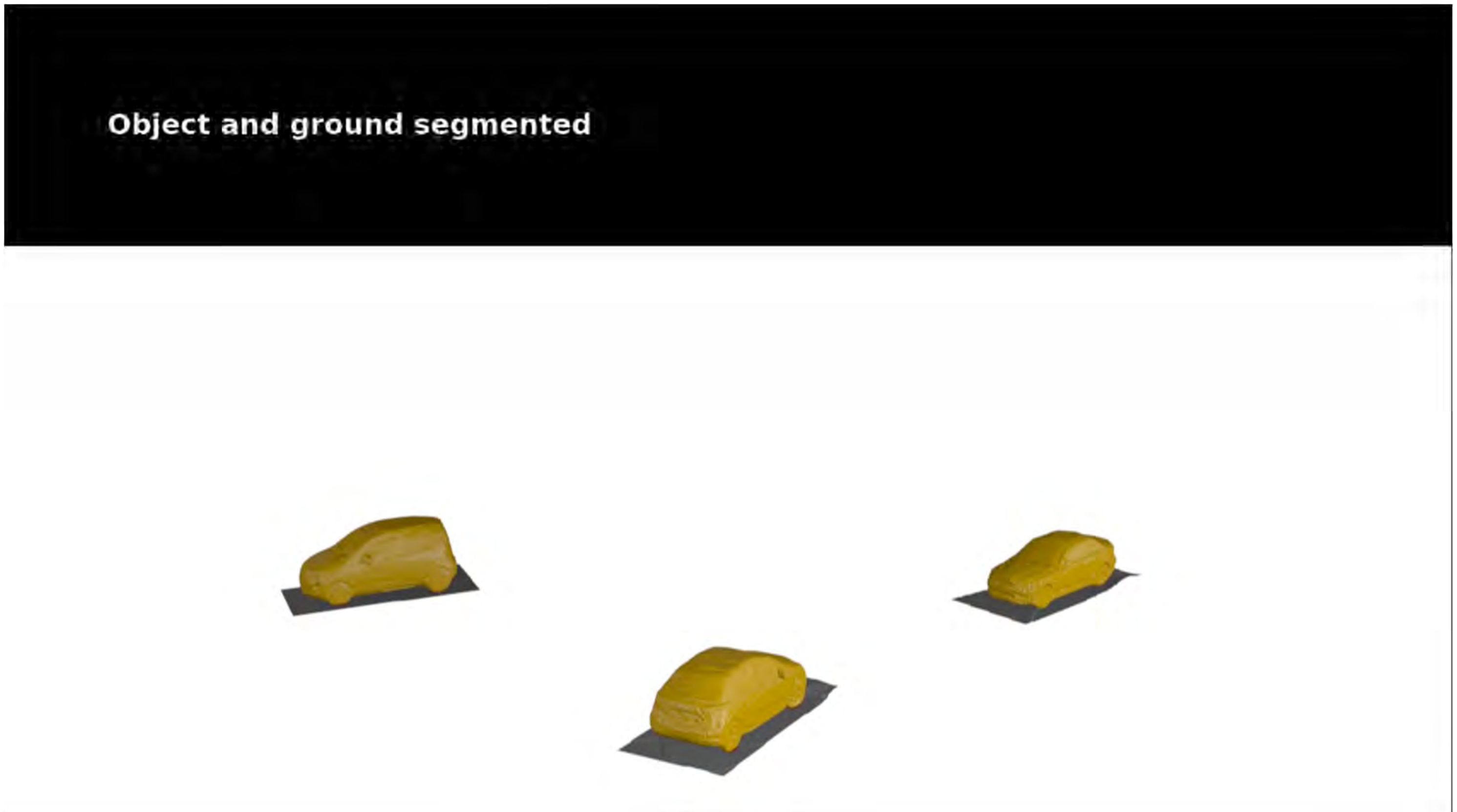


With Shape Prior

Car reconstructions



Advantages of multi-class segmentation



Learn separate statistics for *car-air* and *car-ground* transition likelihoods

Semantic multi-class 3D head reconstruction

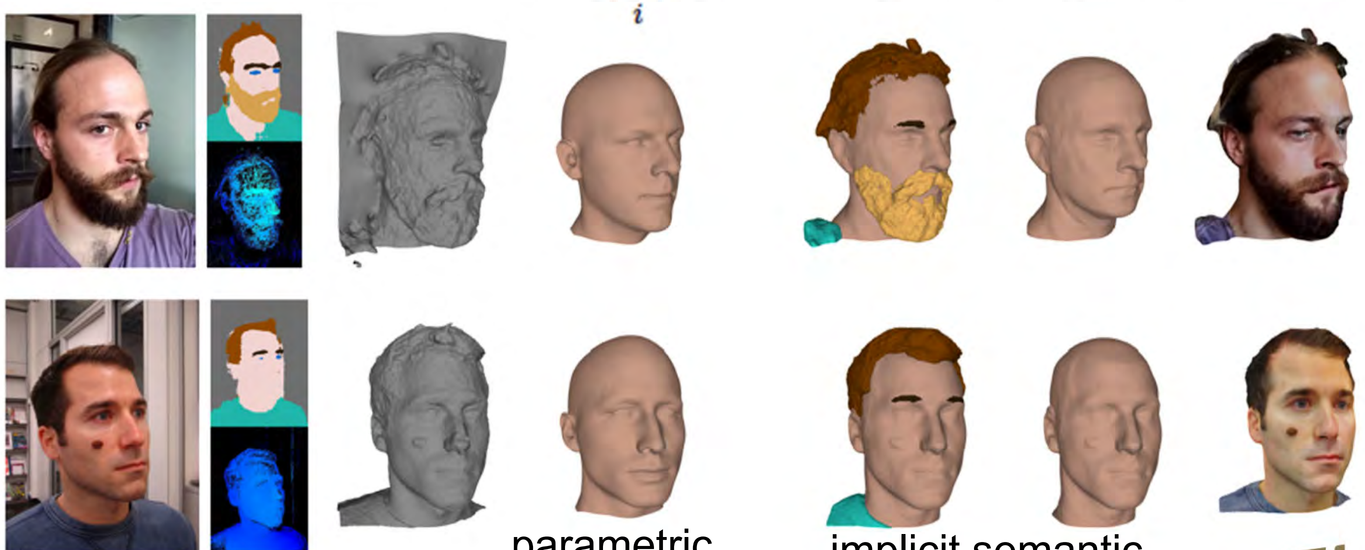
(Maninchedda et al. ECCV2016)

- Align prior to data by minimizing smoothness term towards position

$$E(\mathrm{x}, \underline{\mathcal{T}}) = \sum_{s \in \Omega} \left(\sum_i \rho_s^i(\mathcal{T}) x_s^i + \sum_{i,j:i<j} \phi_s^{ij}(\mathcal{T}, x_s^{ij} - x_s^{ji}) \right)$$

$$\text{s. t. } x_s^i = \sum_j (x_s^{ij})_k, \quad x_s^i = \sum_j (x_{s-e_k}^{ji}),$$

$$\sum_i x_s^i = 1, \quad x_s^i \geq 0, \quad x_s^{ij} \geq 0 . \qquad (1)$$



parametric shape model

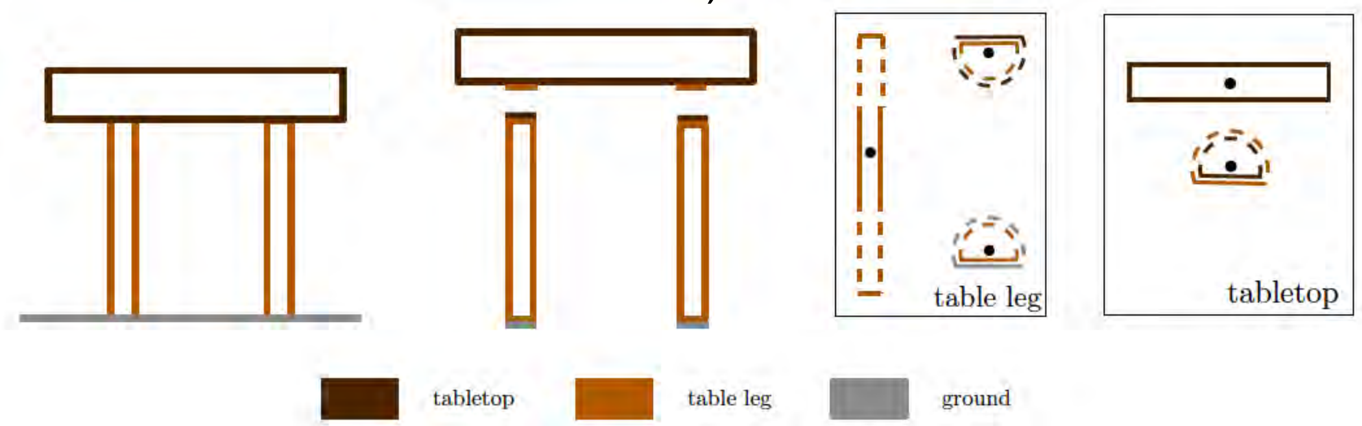
implicit semantic shape model

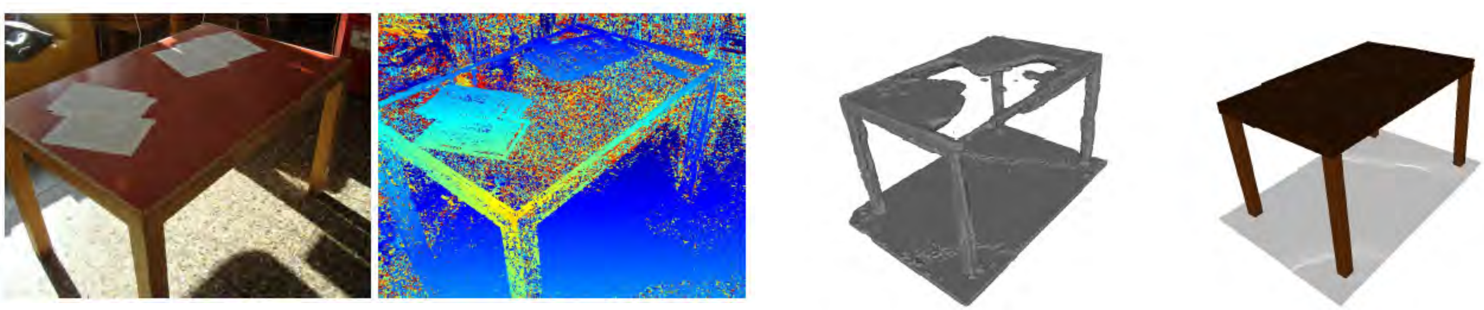
Segment based 3D object shape priors

represent non-convex object shape priors as combination of convex part priors

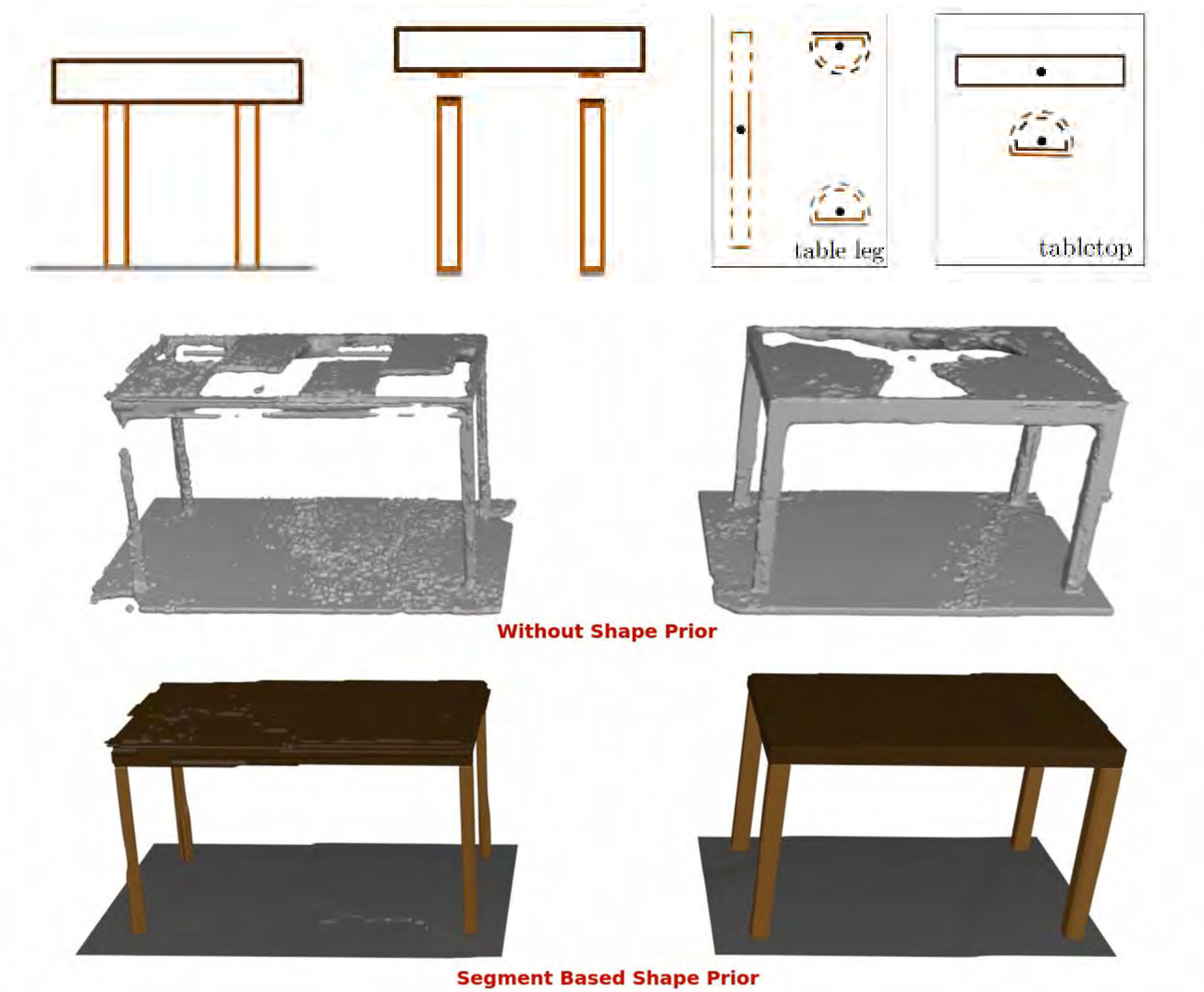
Karimi, Haene, Pollefeys (CVPR15)

build prior by performing (approximate) convex decomposition of example
(and observe which transitions occur)





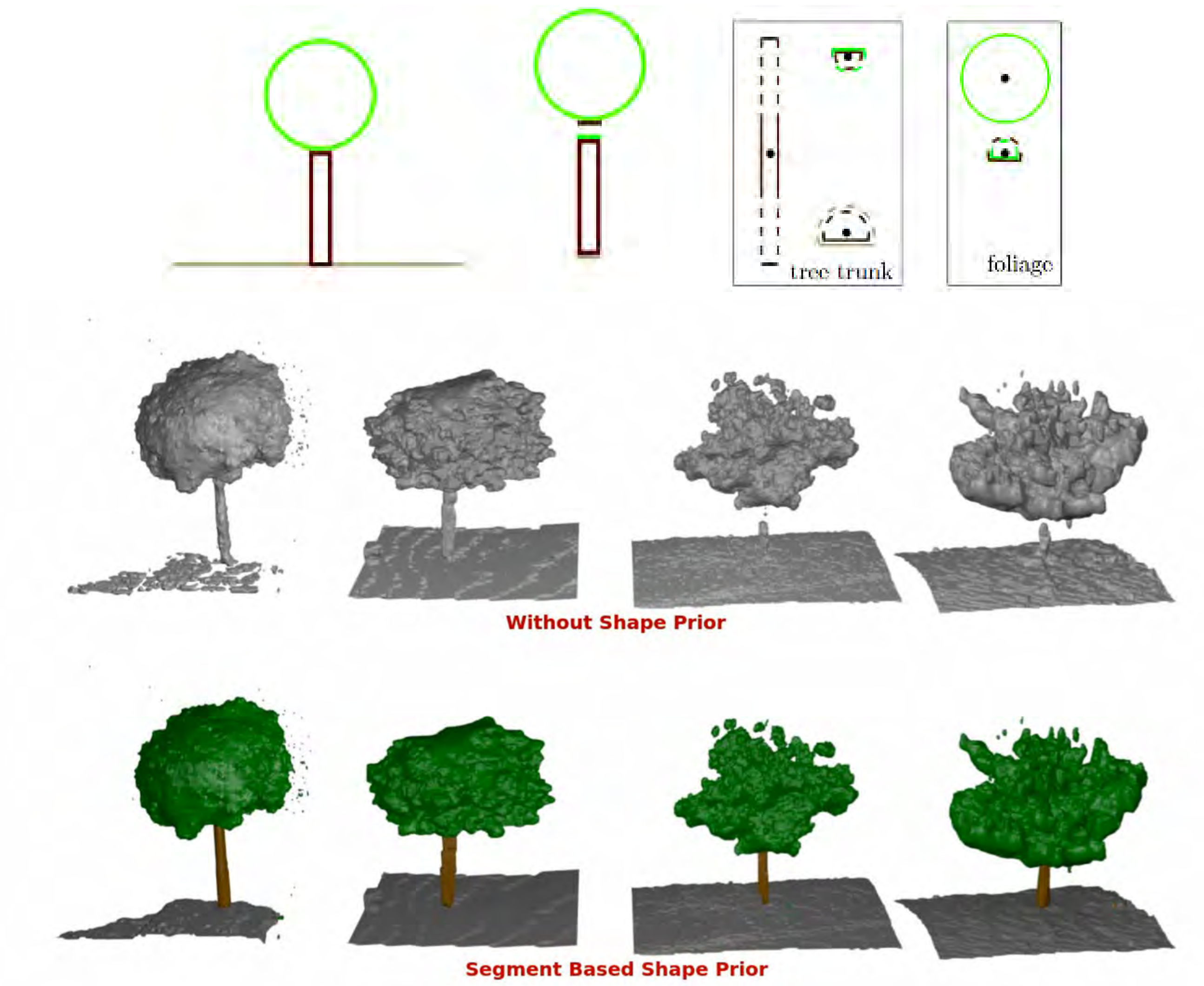
Result: Tables



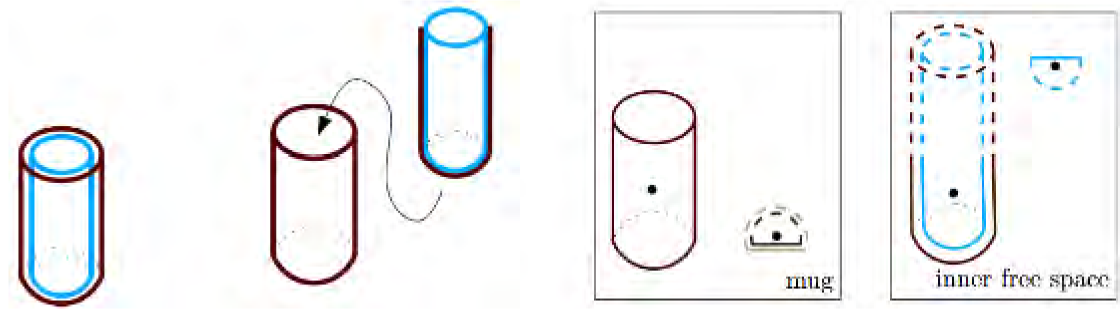
ETH Zürich

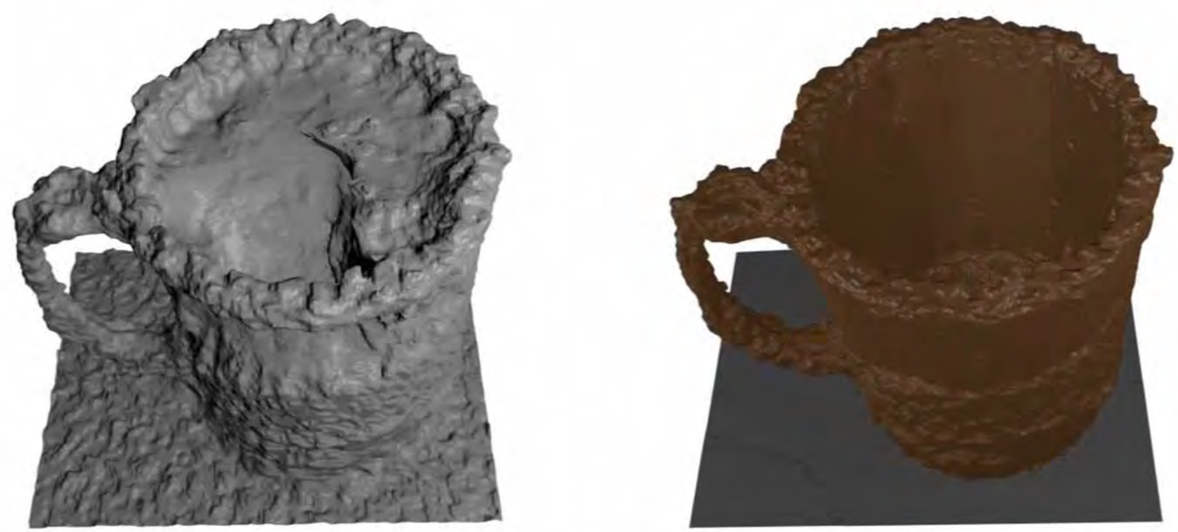
Computer Vision and Geometry Lab

Result: Trees



Result Mug





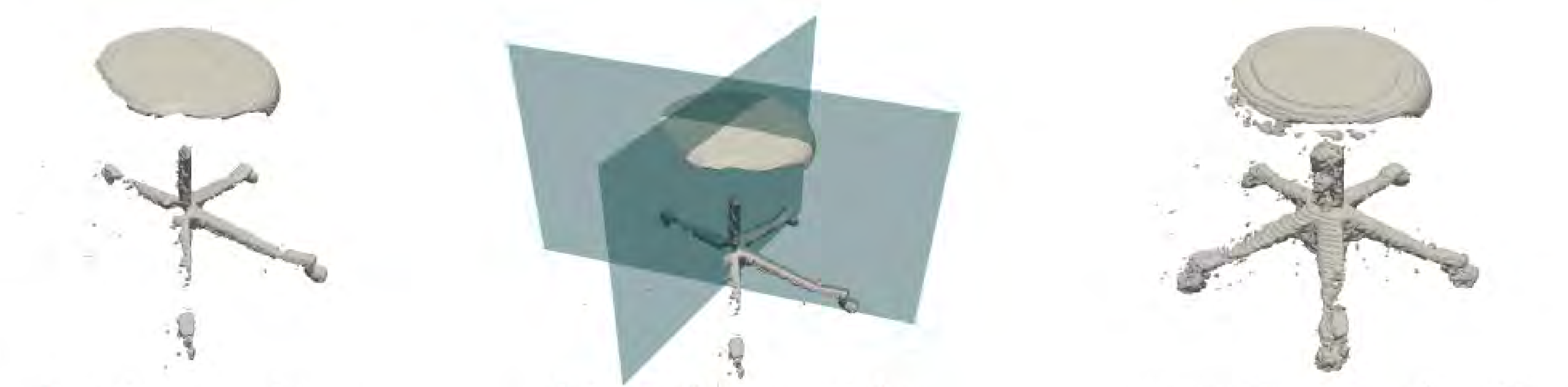
Without Shape Prior

Segment Based Shape Prior

Detect and regularize for symmetries

(Speciale et al. ECCV2016)

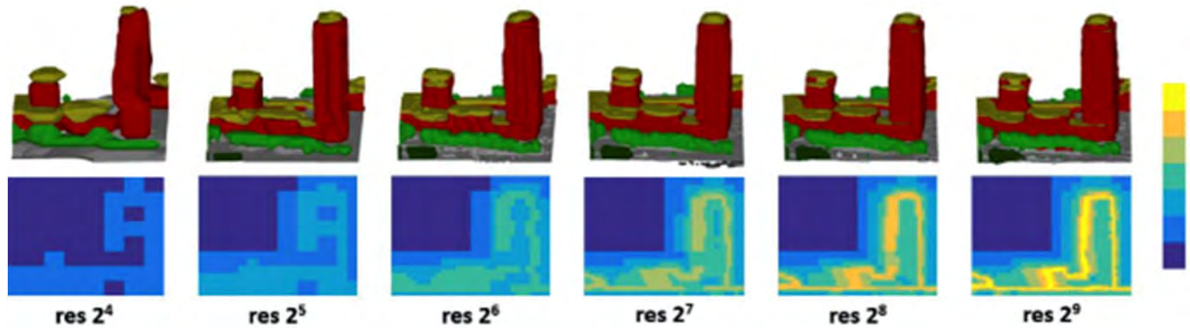
A preference for symmetry can be introduced by adding non-local regularization terms

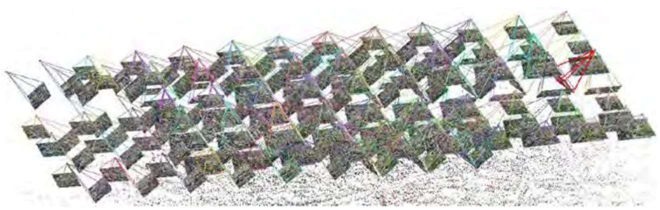


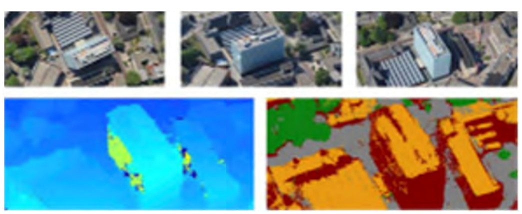
input geometry | detected symmetries | symmetric reconstruction

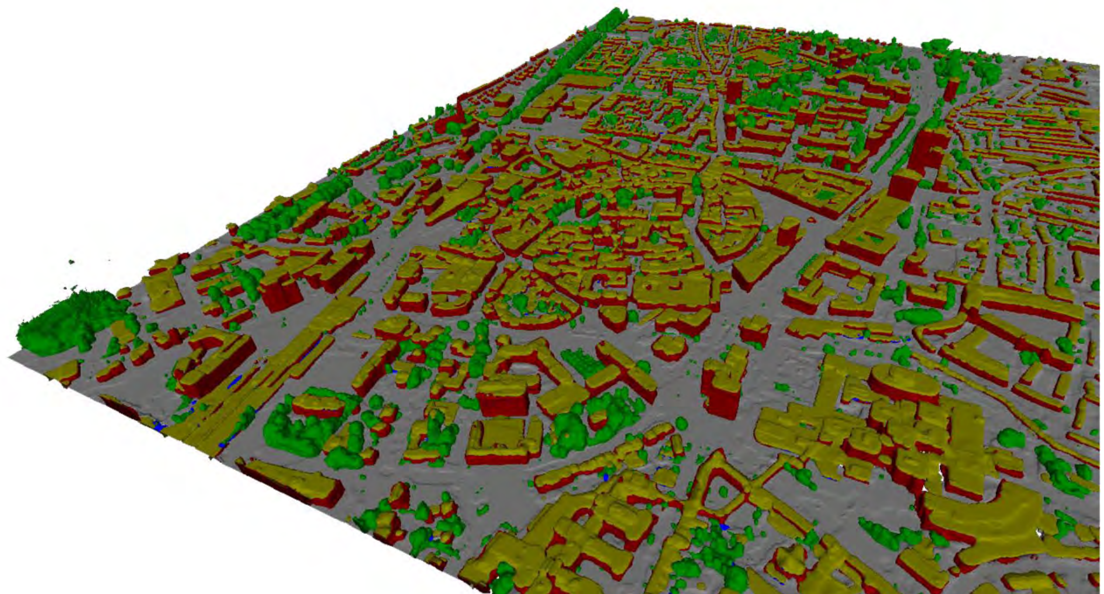
Semantic 3D Reconstruction

→ **Compared to a fixed-grid ≈ 20 x faster, 30 – 40 x less memory**
→ Allows for city scale reconstructions





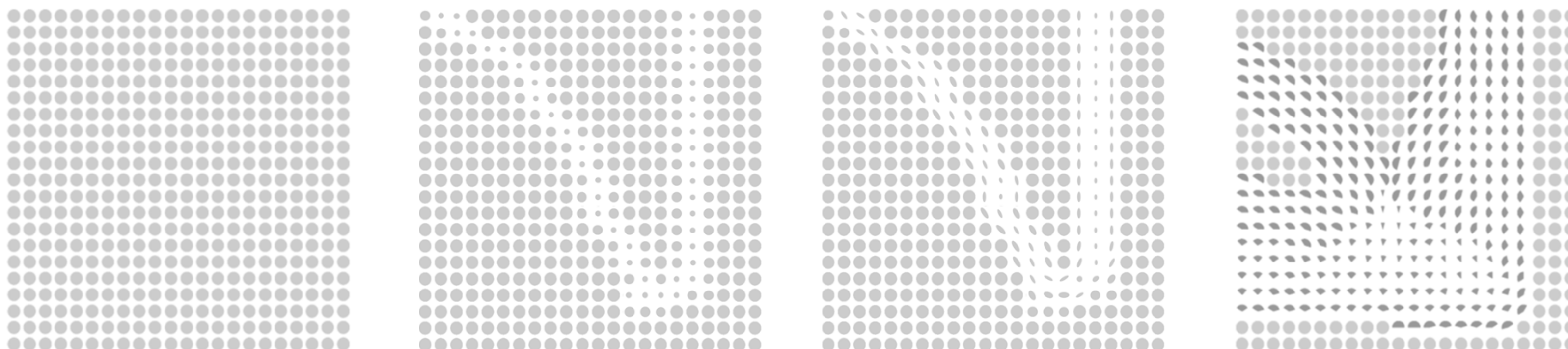




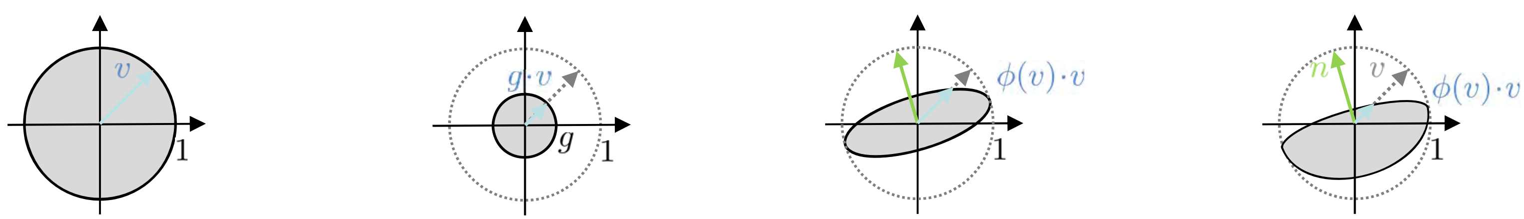
Large-Scale Semantic 3D Reconstruction: an Adaptive Multi-Resolution Model for Multi- Class Volumetric Labeling,
Maros Blaha, Christoph Vogel, Audrey Richard, Jan D. Wegner, Thomas Pock, Konrad Schindler, CVPR 2016

Overview of Regularizers

$$\underset{u}{\text{minimize}} \int_{\Omega} \Big(\underbrace{\phi_{\mathbf{x}}(u)}_{\text{regularization}} + \underbrace{f u}_{\text{data fidelity}} \Big)\, d\mathbf{x} \quad \text{subject to} \quad \forall \mathbf{x} \in \Omega : \sum_{\ell} u_{\ell}(\mathbf{x}) = 1$$



$\phi_{\boldsymbol{x}}(v) = \lVert v \rVert_2$	$\phi_{\boldsymbol{x}}(v) = g(x)\,\lVert v \rVert_2$	$\phi_{\boldsymbol{x}}(v) = \sqrt{v(x)^T D_x v(x)}$	$\phi_{\boldsymbol{x}}(v) = \max\limits_{\mu \in W_{\phi}} \langle \mu, v \rangle$



Isotropic spatially homogeneous TV	**Isotropic spatially varying weighted TV**	**Anisotropic spatially varying weighted TV**	**Anisotropic spatially varying Wulff shape**

Learning Regularization

→ Generalize gradient operator in regularizer
→ **L**earn label interactions

$$\| \nabla u \|_{2,1} \longrightarrow \| Wu \|_{2,1}$$

(Vogel and Pock GCPR2017) (Cherabier et al ECCV2018)

Multi-label segmentation/
3D reconstruction

$$\underset{u}{\text{minimize}} \quad \int_{\Omega} \left(\|Wu\|_2 + fu\right) d\mathbf{x} \quad \text{subject to} \quad \forall \mathbf{x} \in \Omega : \sum_{\ell} u_{\ell}(\mathbf{x}) = 1$$

Saddle-point problem

$$\underset{u}{\text{minimize}} \max_{\|\xi\|_{\infty} \leq 1} \langle Wu, \xi \rangle + \langle f, u \rangle + \nu \left(1 - \sum_{\ell} u_{\ell}\right)$$

Iterate primal-dual
update steps

$$\nu^{t+1} = \nu^t + \sigma \left(\sum_{\ell} \bar{u}_{\ell}^t - 1\right) \qquad u^{t+1} = \Pi_{[0,1]} \left[u^t + \tau(W^* \xi^{t+1} - f)\right]$$

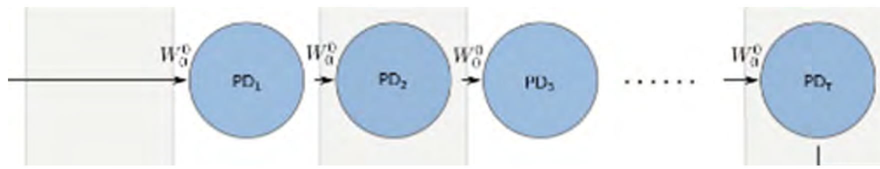
$$\xi^{t+1} = \Pi_{\|\cdot\| \leq 1} \left[\xi^t + \sigma W \bar{u}^t\right] \qquad \bar{u}^{t+1} = 2u^{t+1} - u^t$$

Neural Network via Optimization Unrolling

Iterate primal-dual **update steps**

$$\nu^{t+1} = \nu^t + \sigma\left(\sum_\ell \bar{u}_\ell^t - 1\right) \qquad u^{t+1} = \Pi_{[0,1]}\left[u^t + \tau(W^*\xi^{t+1} - f)\right]$$

$$\xi^{t+1} = \Pi_{\|\cdot\|\leq 1}\left[\xi^t + \sigma W\bar{u}^t\right] \qquad \bar{u}^{t+1} = 2u^{t+1} - u^t$$



Primal Dual

ETH Zürich

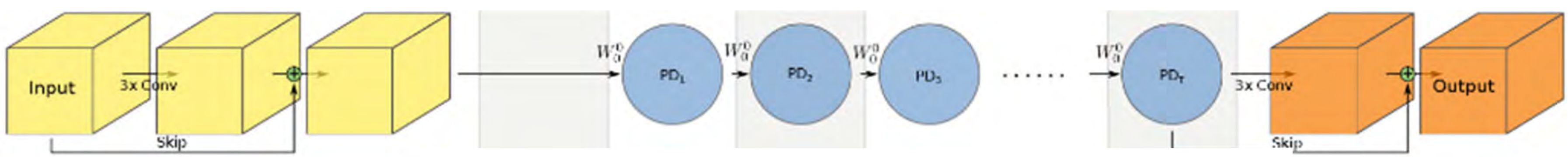
Computer Vision and Geometry Lab

Neural Network via Optimization Unrolling

Iterate primal-dual **update steps**

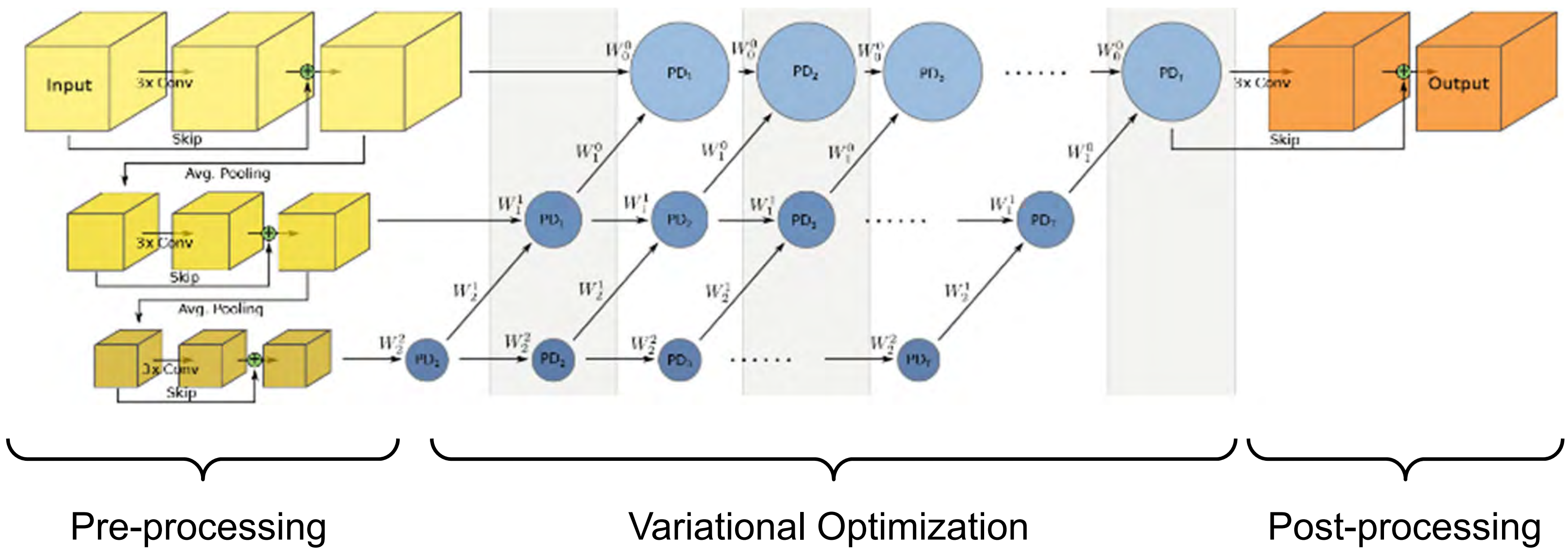
$$\nu^{t+1} = \nu^t + \sigma\left(\sum_\ell \bar{u}_\ell^t - 1\right) \qquad u^{t+1} = \Pi_{[0,1]}\left[u^t + \tau(W^*\xi^{t+1} - f)\right]$$

$$\xi^{t+1} = \Pi_{\|\cdot\|\leq 1}\left[\xi^t + \sigma W \bar{u}^t\right] \qquad \bar{u}^{t+1} = 2u^{t+1} - u^t$$

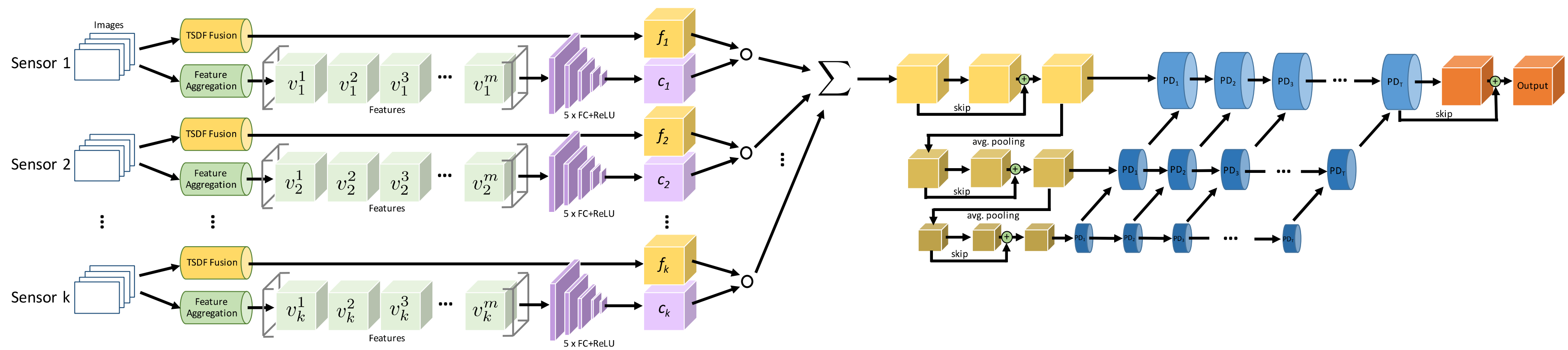


Pre-processing **Primal Dual** **Post-processing**

Multi-scale Architecture



Potential to combine with Multi-Sensor Depth Map Fusion



Multi-Sensor Aggregation

Semantic 3D Reconstruction

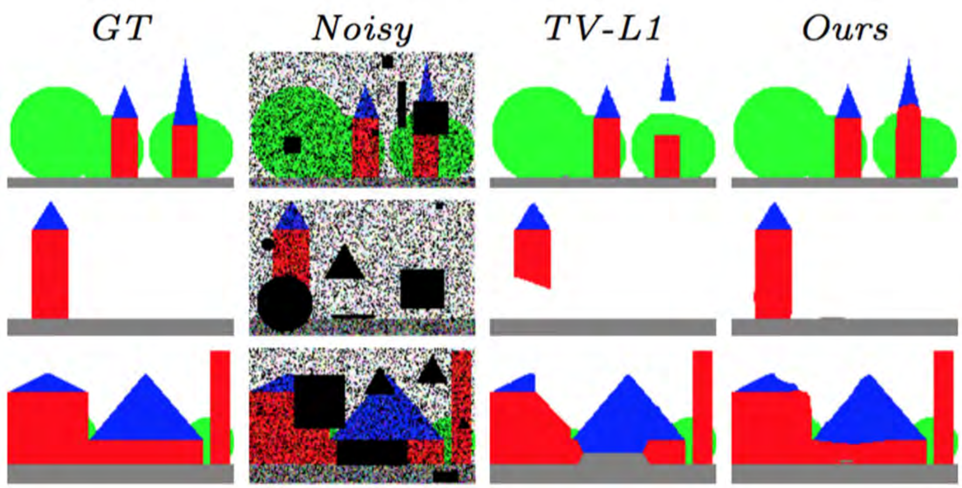
$$\underset{u}{\text{minimize}} \quad \int_{\Omega} \Big(\|W u\|_2 + \sum_{s \in \mathcal{S}} (c_s \circ f_s)\ u \Big)\, d\mathbf{x}$$

$$\text{subject to} \quad \forall \mathbf{x} \in \Omega : \sum_{\ell \in \mathcal{L}} u_\ell\, (\mathbf{x}) = 1$$

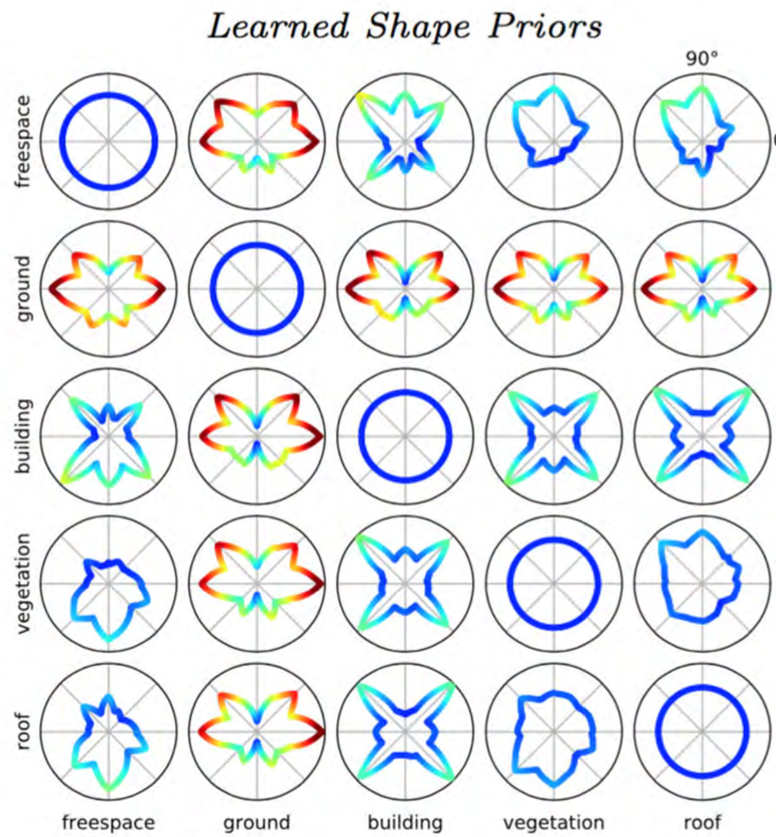
ETH Zürich

Computer Vision and Geometry Lab

2D Experiments



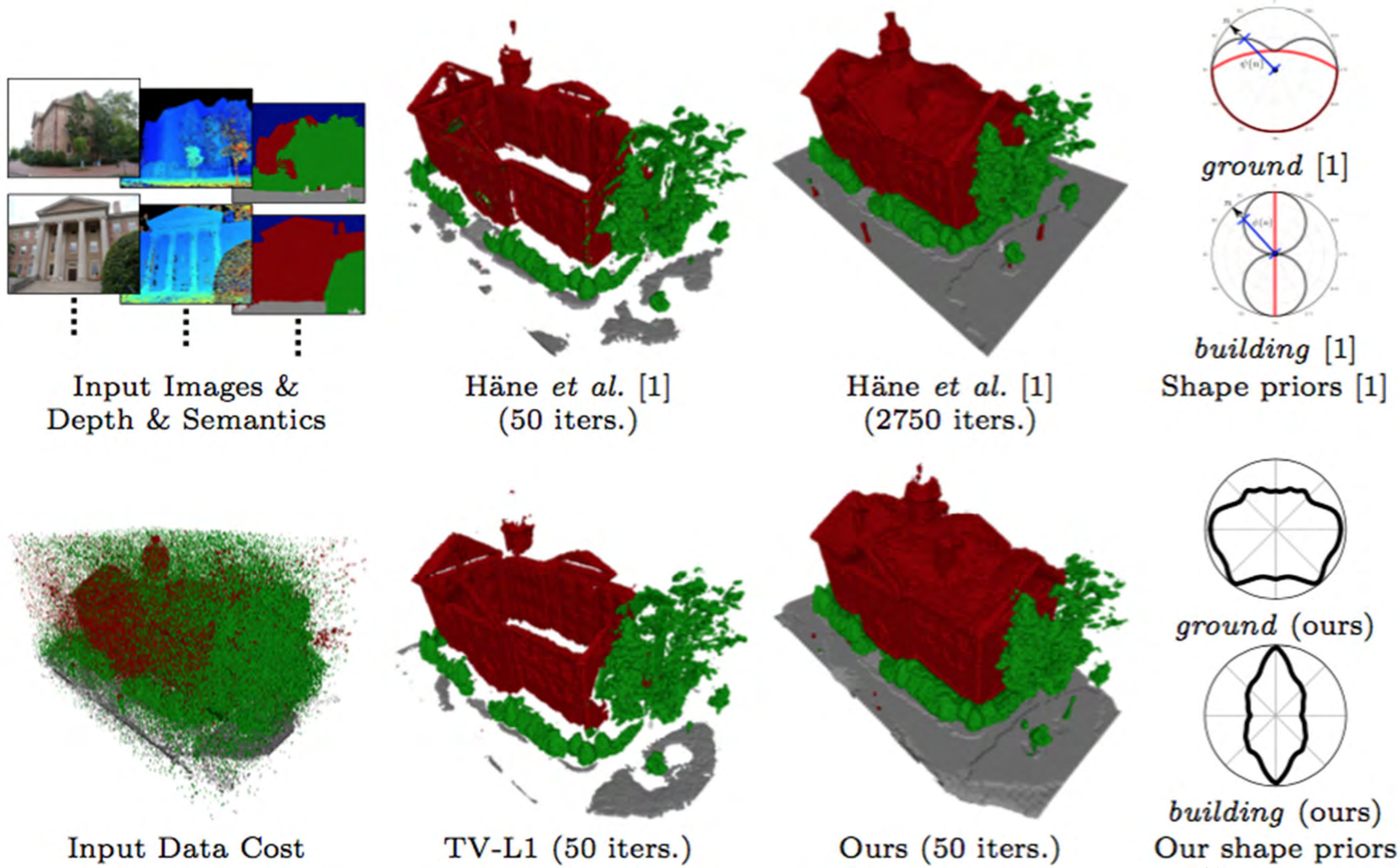
T \ S	1	2	3	4	3+E	3+E+D	TV-L1
10	76.83	97.57	98.10	98.34	**99.37**	99.32	97.73
	38.58	82.11	87.43	88.74	94.94	**95.14**	79.50
20	90.76	98.26	98.85	98.86	99.38	**99.41**	98.40
	49.13	88.80	91.42	91.83	95.16	**95.23**	85.94
50	97.21	98.99	99.19	99.21	99.20	**99.38**	98.70
	74.36	91.56	91.42	93.20	93.57	**94.86**	88.31



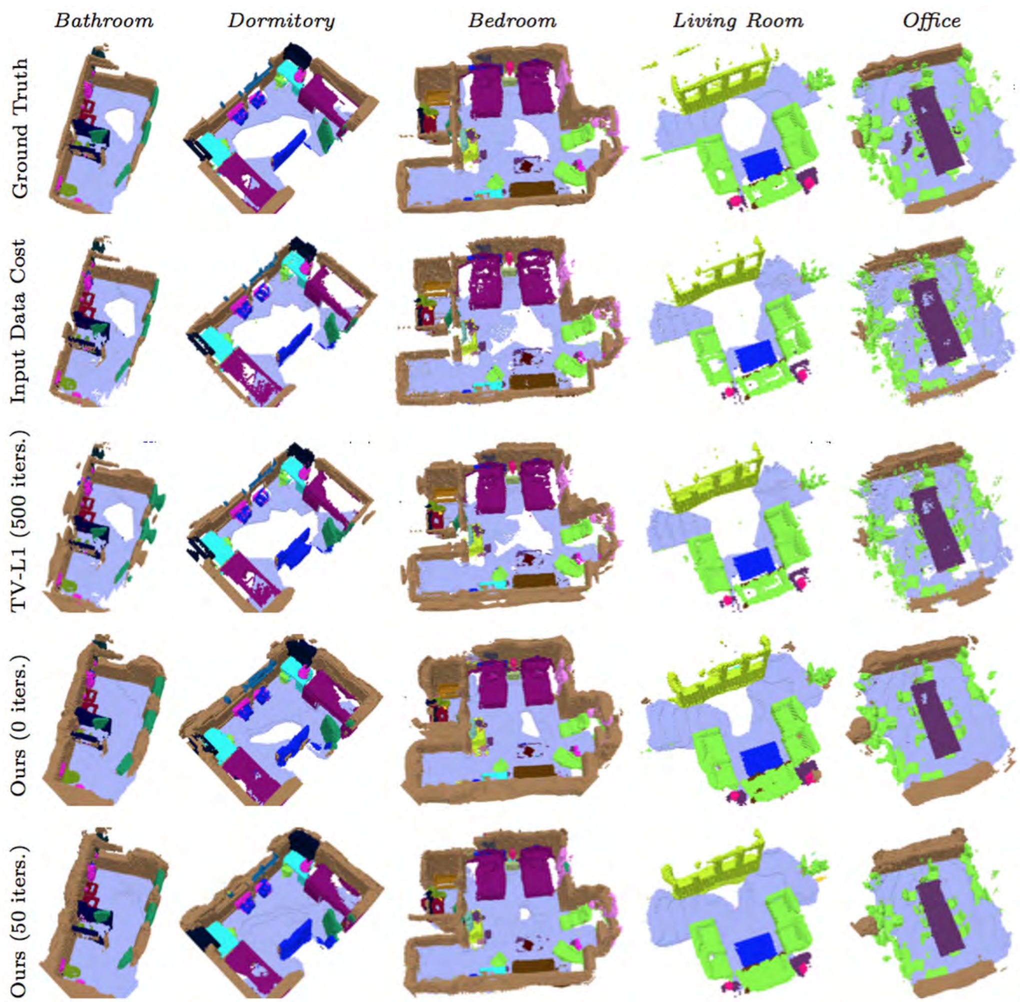
ETH Zürich

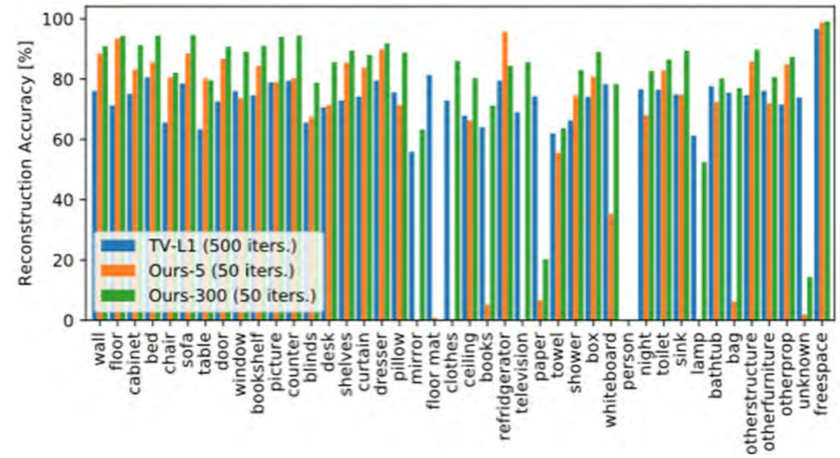
Computer Vision and Geometry Lab

3D Experiments



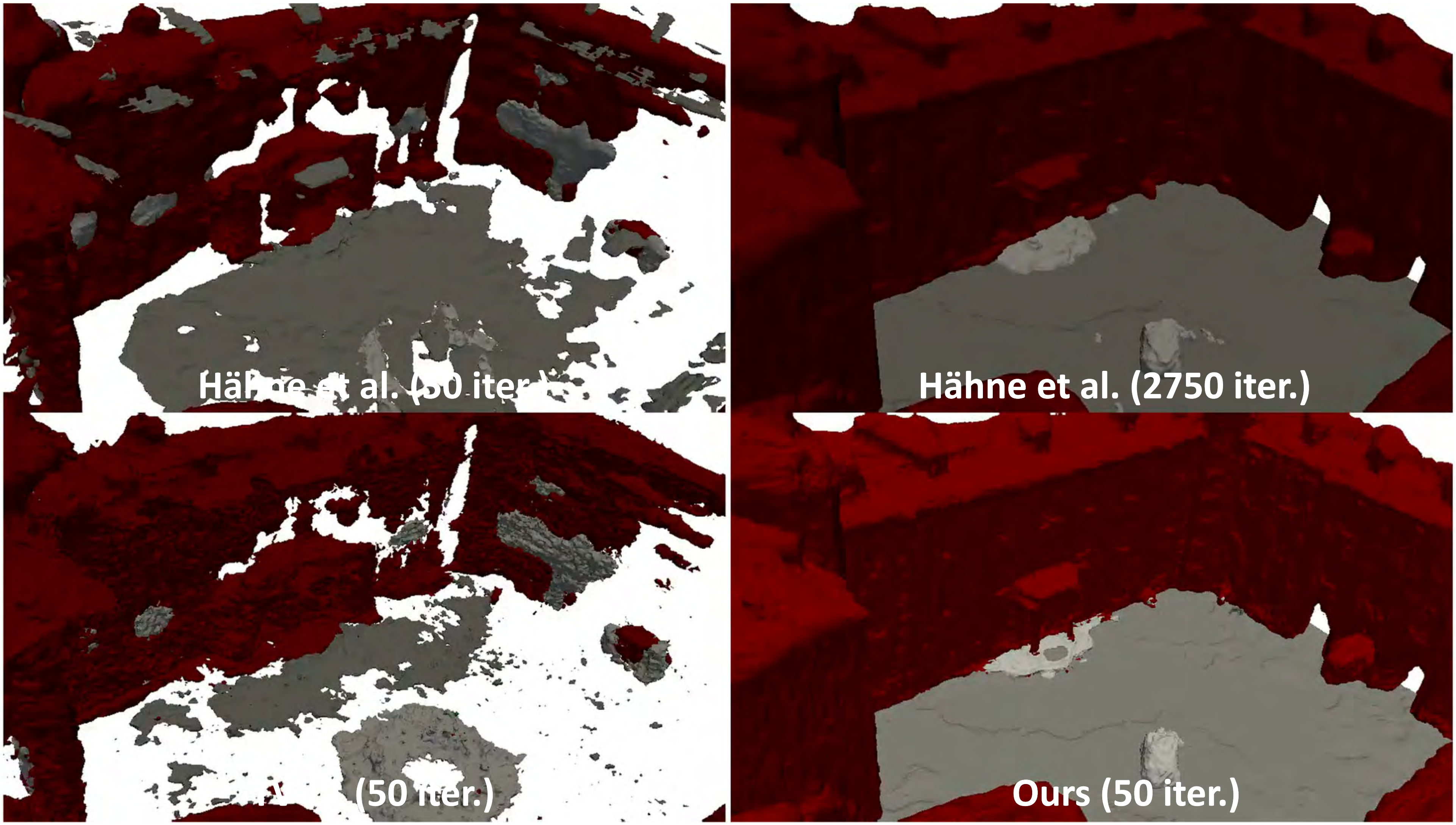
3D Experiments on ScanNet





Methods	Overall	Freespace	Occupied	Semantic
Input data	59.8	39.1	99.7	68.4
TV-L1 (50 it.)	92.8	71.0	91.4	87.8
TV-L1 (500 it.)	95.8	86.4	92.3	88.5
C2F (50 it.)	21.0	26.7	99.9	31.4
Ours-5 (50 it.)	96.7	95.8	93.9	86.4
Ours-300 (0 it.)	97.3	97.6	92.3	90.2
Ours-300 (50 it.)	**98.7**	**98.6**	**94.4**	**91.5**

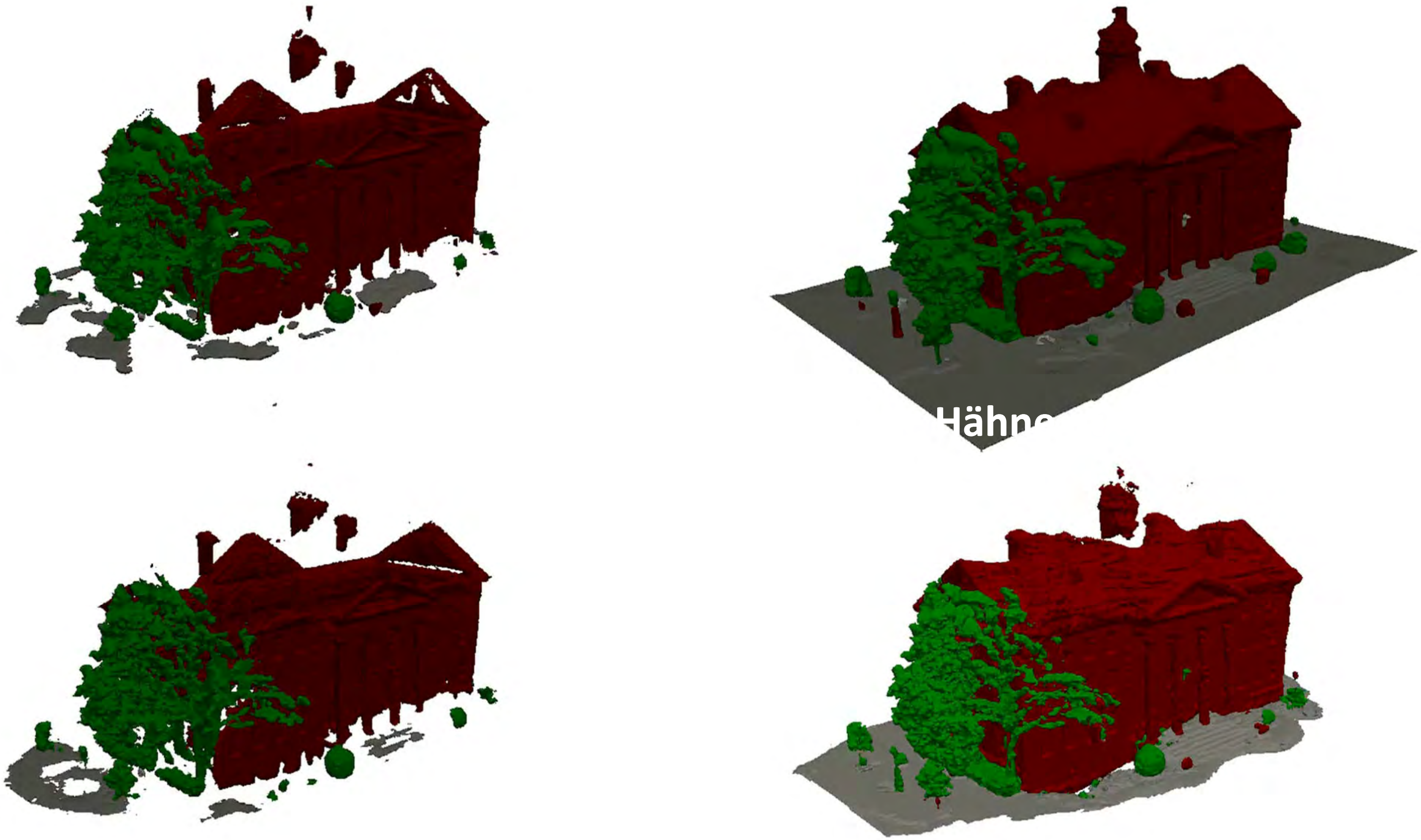
Learning Regularization



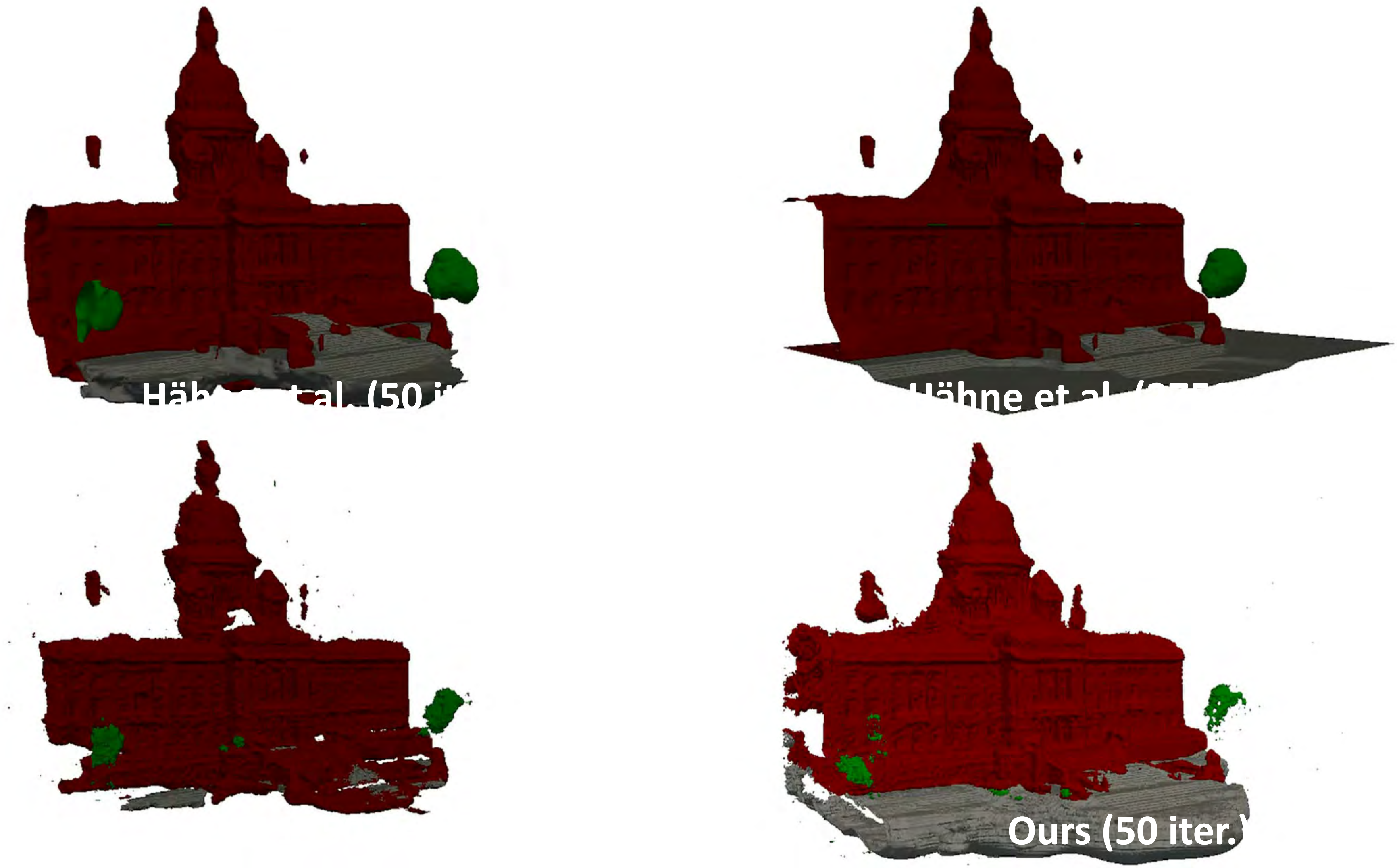
ETH Zürich

Computer Vision
and Geometry Lab

Learning Regularization

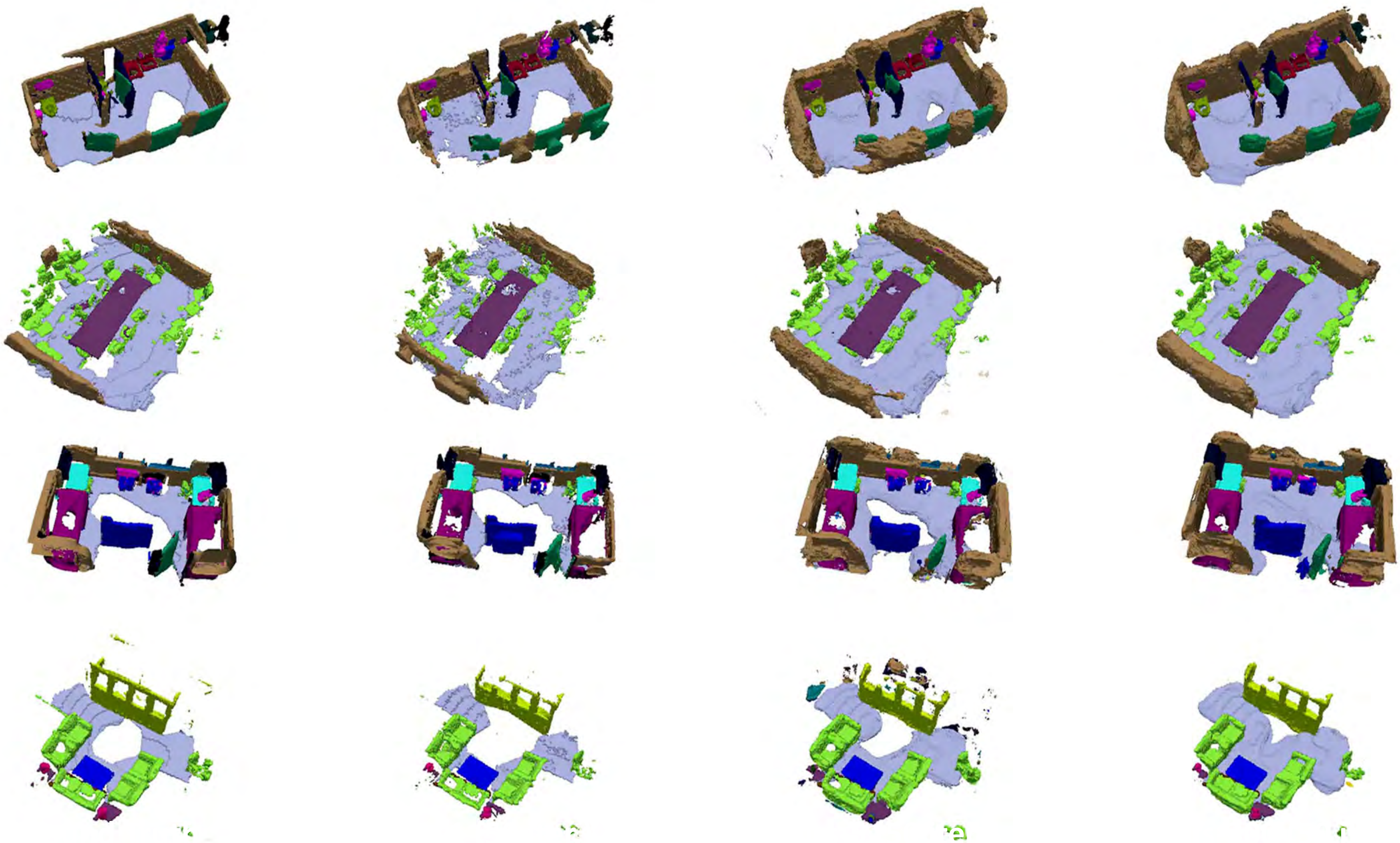


Learning Regularization

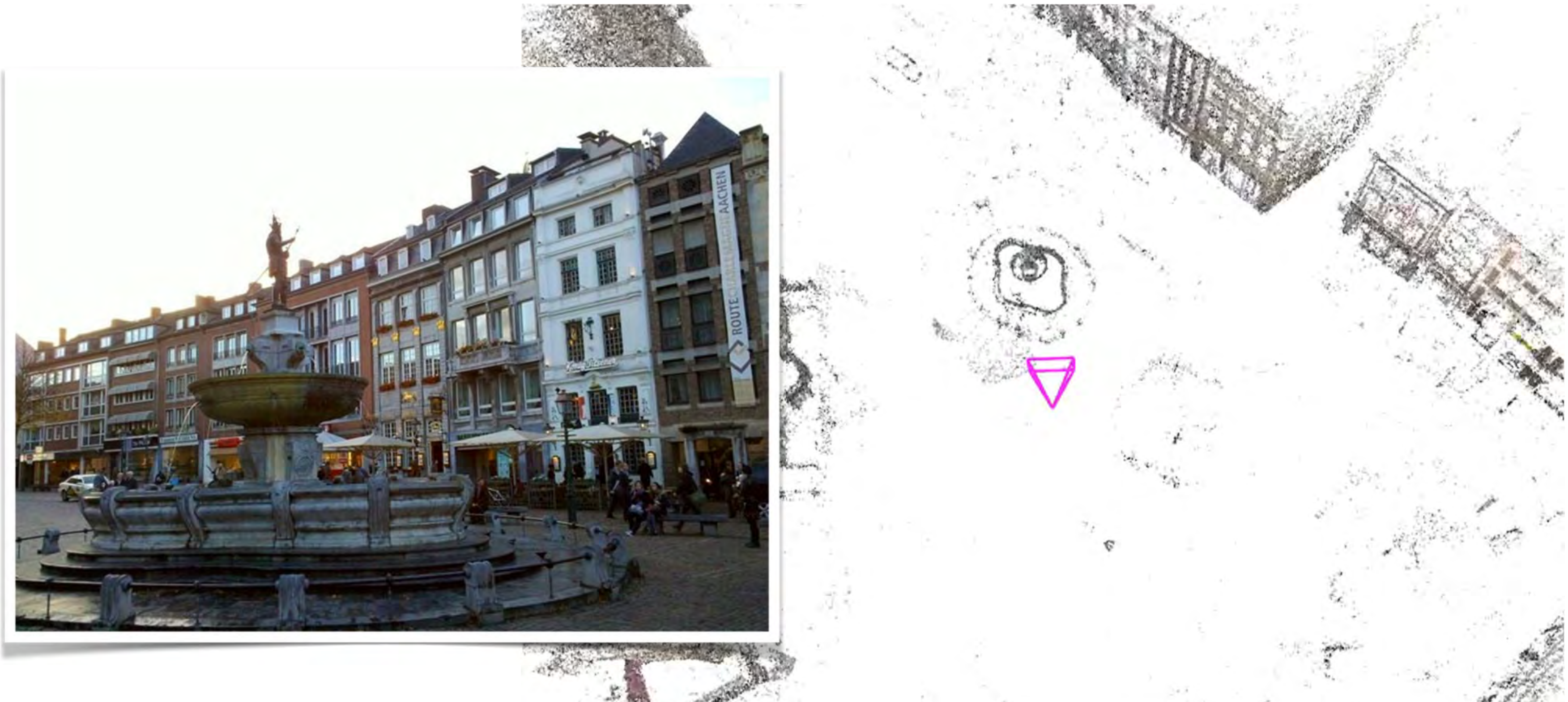


Learning Regularization

Learning Regularization

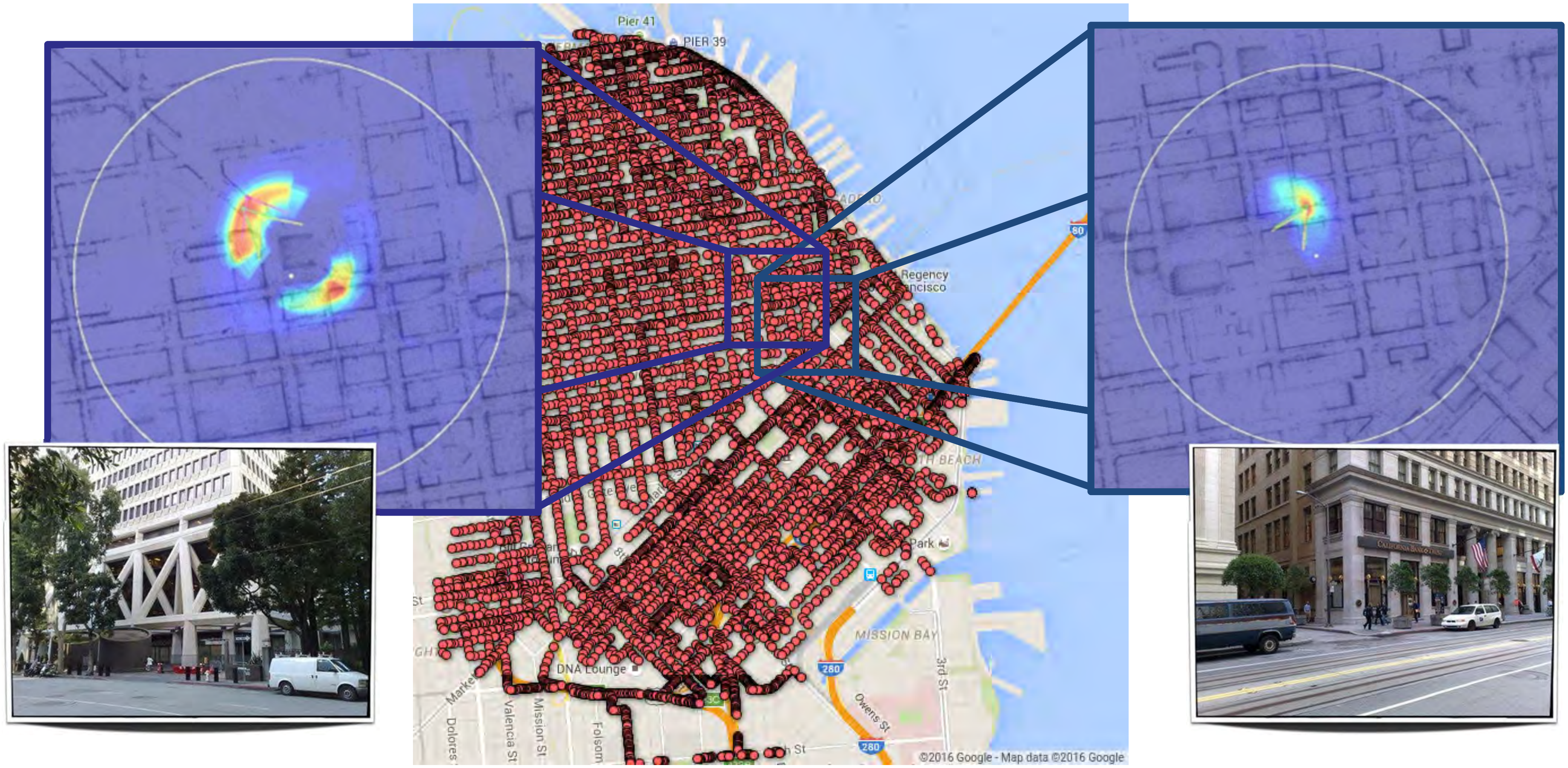


Visual Localization



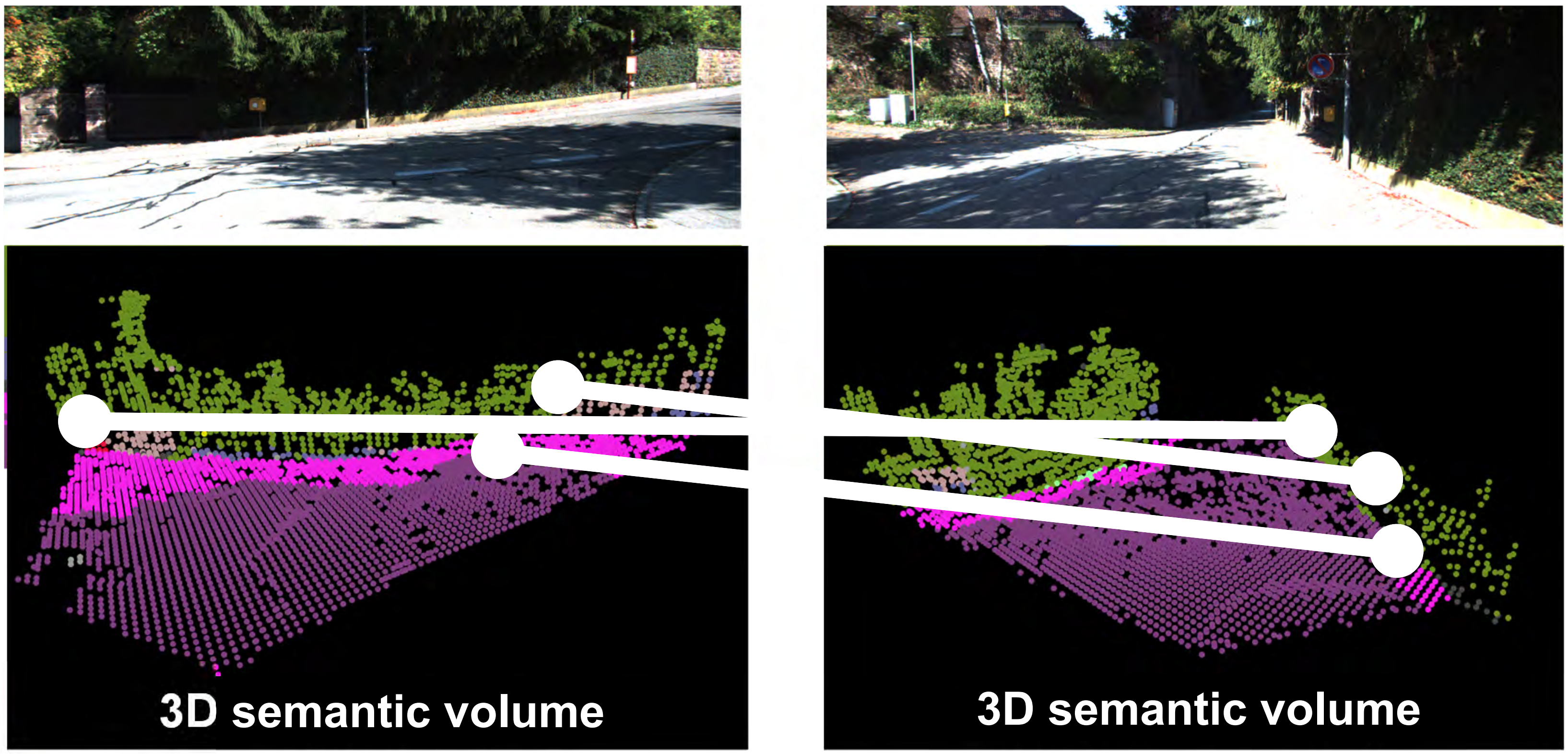
Compute exact position and orientation of query image.

Visual Localization



[Bernhard Zeisl, Torsten Sattler and Marc Pollefeys.
Camera Pose Voting for Large-Scale Image-Based Localization. ICCV, 2015]

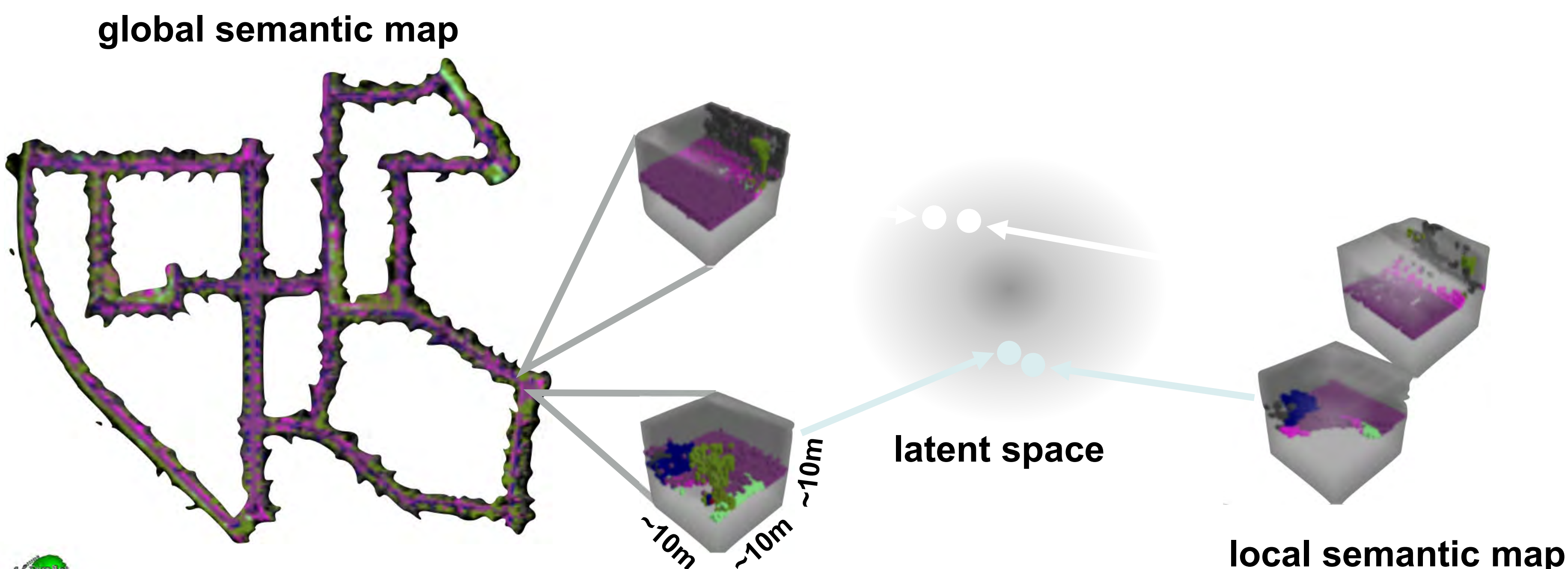
3D Semantic Localization



Horizon 2020
European Union funding
for Research & Innovation

[Schönberger, Pollefeys, Geiger, Sattler, Semantic Visual Localization, CVPR 2018]

3D Semantic Localization



[Schönberger, Pollefeys, Geiger, Sattler, Semantic Visual Localization, CVPR 2018]

Horizon 2020
European Union funding
for Research & Innovation

The 180º Case

Horizon 2020
European Union funding
for Research & Innovation

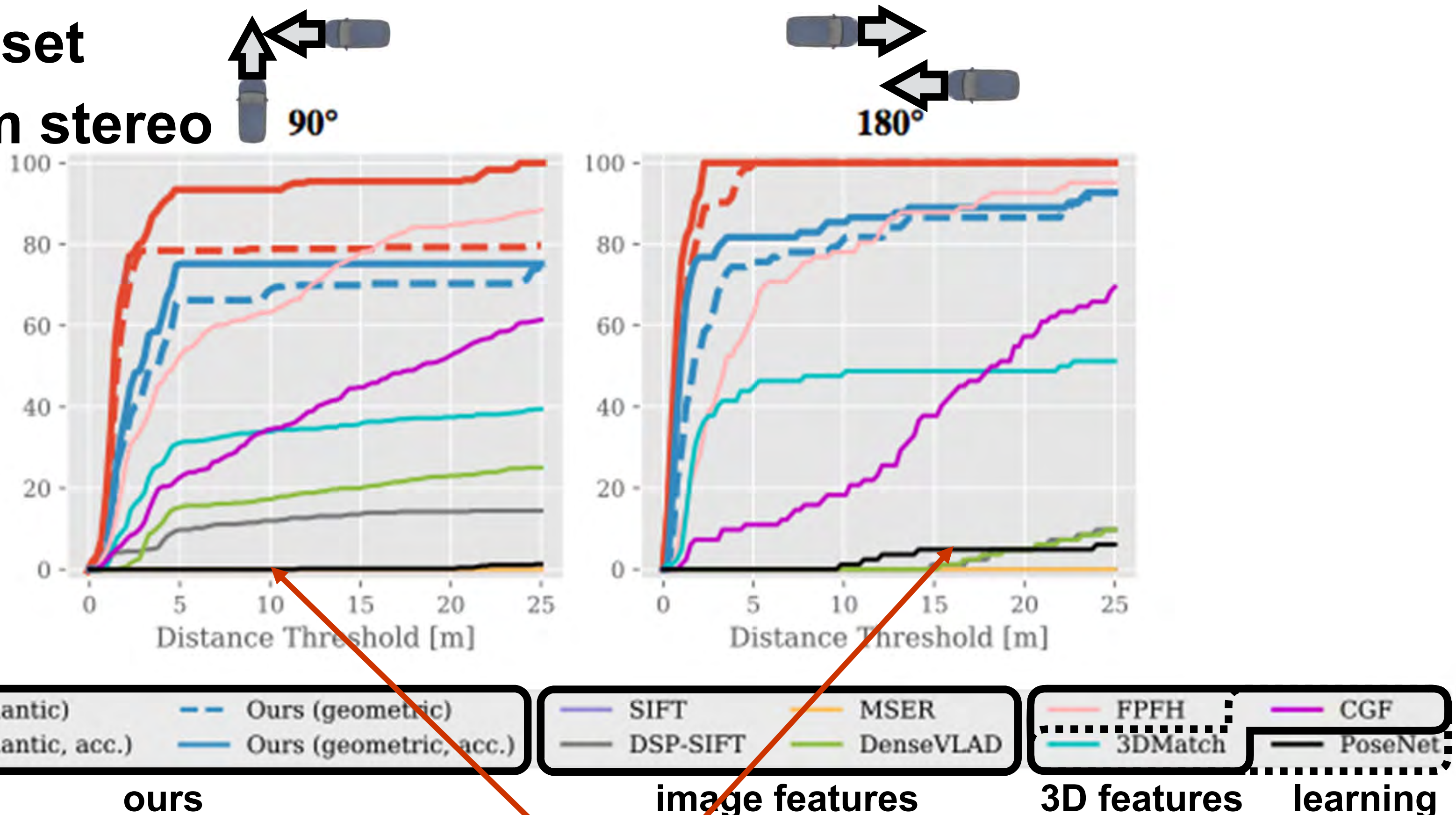
[Schönberger, Pollefeys, Geiger, Sattler, Semantic Visual Localization, CVPR 2018]

ETH Zürich

Computer Vision
and Geometry Lab

Strong Viewpoint Change

KITTI dataset
depth from stereo



[Schönberger, Pollefeys, Geiger, Sattler, Semantic Visual Localization, CVPR 2018]

Understanding the Limitations of CNN-based Absolute Camera Pose Regression, Sattler et al CVPR 2019

Questions?