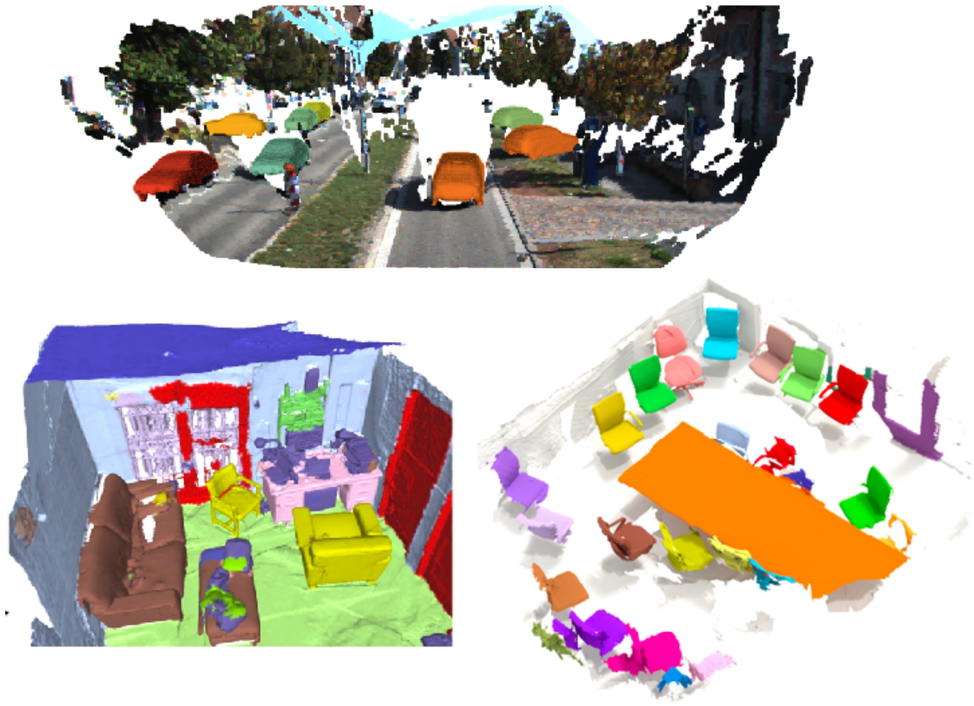
3D Scene Understanding



Francis Engelmann

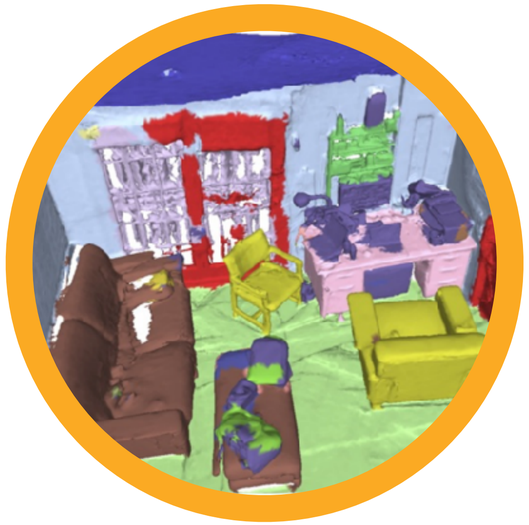
RWTH Aachen University

Visual Computing Institute

December 2020

Overview

Shape and Pose Reconstruction



3D Semantic Segmentation

3D Instance Segmentation

Joint Object Pose Estimation and Shape Reconstruction

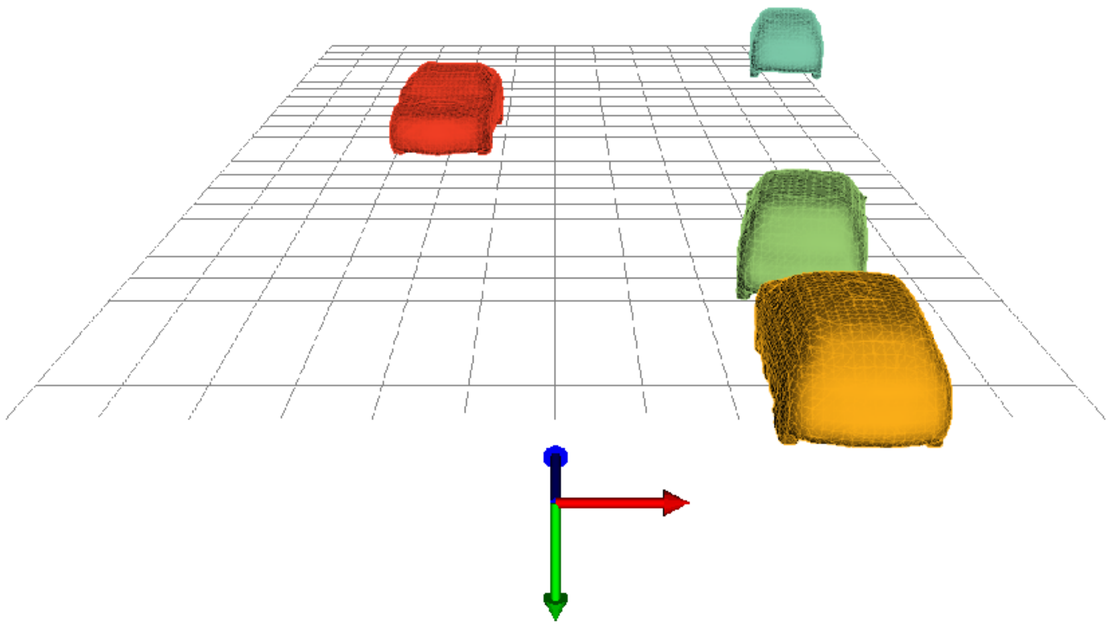
Input

Stereo Image Pair

Left Camera Right Camera

Output

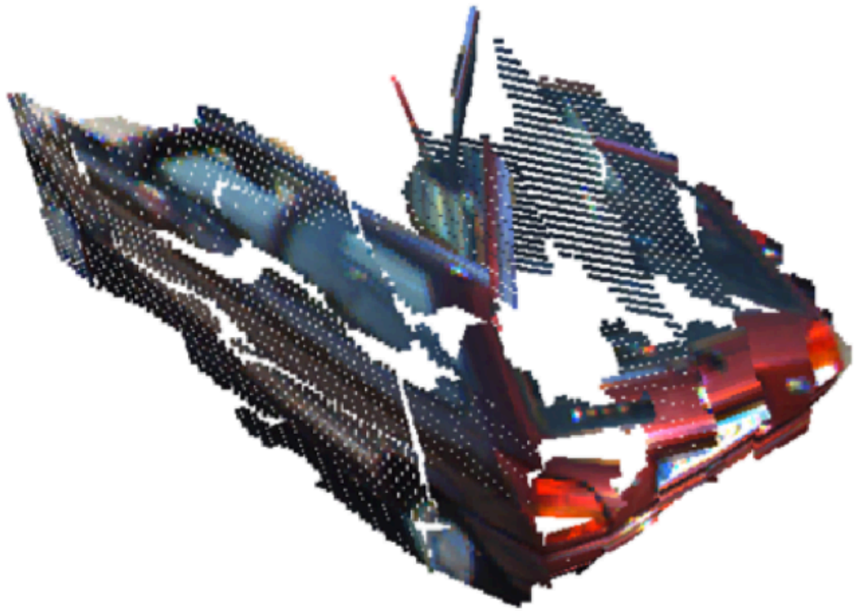
3D Shape and 3D Pose



Joint Object Pose Estimation and Shape Reconstruction in Urban Street Scenes Using 3D Shape Priors
Engelmann Francis, Stückler Jörg, Leibe Bastian, GCPR'16

Shape Completion

Input



Noisy Stereo Depth

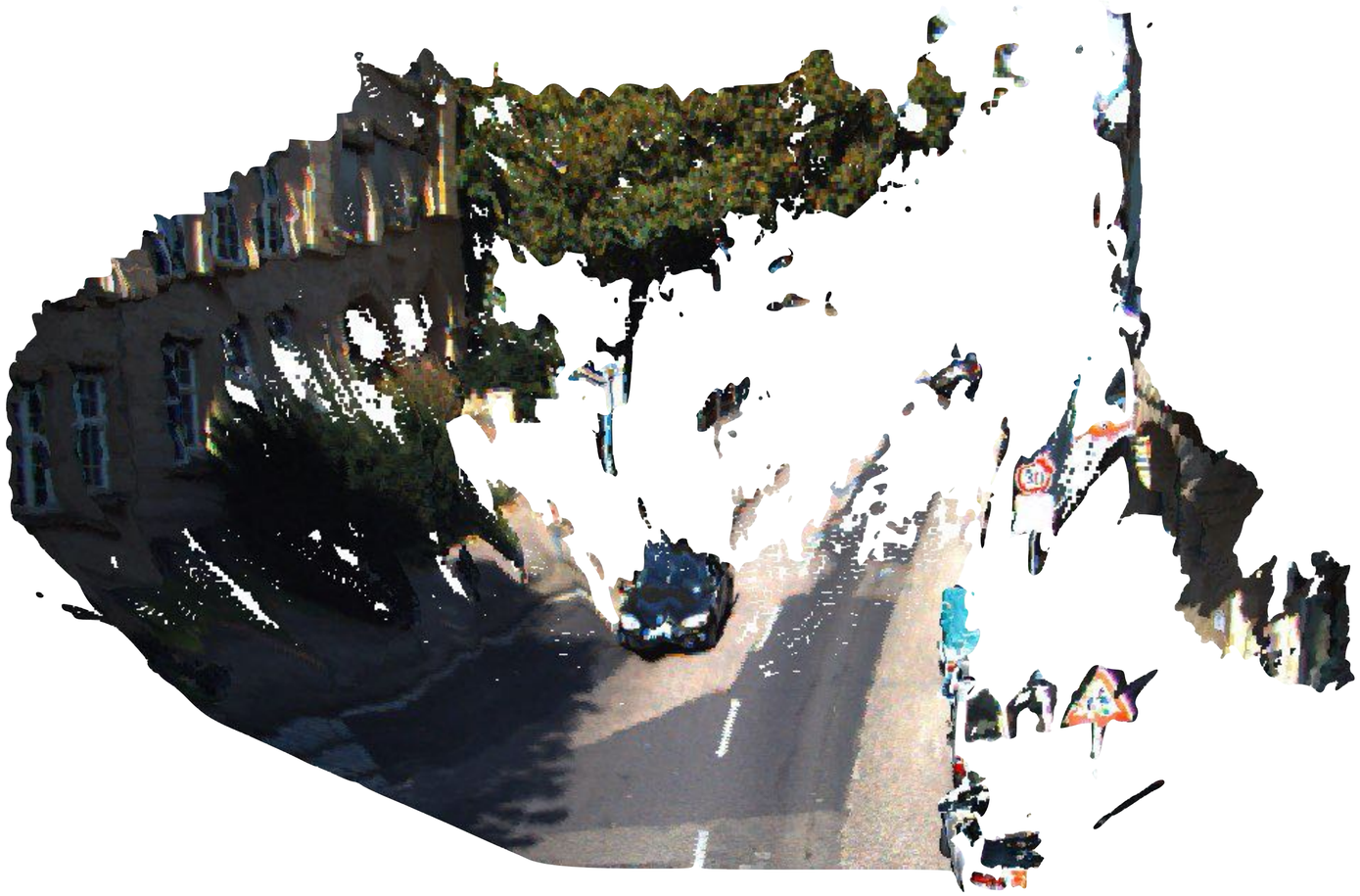
Output

Shape Completion

Joint Object Pose Estimation and Shape Reconstruction

Input: 3D Stereo Reconstruction

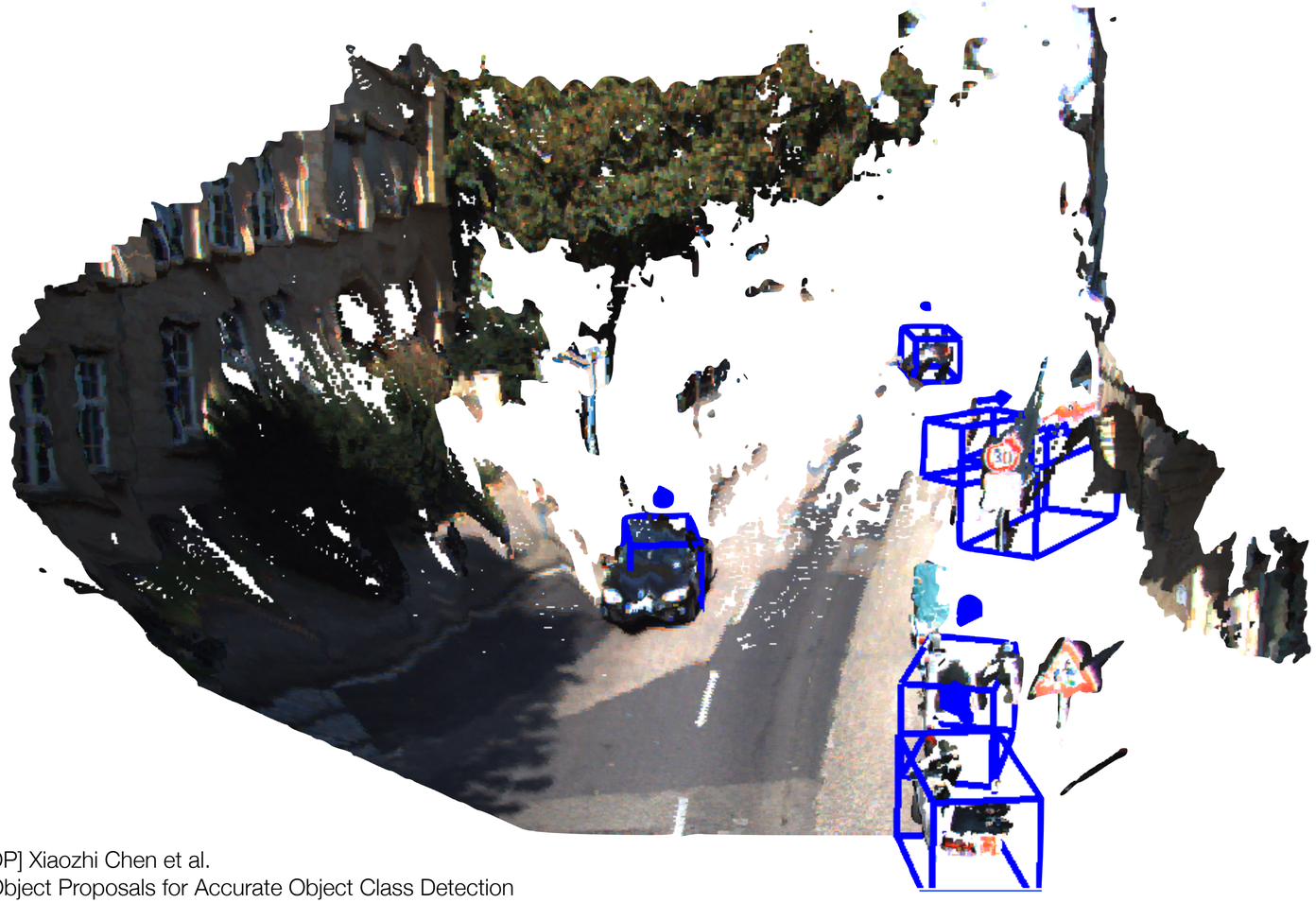
3D Stereo Reconstruction



ELAS (Geiger et al.)
SPS-Stereo (Yamaguchi et al.)

Joint Object Pose Estimation and Shape Reconstruction

Input: Car detections from 3DOP



3D Stereo Reconstruction

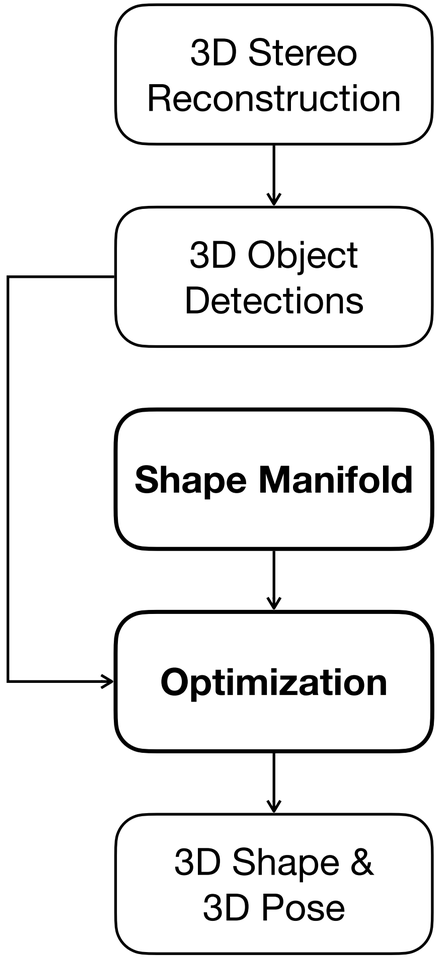
↓

3D Object Detections

[3DOP] Xiaozhi Chen et al.
3D Object Proposals for Accurate Object Class Detection
Neural Information Processing Systems (NIPS), 2015

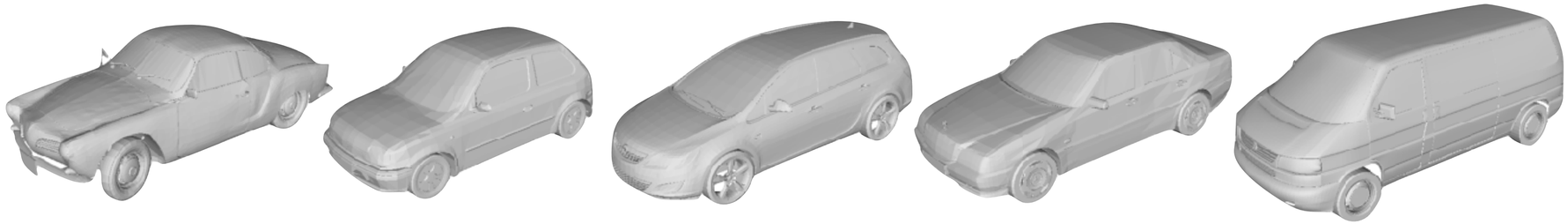
Joint Object Pose Estimation and Shape Reconstruction

Inference: Align shape manifold with noisy measurements.

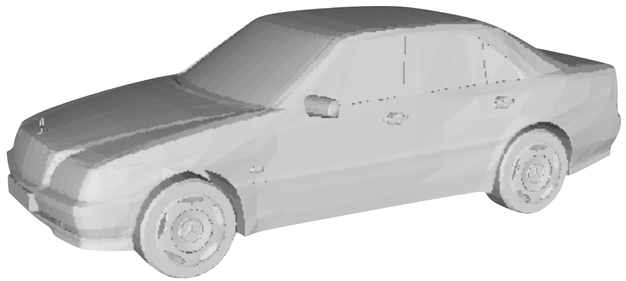


Shape Manifold

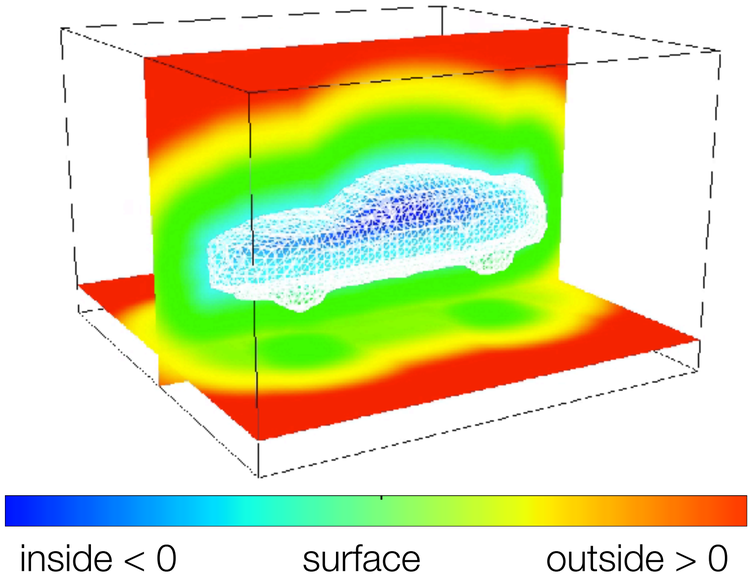
1. Collection of CAD Models



2. **Compute 3D TSDF** (Truncated Signed Distance Function ϕ)

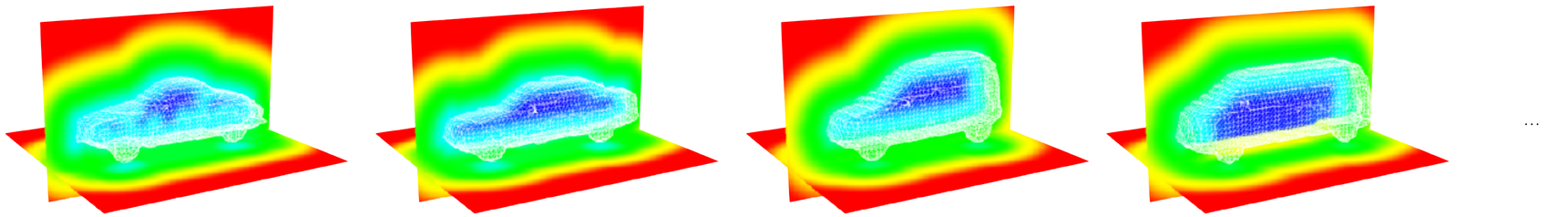


$$\phi : \mathbb{R}^3 \mapsto \mathbb{R}$$



Shape Manifold

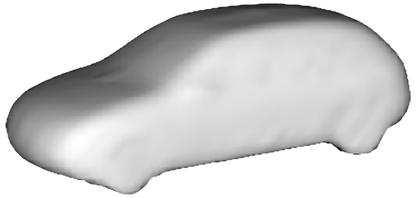
Linear Subspace



perform PCA

Reconstruct

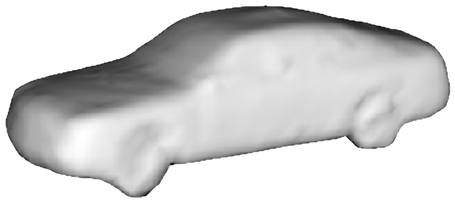
$$\mu + \sum_i z_i \cdot \mathrm{PC}_i = y$$



μ: mean shape

$\mathbf{z}$: shape parameters

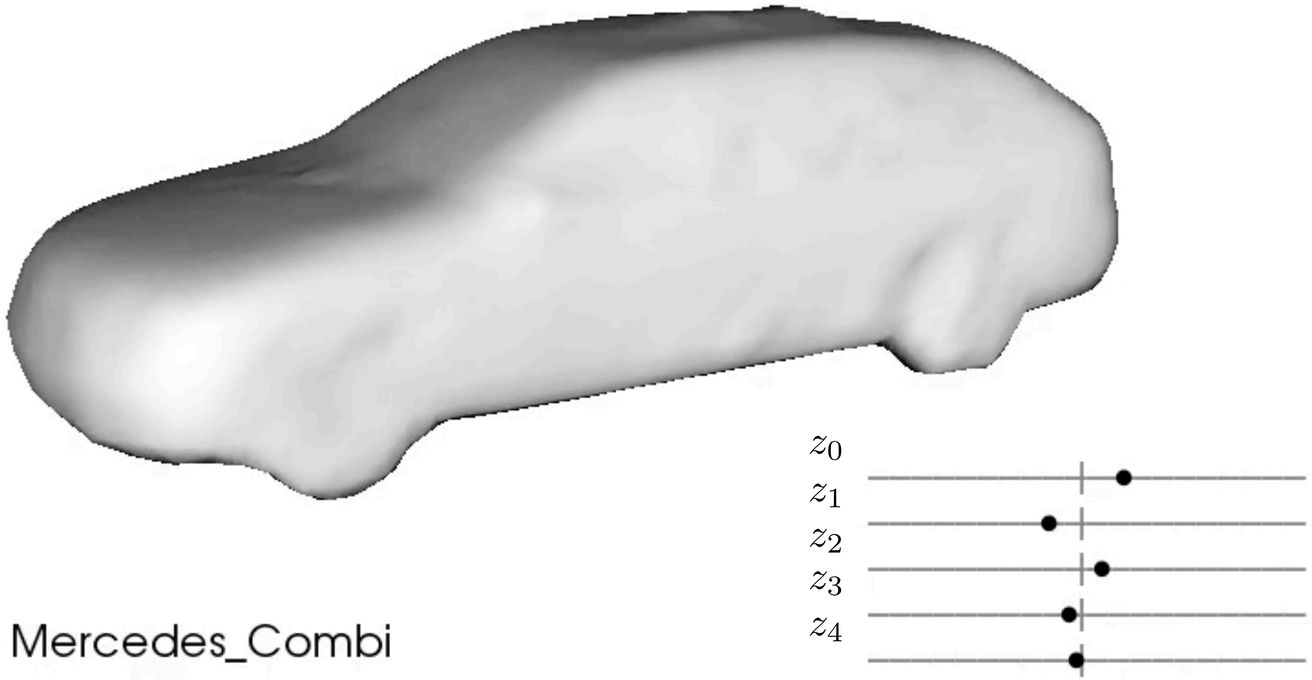
$\mathbf{PC}$: principal components



y: example reconstruction

Shape Manifold

Reconstructed Models (first 5 principal components, $\mathbf{z} \in \mathbb{R}^5$)

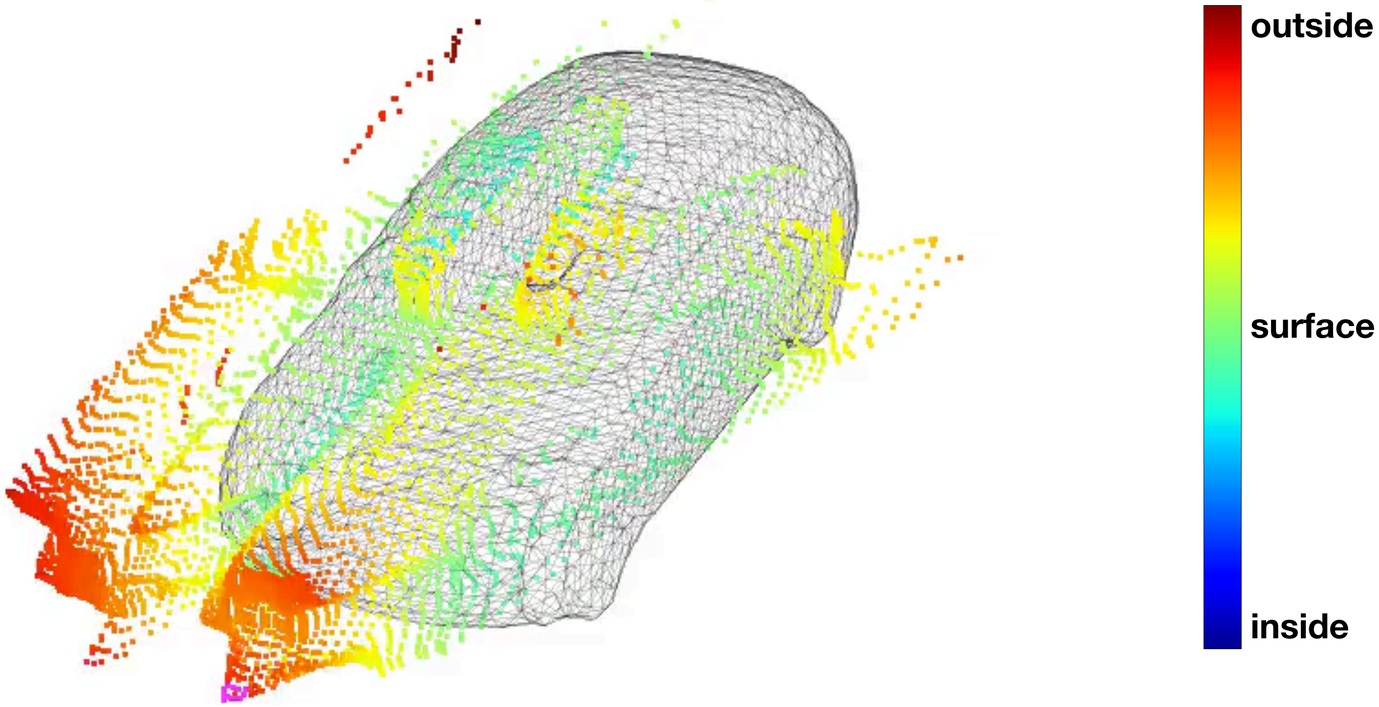


10

Inference

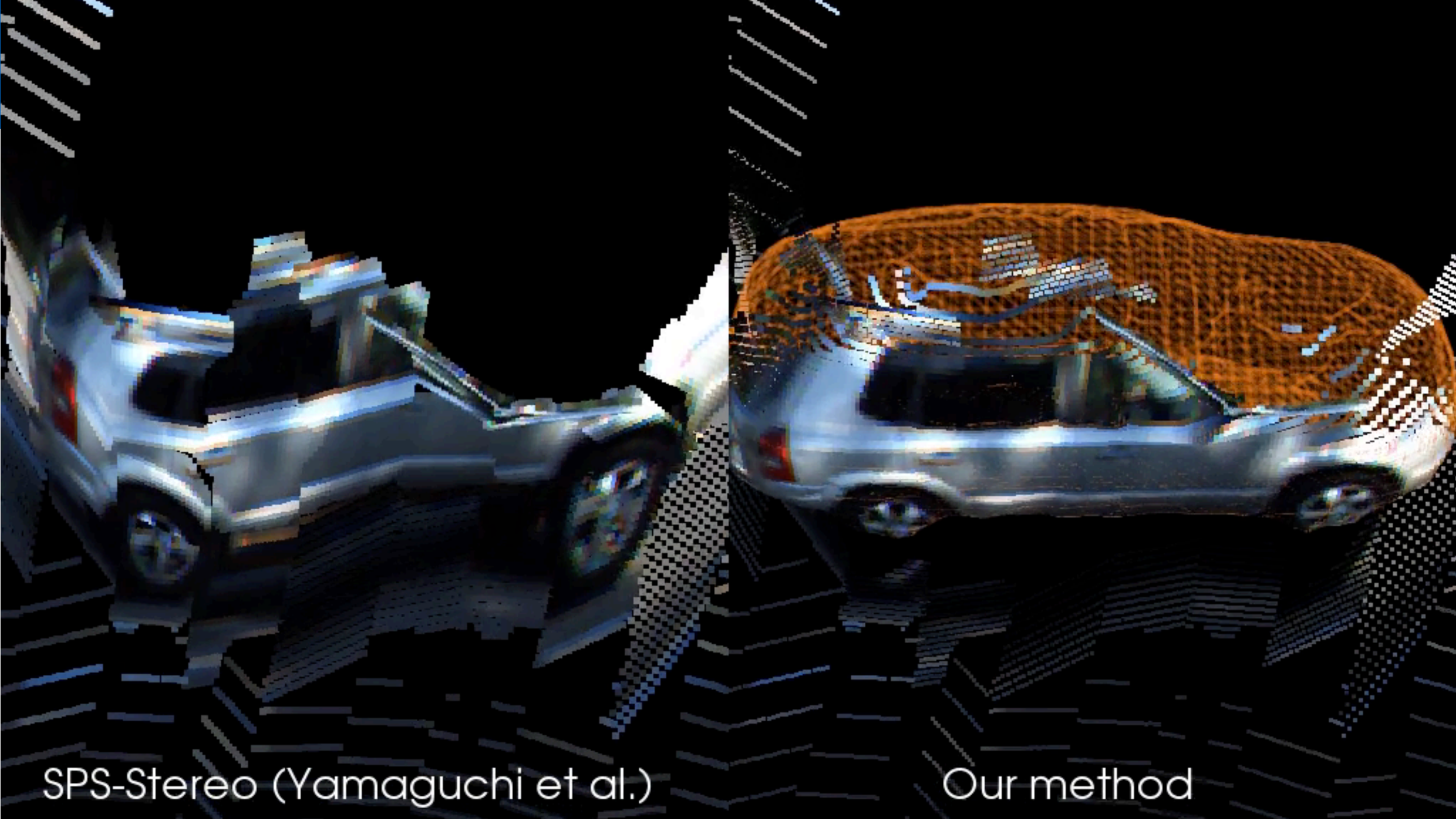
Key Idea: Surface points have SDF of zero.

$$\mathrm{E}(\mathbf{z}, \boldsymbol{\xi} \mid \mathrm{X}) = \frac{1}{N} \sum_{i}^{N} \frac{1}{\sigma_d^2} \rho\Big(\phi_{\mathbf{z}}(T_{\boldsymbol{\xi}}^{-1} \mathbf{x}_i)\Big) + \sum_{j}^{K} \left(\frac{z_j}{\sigma_j}\right)^2 + \left(\frac{\mathrm{t}_y - g(\mathbf{t})}{\sigma_g}\right)^2$$



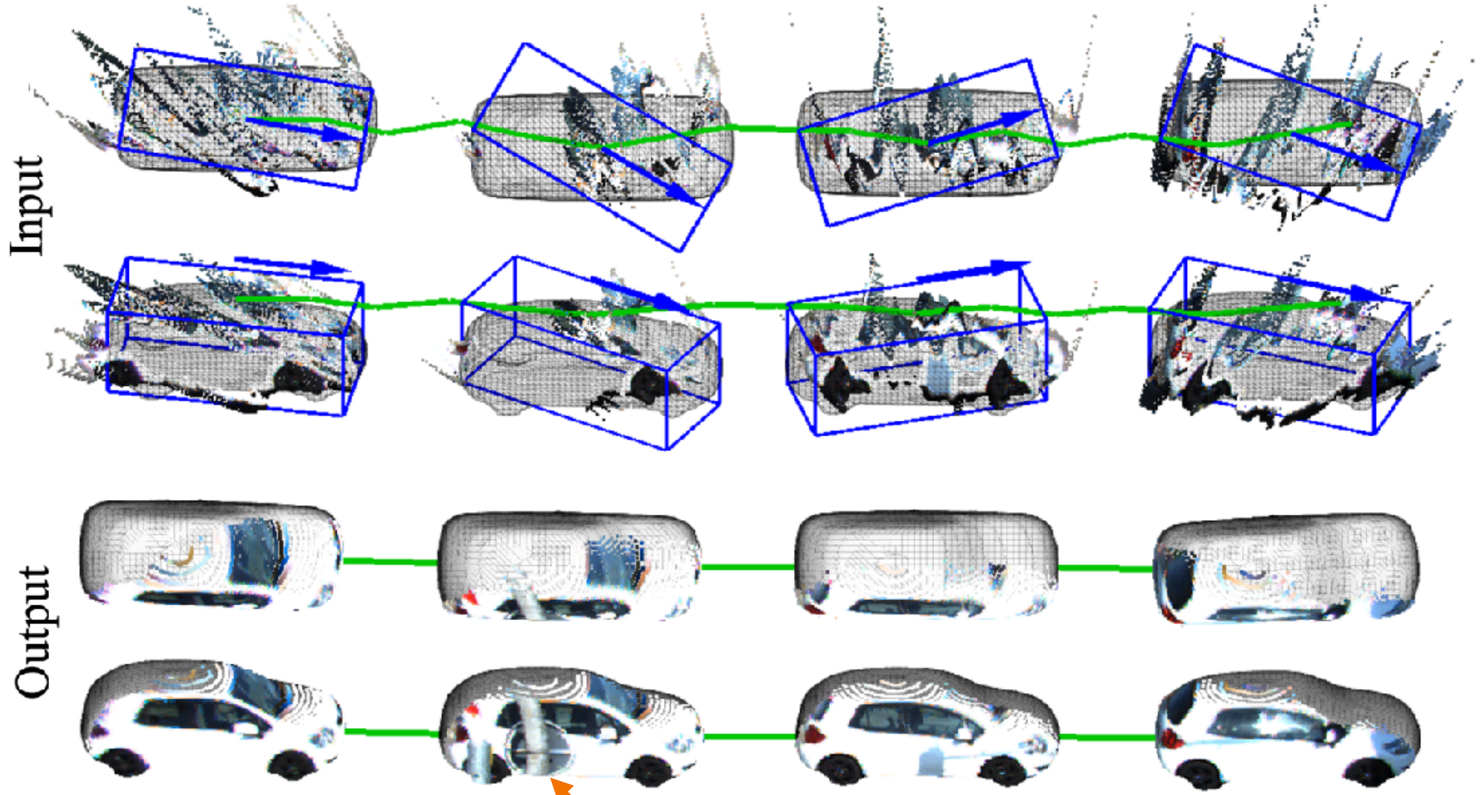
libELAS (Geiger et al.)

Our method


SPS-Stereo (Yamaguchi et al.)
Our method

Extension: Temporal Data

Motivation:
- Additional observations over time
- Handle (short) occlusions
- Motion is a strong prior for pose

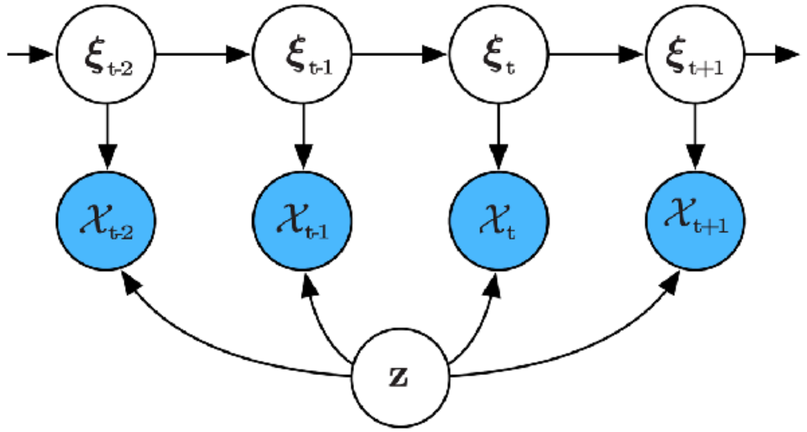


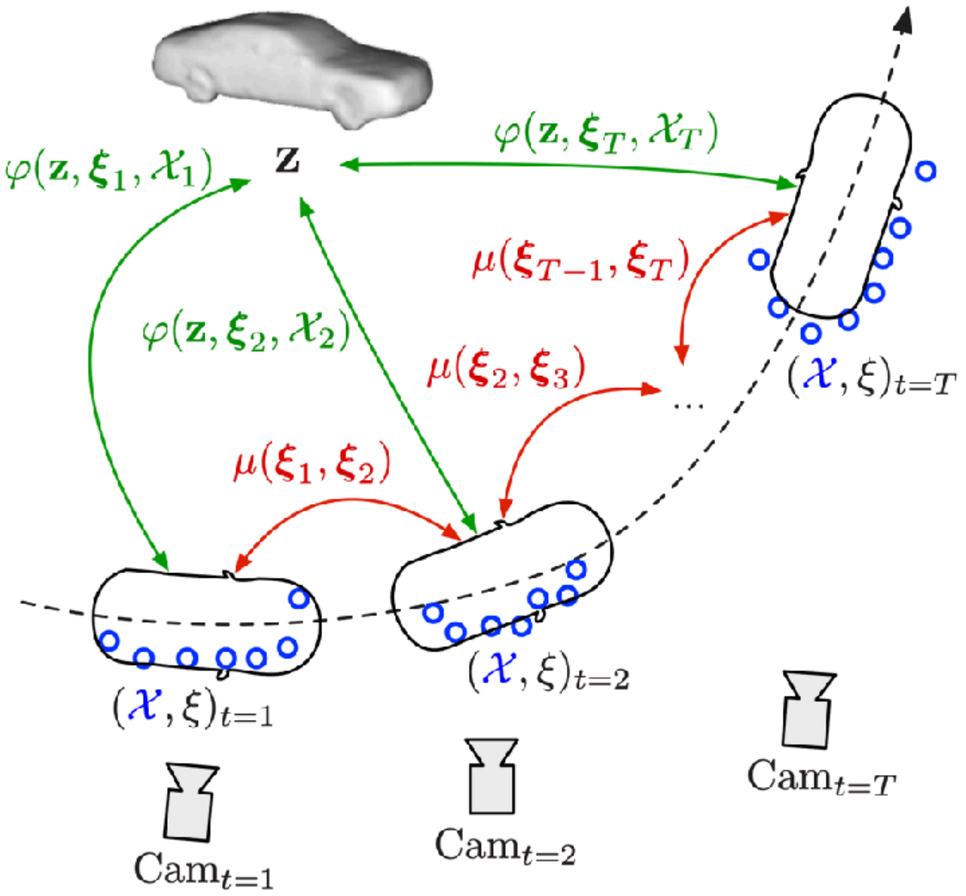
occlusion

SAMP: Shape and Motion Priors for 4D Vehicle Reconstruction
Engelmann Francis, Stückler Jörg, Leibe Bastian, WACV'17

Extension: Temporal Data

Graphical Model



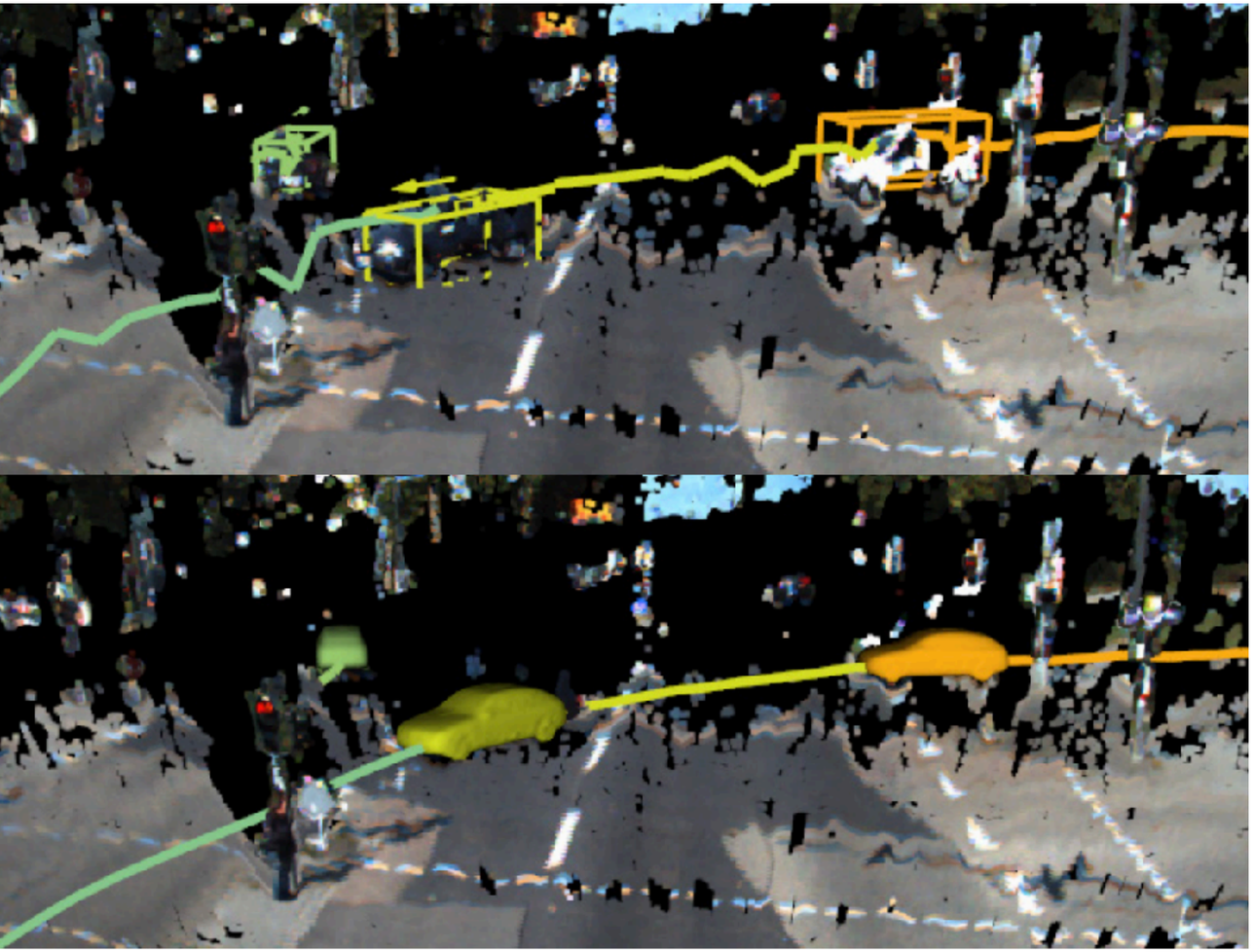


$$\mathrm{E}(\mathbf{z}, \boldsymbol{\xi}) = \frac{1}{T} \sum_t [\underbrace{\varphi(\mathbf{z}, \boldsymbol{\xi}_t, \mathcal{X}_t)}_{\text{data term}} + \underbrace{\mu(\boldsymbol{\xi}_t, \boldsymbol{\xi}_{t-1})}_{\text{motion term}}] + \underbrace{\kappa(\mathbf{z})}_{\text{shape regul.}}$$

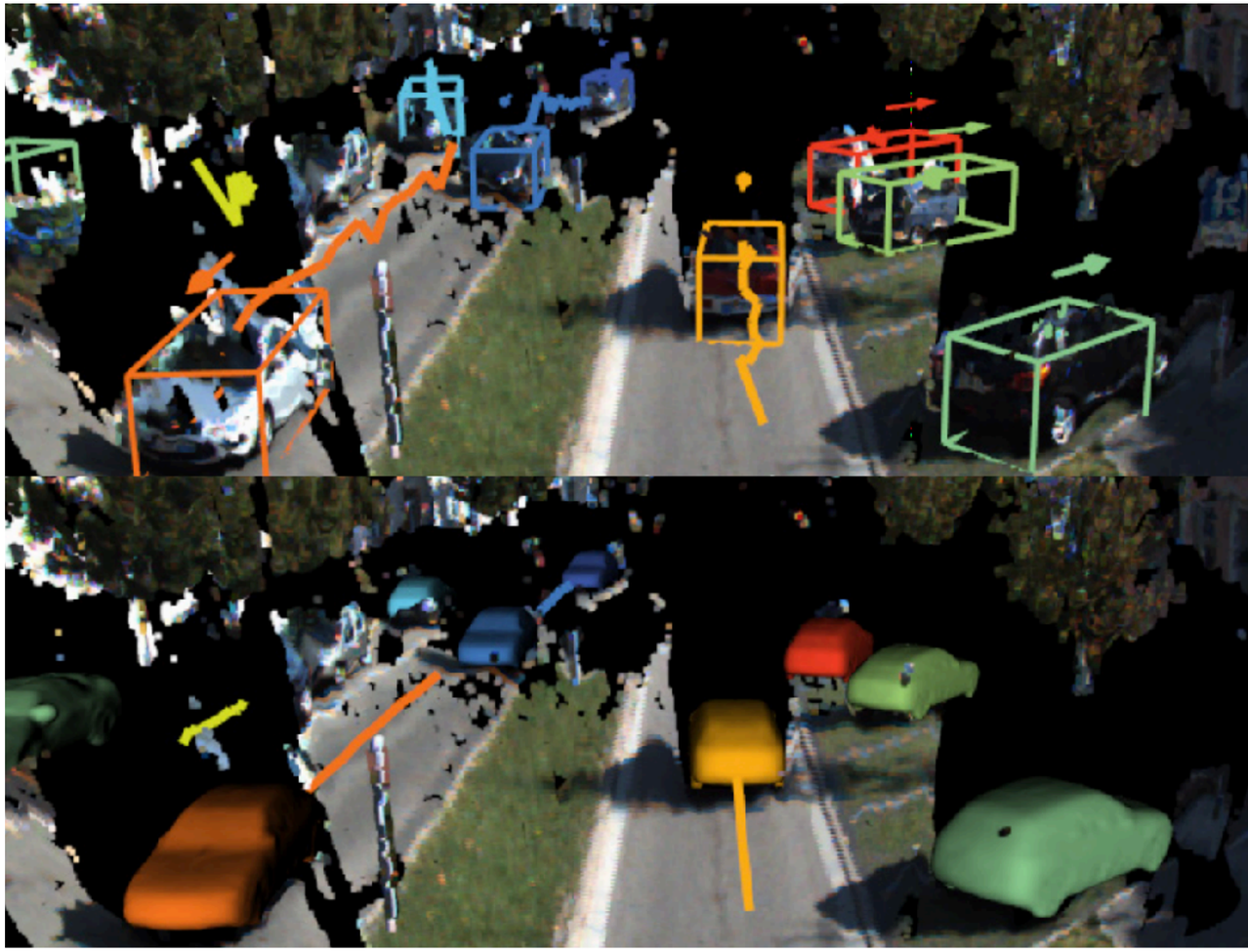
SAMP: Shape and Motion Priors for 4D Vehicle Reconstruction
Engelmann Francis, Stückler Jörg, Leibe Bastian, WACV'17

Qualitative Results

Input Sequence 1



Input Sequence 2



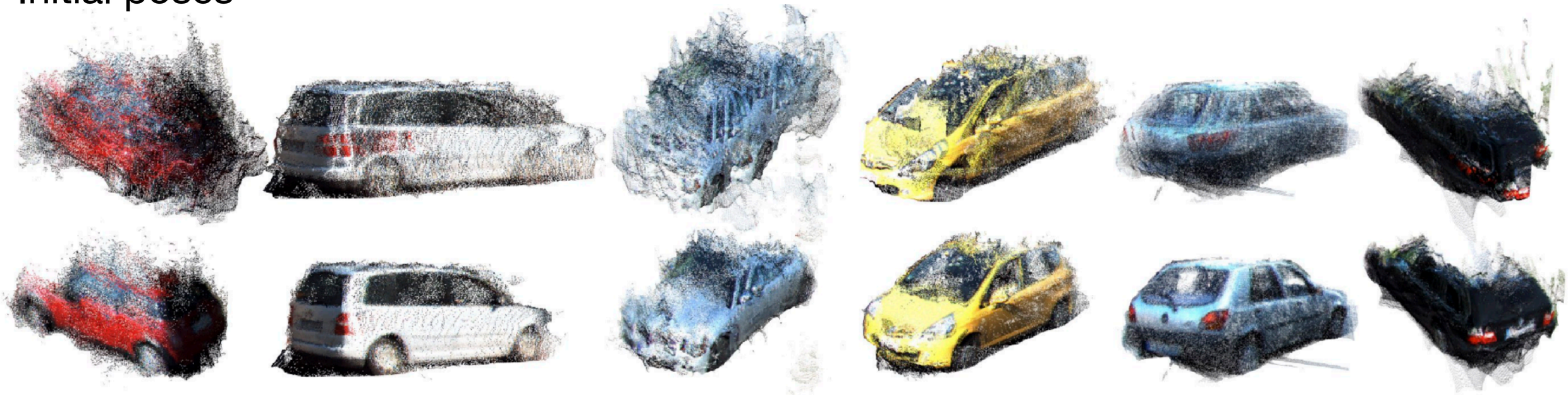
Output Sequence 1

Output Sequence 2

Qualitative Results

Multiple measurements overlaid (Canonical Coordinate System)

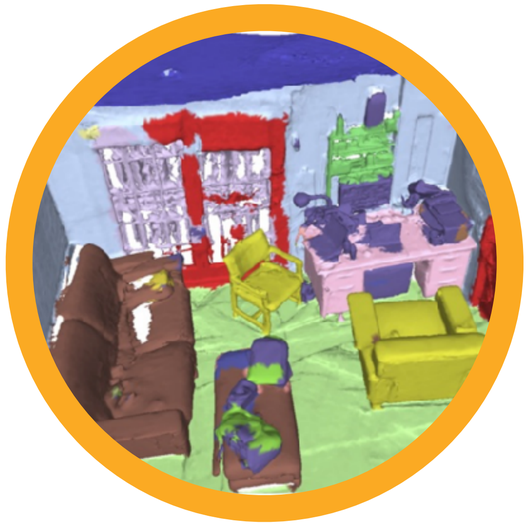
Initial poses



Optimized poses

Overview

Shape and Pose Reconstruction

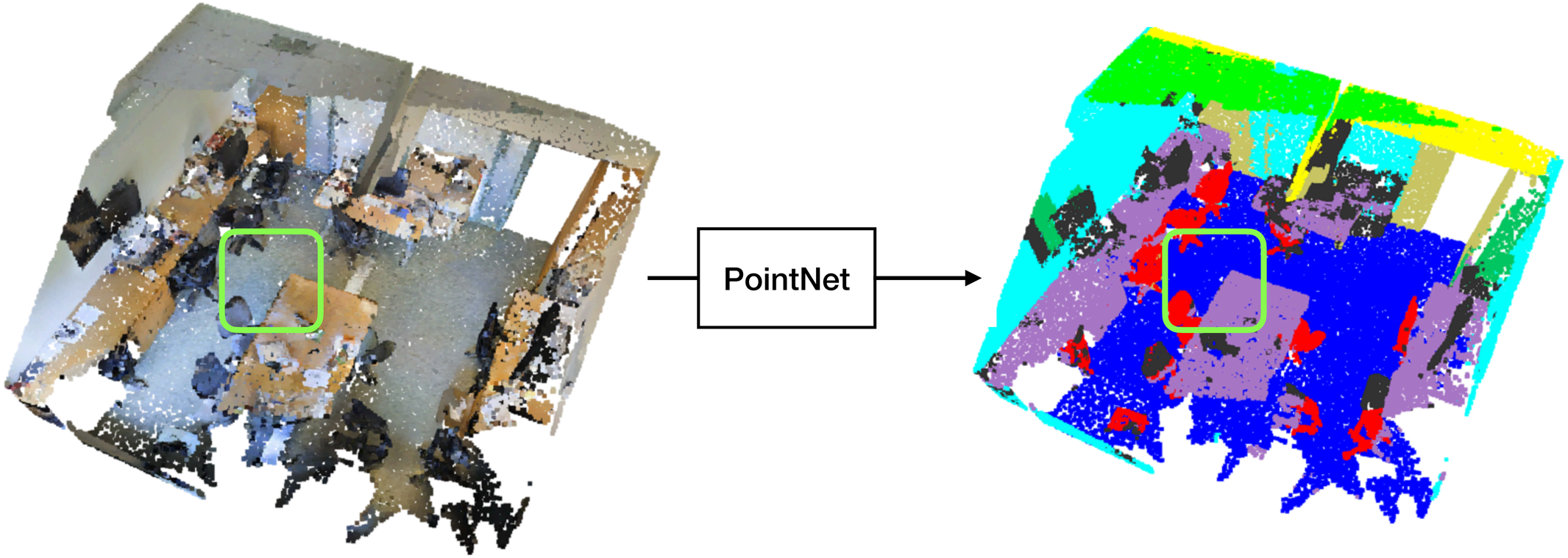


3D Semantic Segmentation

3D Instance Segmentation

3D Semantic Segmentation

Problem: Limited Context

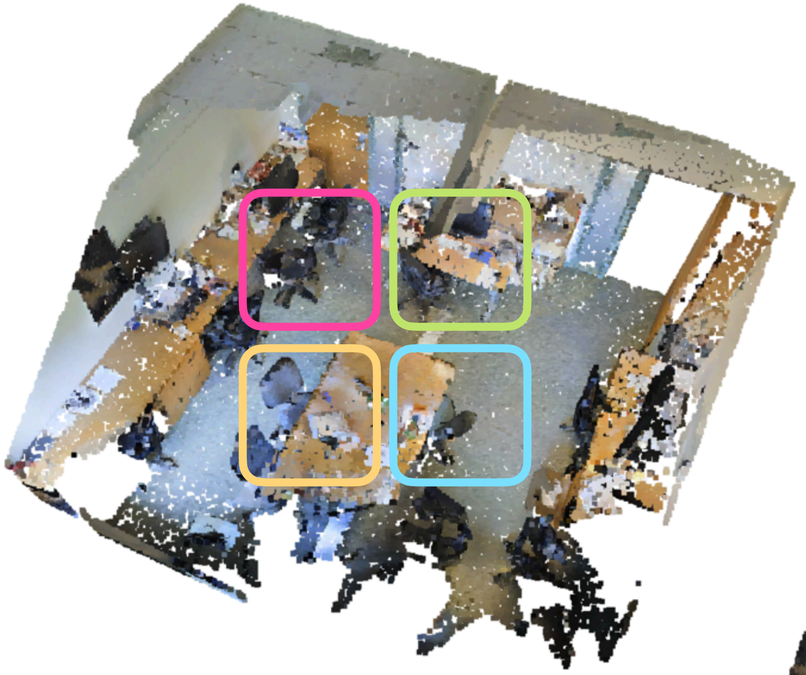


Input: Point Cloud

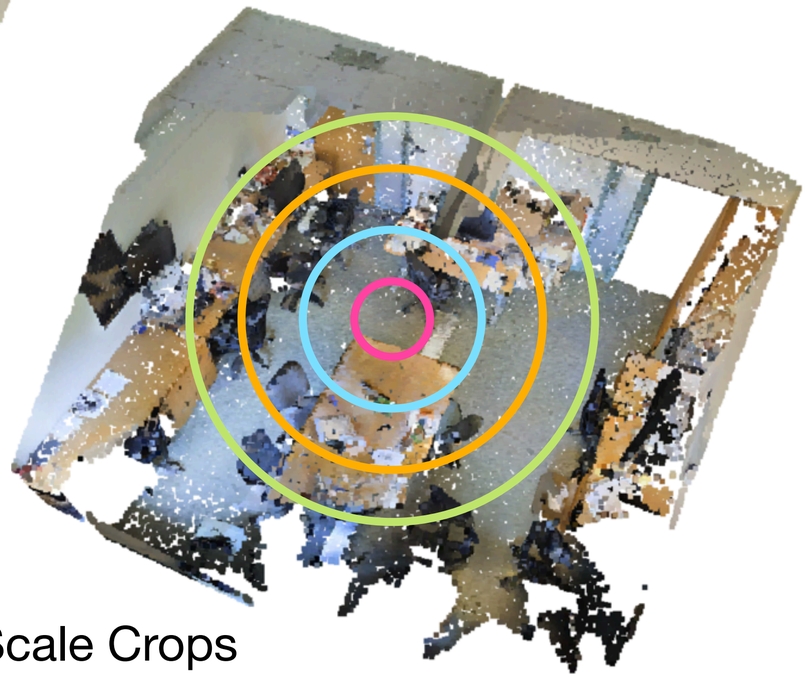
Output: Semantic Labels

Exploring Spatial Context for 3D Semantic Segmentation of Point Clouds
Engelmann Francis, Kontogianni Theodora, Alexander Hermans, Leibe Bastian, ICCV'W'17

Exploring Spatial Context for 3D Semantic Segmentation

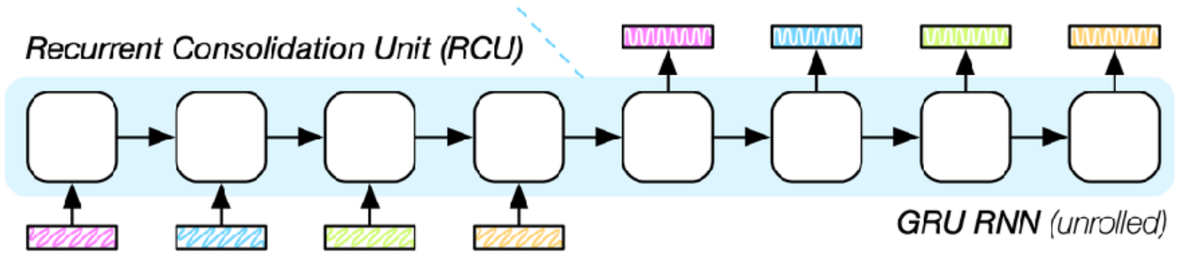


Grid Neighborhood

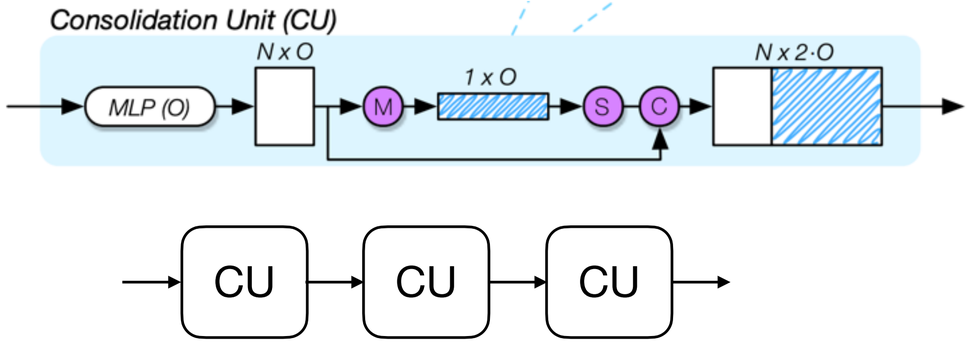


Multi-Scale Crops

Aggregation using RNN



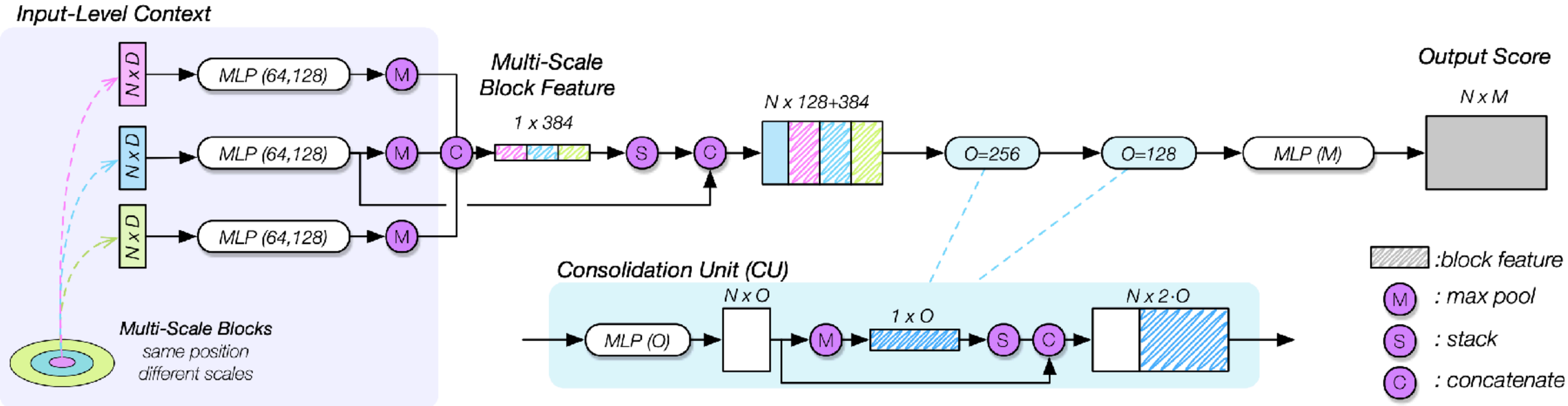
Stacking PointNets



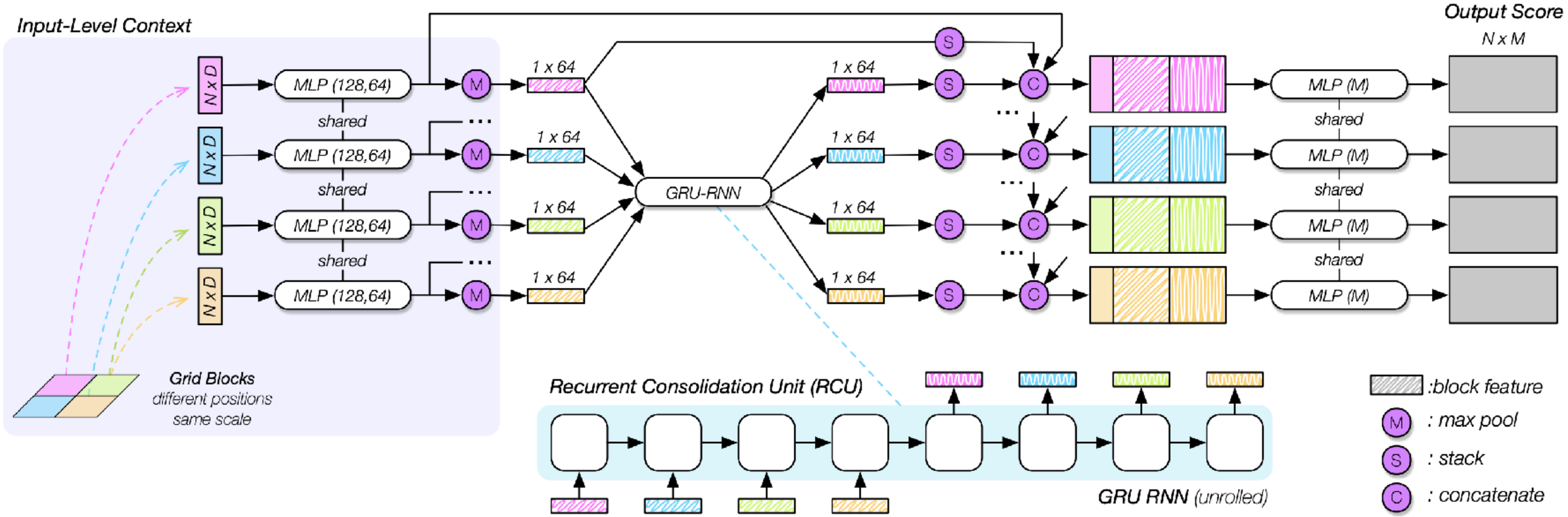
Consolidation Unit (CU) = mini PointNet

Exploring Spatial Context for 3D Semantic Segmentation of Point Clouds
Engelmann Francis, Kontogianni Theodora, Alexander Hermans, Leibe Bastian, ICCV'W'17

Multi-Scale / Stacked PointNets Architecture



Grid Blocks / Recurrent Consolidation Unit



Experiments

Comparison to PointNet

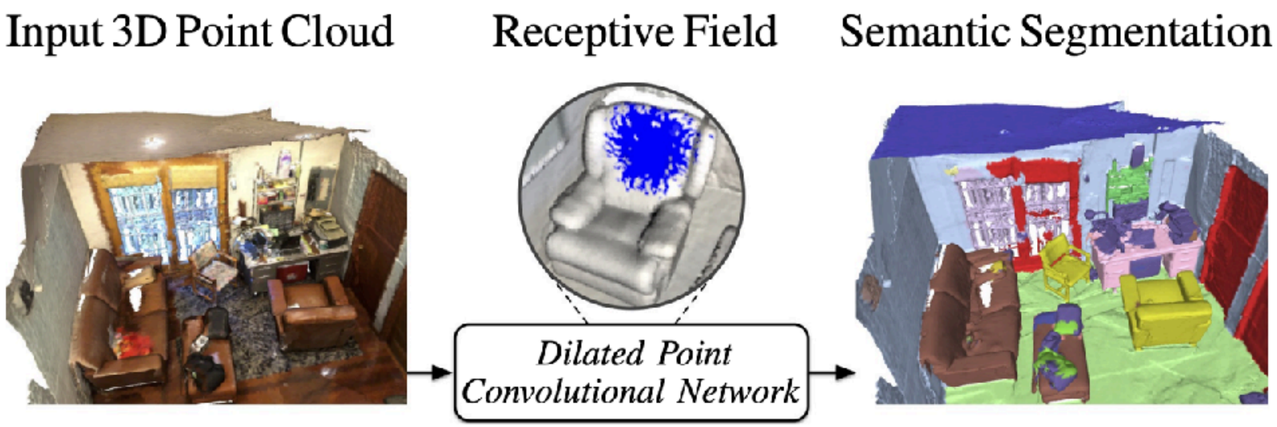
S3DIS Dataset [1] XYZ-RGB	mean IoU	overall accuracy	avg. class accuracy
*PointNet [26]	43.5	75.0	55.5
*MS	44.4	75.5	57.6
*MS + RCU	45.5	77.2	57.2
*SS + CU(1)	45.9	77.8	57.7
*MS + CU(2)	**47.8**	**79.2**	**59.7**
PointNet [26]	47.6	78.5	66.2
G + RCU	**49.7**	**81.1**	**66.4**

Ablation Study

S3DIS Dataset [1]	mean IoU	*Ceiling*	*Floor*	*Wall*	*Beam*	*Column*	*Window*	*Door*	*Table*	*Chair*	*Sofa*	*Bookcase*	*Board*	*Clutter*
*PointNet [26]	43.5	81.5	86.7	64.8	29.4	16.3	39.1	48.1	52.5	42.5	5.4	37.6	30.4	31.4
*MS	44.4	82.2	86.9	64.2	33.8	22.8	43.3	52.0	51.0	38.6	9.2	36.1	23.6	33.7
*MS + RCU	45.5	83.6	86.9	**67.5**	**40.5**	17.1	37.0	48.8	**53.9**	42.3	6.8	**39.7**	**32.8**	34.2
*SS + CU(1)	45.9	**88.6**	92.6	66.3	36.2	23.6	47.1	51.2	50.2	36.9	**12.6**	33.7	22.7	35.3
*MS + CU(2)	**47.8**	**88.6**	**95.8**	67.3	36.9	**24.9**	**48.6**	**52.3**	51.9	**45.1**	10.6	36.8	24.7	**37.5**
PointNet [26]	47.6	88.0	88.7	**69.3**	42.4	23.1	47.5	**51.6**	54.1	42.0	**9.6**	38.2	29.4	35.2
G + RCU	**49.7**	**90.3**	**92.1**	67.9	**44.7**	**24.2**	**52.3**	51.2	**58.1**	**47.4**	6.9	**39.0**	**30.0**	**41.9**

Dilated Point Convolutions

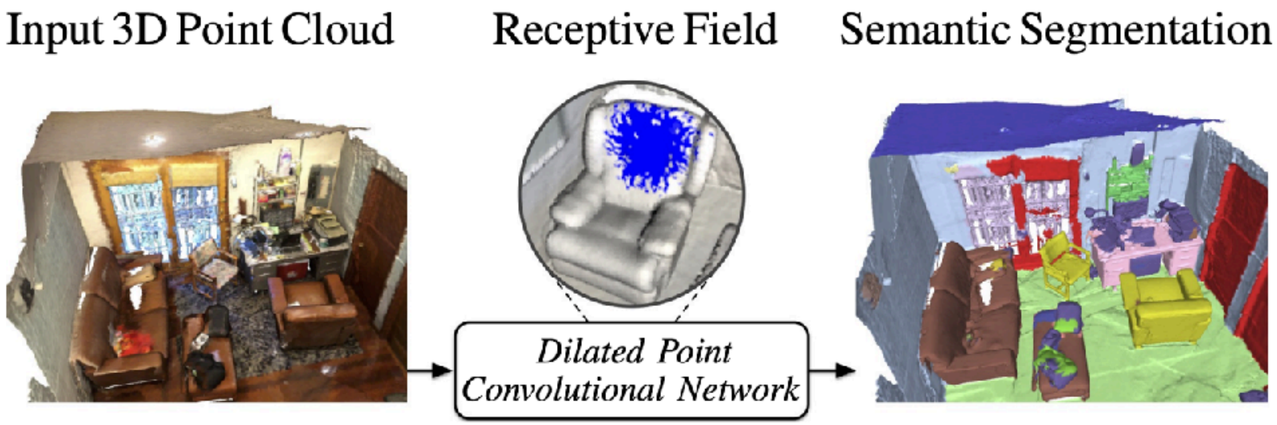
Point Convolutions



Dilated Point Convolutions: On the Receptive Field Size of Point Convolutions on 3D Point Clouds
Engelmann Francis, Kontogianni Theodora, Leibe Bastian, ICRA'20

Dilated Point Convolutions

Point Convolutions



General Convolution

kernel function

$$(f * g)(p_i) = \int_{-\infty}^{+\infty} f(p_j) \odot g(p_i - p_j)\, dp_j$$

$$(f * g)(p_i) \approx \frac{1}{N} \sum_{n=1}^{N} f(p_n) \odot g(p_i - p_n)$$

$$(f * g)(p_i) \approx \frac{1}{|\mathcal{N}_i|} \sum_{p_k \in \mathcal{N}_i} f(p_k) \odot g(p_i - p_k)$$

KNN neighborhood

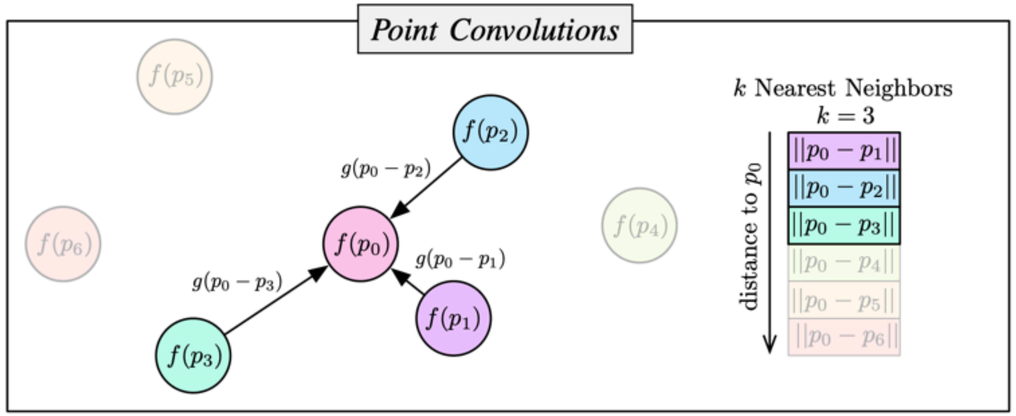
Wang et al. Dynamic Graph CNN for Learning on Point Clouds

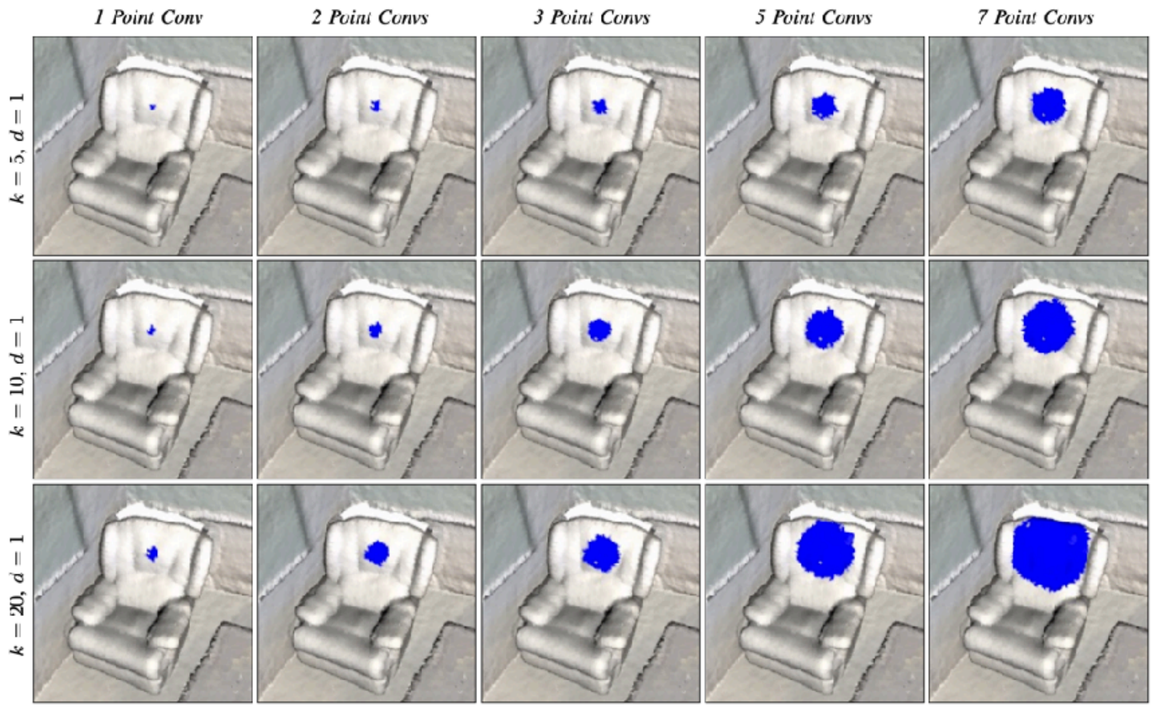
Dilated Point Convolutions: On the Receptive Field Size of Point Convolutions on 3D Point Clouds
Engelmann Francis, Kontogianni Theodora, Leibe Bastian, ICRA'20

Dilated Point Convolutions: Idea

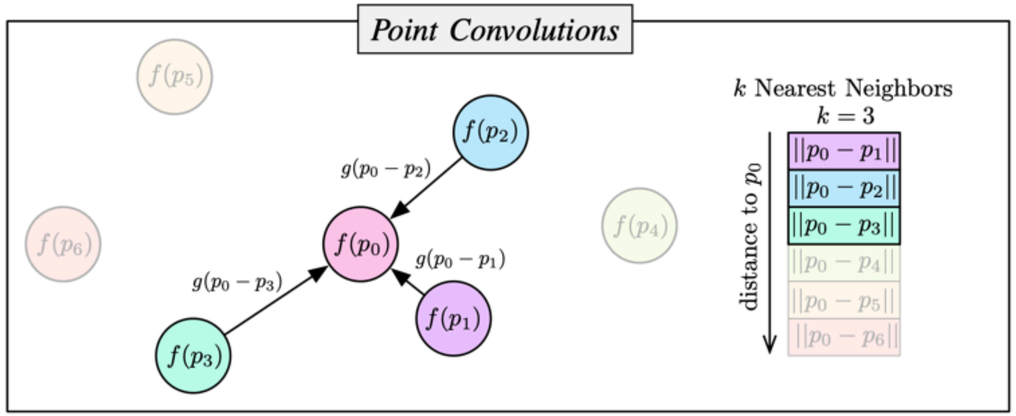
How to increase the receptive field size?

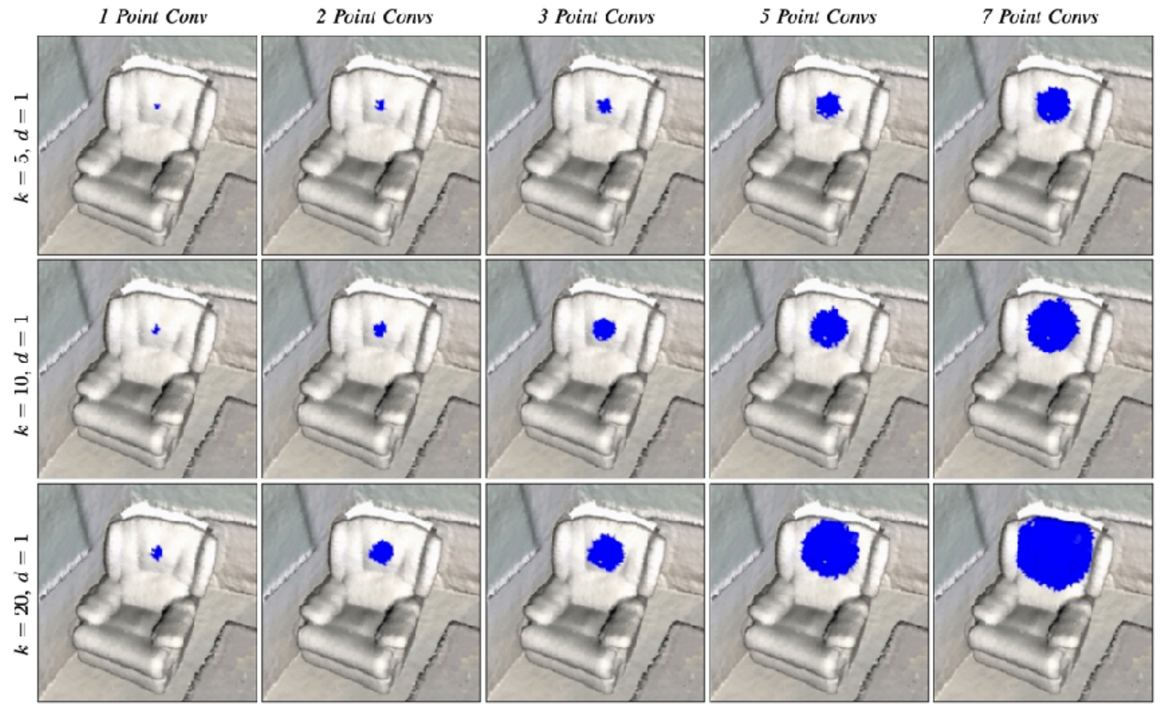
- Larger kernel —> increase k
- Stack more convolutions



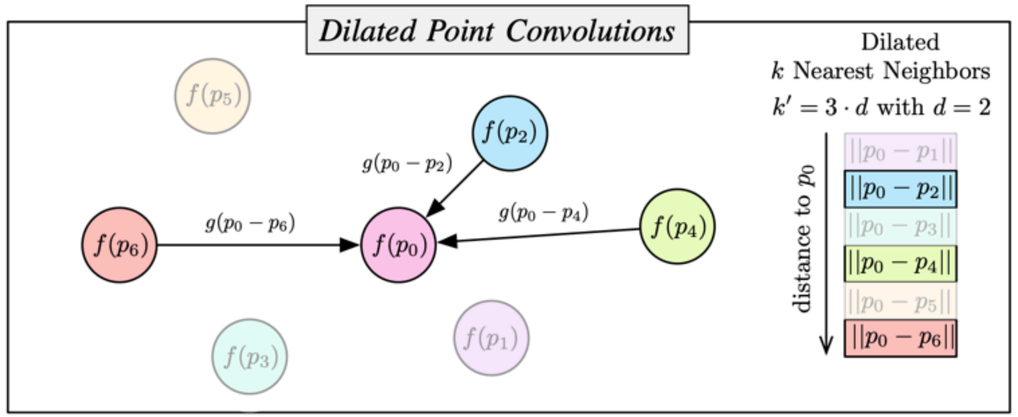


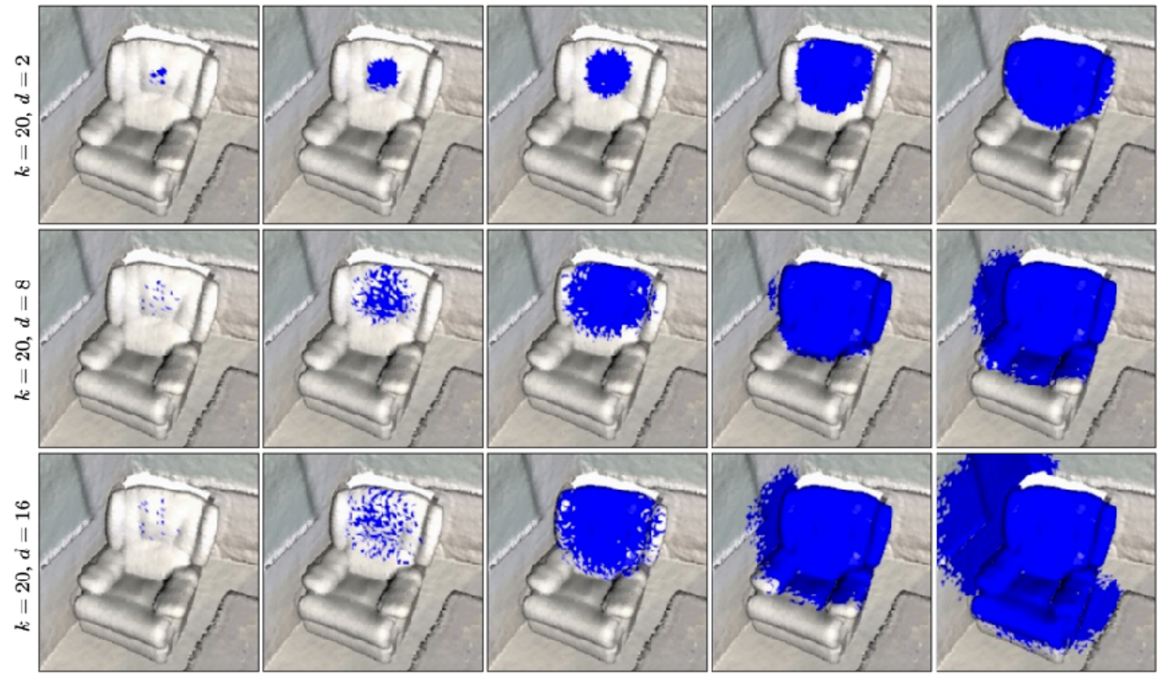
Dilated Point Convolutions: On the Receptive Field Size of Point Convolutions on 3D Point Clouds
Engelmann Francis, Kontogianni Theodora, Leibe Bastian, ICRA'20





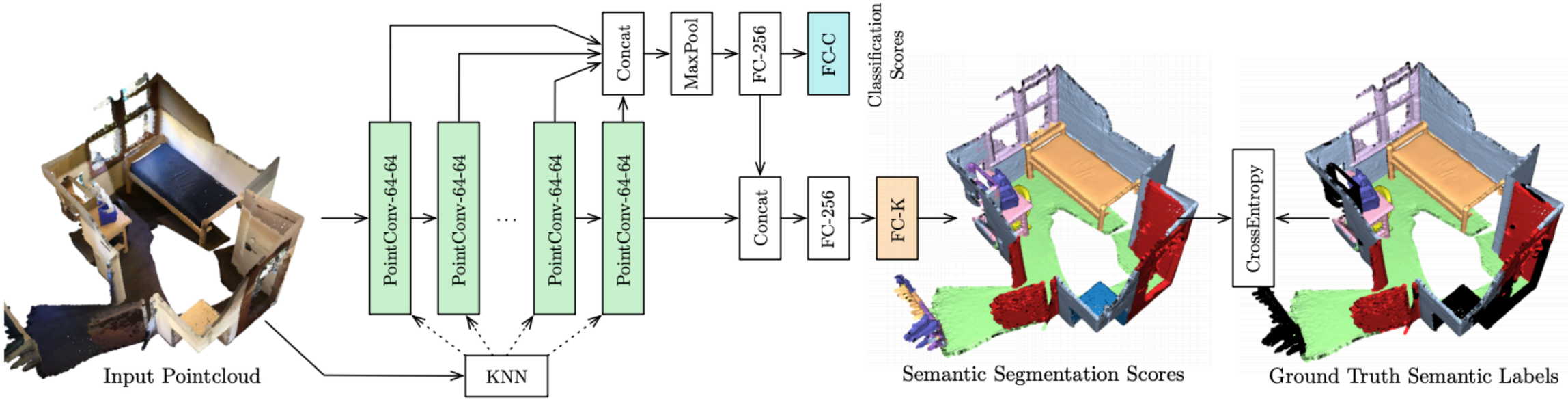
Even better: Dilation!





Dilated Point Convolutions: On the Receptive Field Size of Point Convolutions on 3D Point Clouds
Engelmann Francis, Kontogianni Theodora, Leibe Bastian, ICRA'20

Dilated Point Convolutions: Model

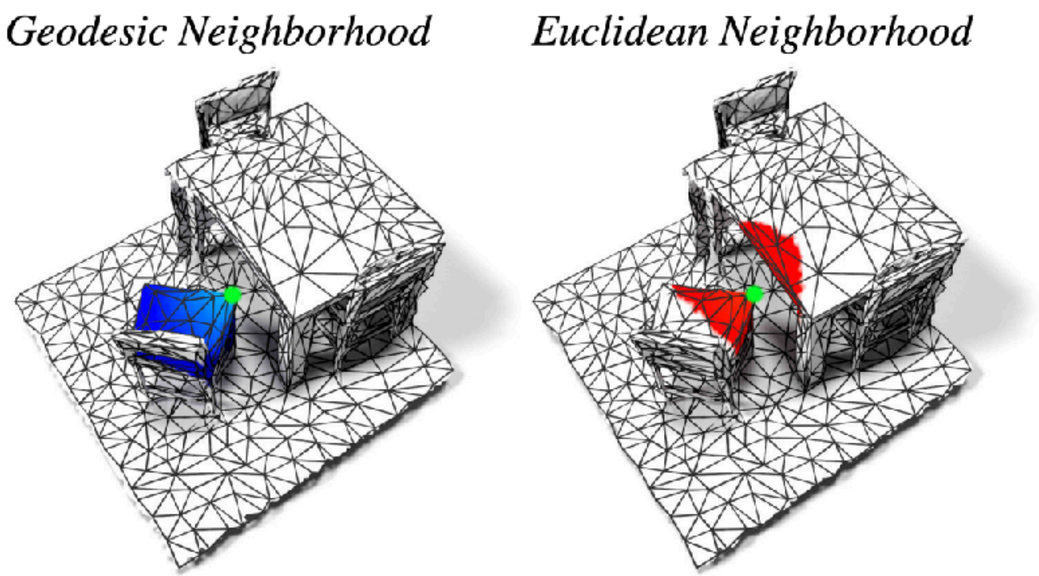


Dilated Point Convolutions: On the Receptive Field Size of Point Convolutions on 3D Point Clouds
Engelmann Francis, Kontogianni Theodora, Leibe Bastian, ICRA'20

DualConvMesh (DCM-Net)

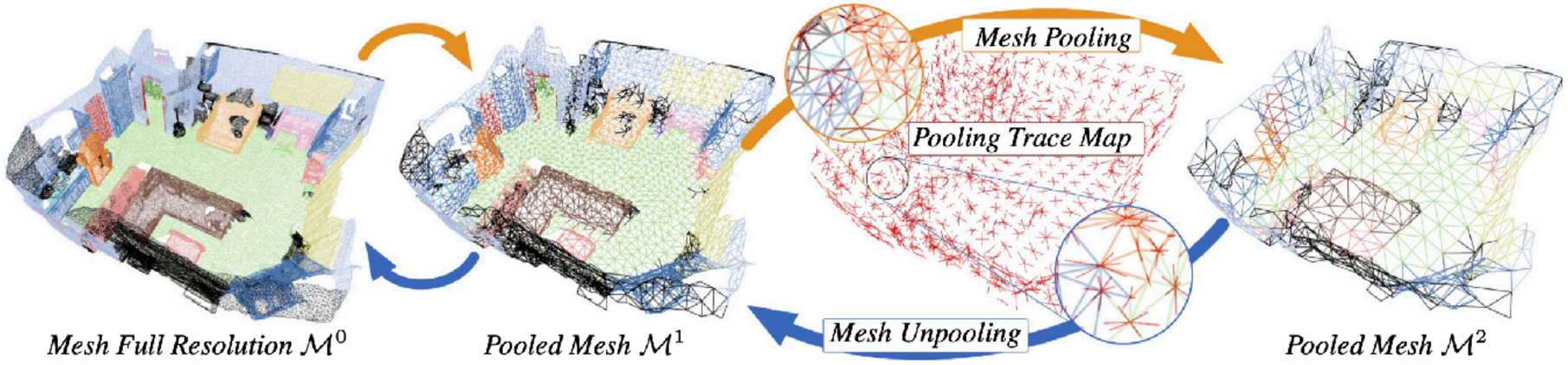
So far: Convolutions based on Euclidean distance between points

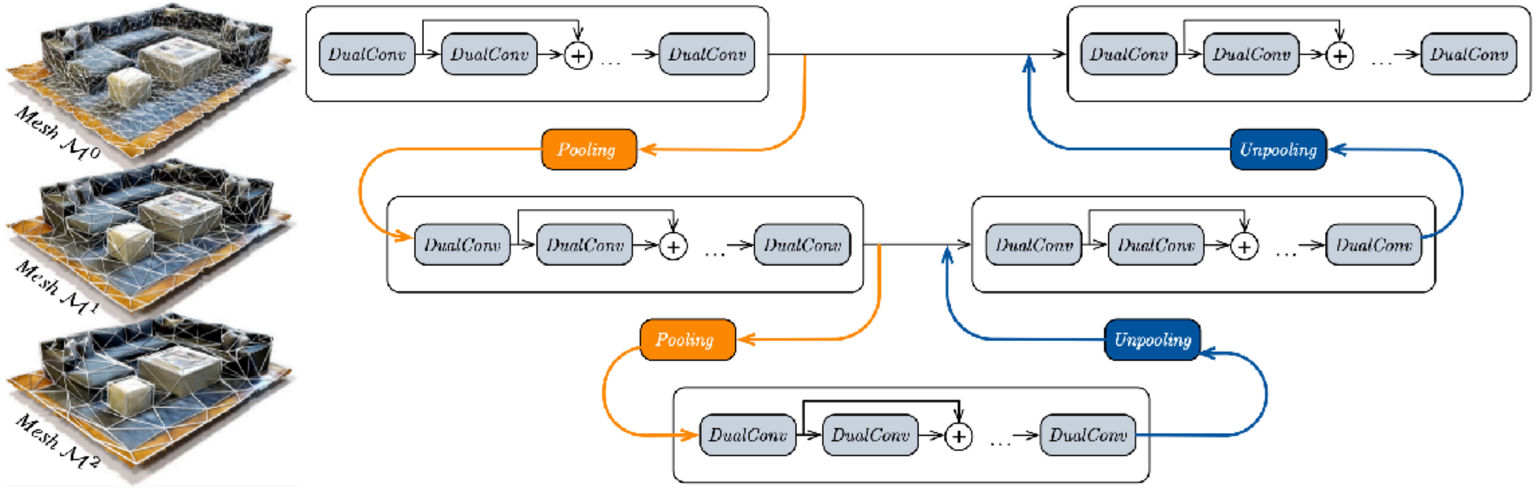
Question: Can we use the mesh as well?



DualConvMesh-Net: Joint Geodesic and Euclidean Convolutions on 3D Meshes
Jonas Schult*, Engelmann Francis*, Kontogianni Theodora, Leibe Bastian, CVPR'20

Dilated Point Convolutions

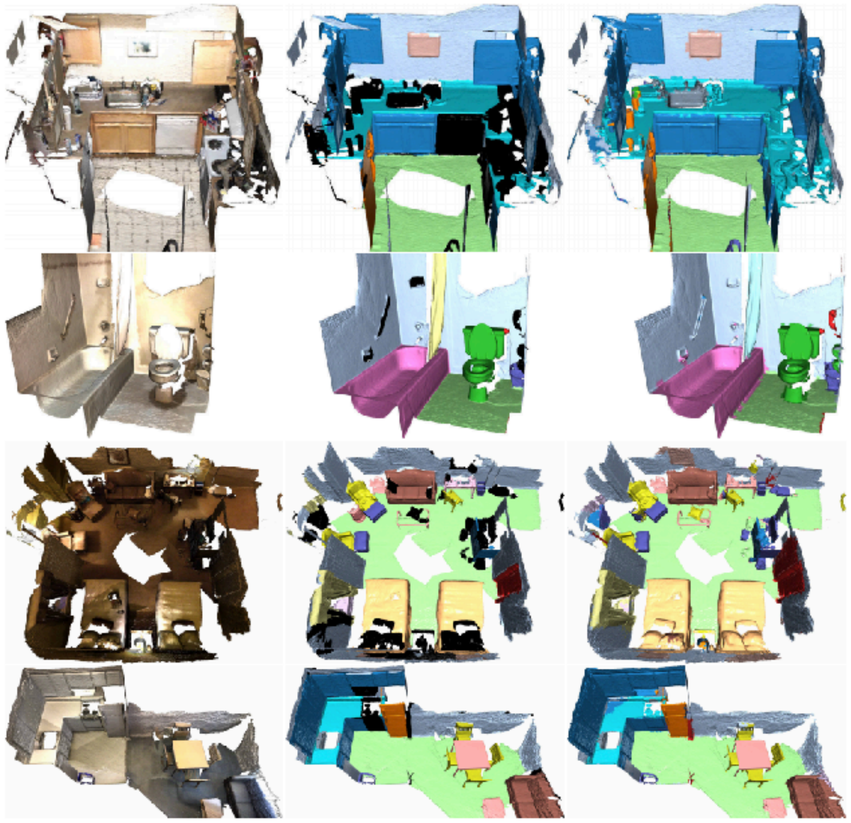




DualConvMesh-Net: Joint Geodesic and Euclidean Convolutions on 3D Meshes
Jonas Schult*, Engelmann Francis*, Kontogianni Theodora, Leibe Bastian, CVPR'20

Experiments

Method	mIoU ScanNet	mIoU S3DIS	Convolutional Category
PointNet [46]	-	41.1	Permutation Invariant Networks
PointNet++ [47]	33.9	-	
FCPN [11]	44.7	-	
3DMV [9]	48.3	-	2D-3D
JPBNet [6]	63.4	-	
MVPNet [32]	64.1	62.4	
TangentConv [55]	43.8	52.6	SurfaceConv
SurfaceConvPF* [24]	44.2	-	
TextureNet [30]	56.6	-	
PointCNN [40]	45.8	57.3	PointConv
ParamConv [59]	-	58.3	
DPC [16]	59.2	61.3	
MCCN [27]	63.3	-	
PointConv [63]	66.6	-	
KPConv [57]	68.4	**67.1**	
SparseConvNet [21]	72.5	-	Voxelized SparseConv
MinkowskiNet [7]	**73.4**	65.3	
DeepGCN [38]	-	52.5	GraphConv
SPGraph [41]	-	58.0	
SPH3D-GCN* [37]	61.0	59.5	
HPEIN [33]	61.8	61.9	
DCM-Net (**Ours**)	**65.8**	**64.0**	



pool	arch	neighb	mIoU (± stdev)	Δ
VC	Single	geo	57.1	−0.3
QEM	Single	geo	56.8	
VC	Single	knn	60.1	+0.8
QEM	Single	knn	60.9	
VC	Single	rad	61.9 (± 0.20)	+2.0
FPS	Single	rad	63.5 (± 0.13)	+0.4
QEM	Single	rad	63.9 (± 0.20)	
VC	Dual	knn/geo	59.7	+3.2
QEM	Dual	knn/geo	62.9	
VC	Dual	rad/geo	62.8 (± 0.12)	+4.5
QEM	Dual	rad/geo	67.3 (± 0.22)	

Table 3: **Comparison of pooling methods.** We compare Vertex Clustering (VC), Farthest Point Sampling (FPS), and Quadric Error Metrics (QEM) as pooling methods.

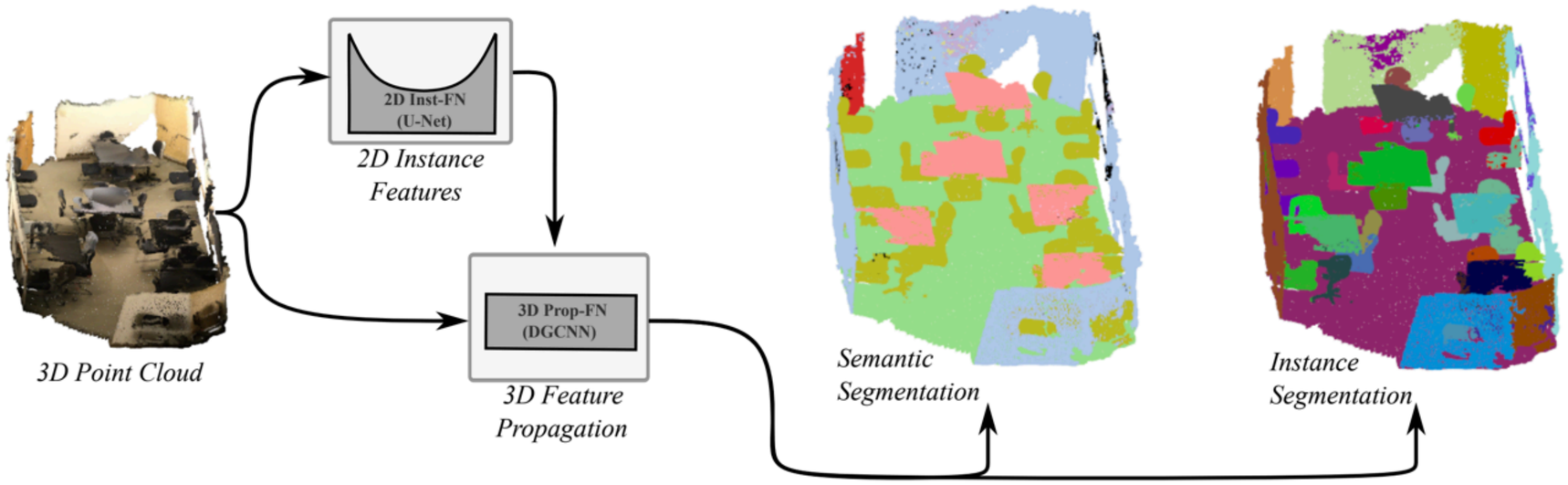
Overview

Shape and Pose Reconstruction

3D Semantic Segmentation

3D Instance Segmentation

3D-BEVIS



Input: $\mathcal{B}$ GT instance seg. Output: $\mathcal{E}$

3D-BEVIS: Bird's-Eye-View Instance Segmentation
Cathrin Elich, Francis Engelmann, Jonas Schult, Theodora Kontogianni, Bastian Leibe (GCPR'19)

3D-MPA: Multi Proposal Aggregation for 3D Semantic Instance Segmentation

Input: 3D Point Cloud *Object Center Votes & Aggregated Proposals* *Output: 3D Semantic Instances*

3D-MPA: Multi Proposal Aggregation for 3D Semantic Instance Segmentation
Engelmann Francis, Martin Bokeloh, Alireza Fathi, Leibe Bastian, Matthias Nießner, CVPR'20

Motivation

Existing Methods:

- Top-down: first detect objects, then compute mask (Mask-RCNN, 3D-SIS, …)
- *Sensitive to missed detections*

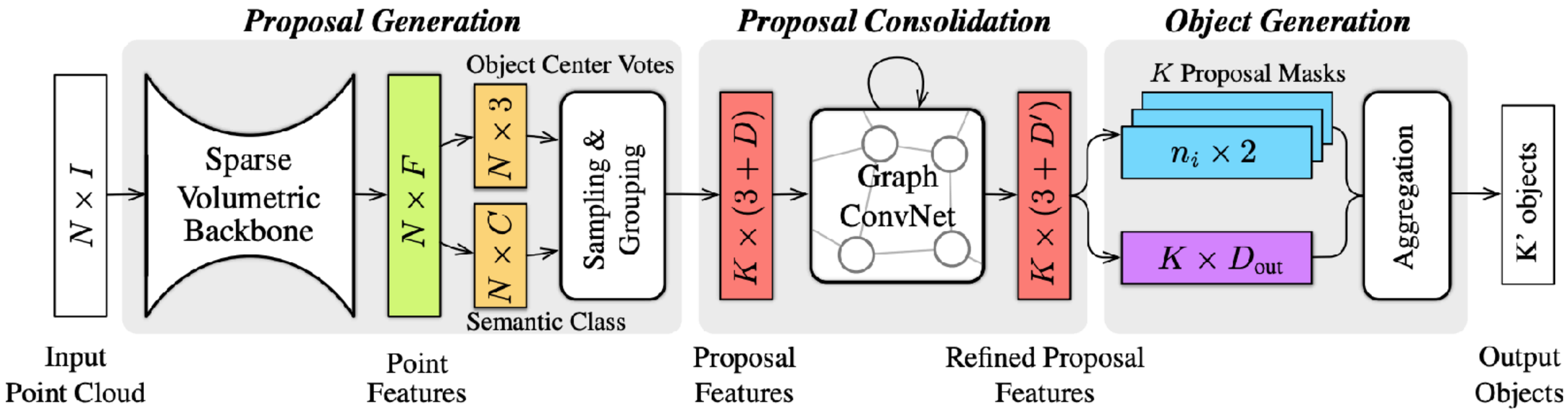
- Bottom-up: learn feature embedding space and apply clustering (MTML)
- *Clustering computationally expensive*

Motivation

Our Approach:

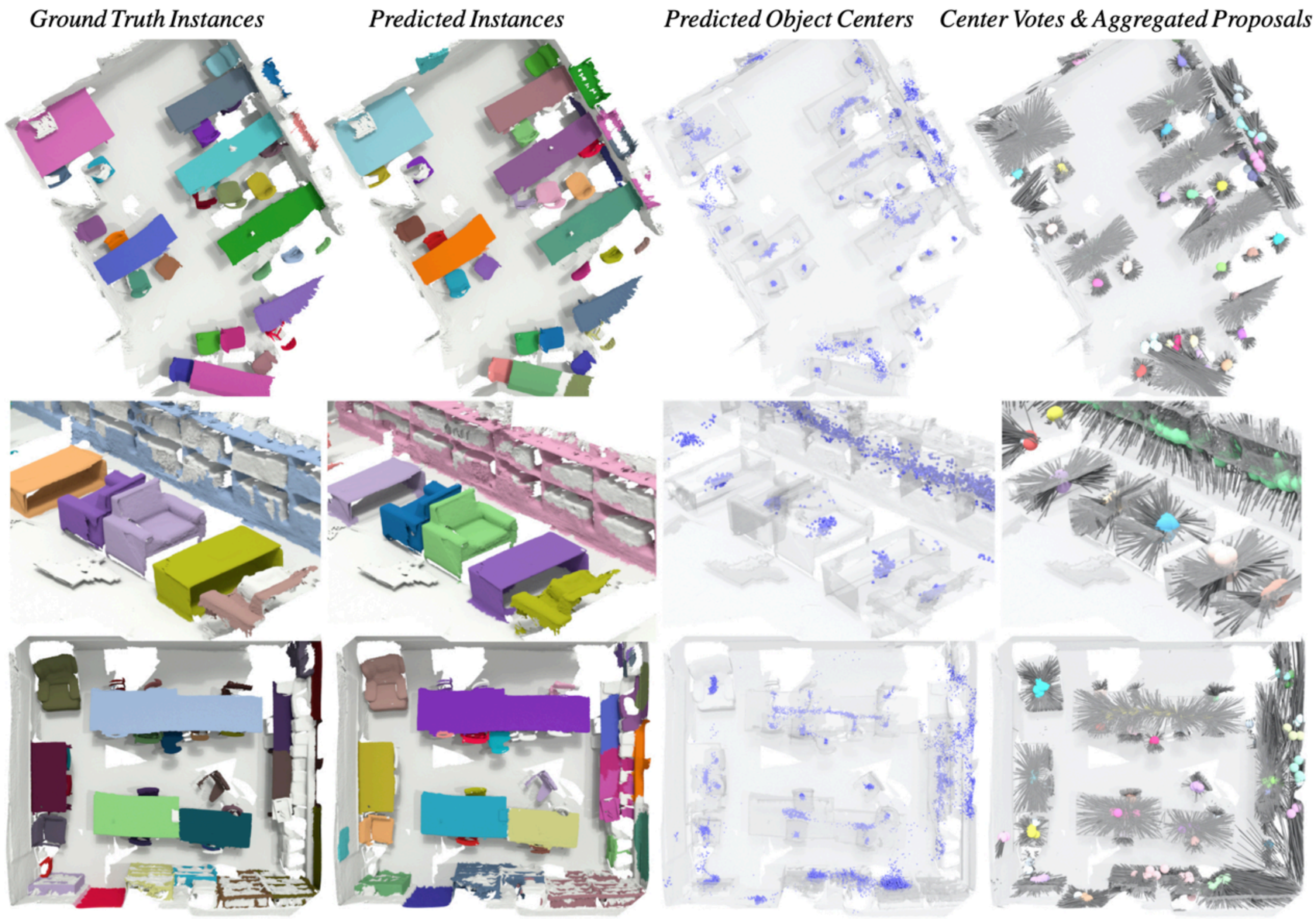
- First, group points into proposals based on predicted objects centers
- Then, aggregate / cluster proposals based on learned features

3D-MPA: Architecture



Input: 3D Point Cloud *Object Center Votes & Aggregated Proposals* *Output: 3D Semantic Instances*

Qualitative Results



Quantitative Results

3D Instance Segmentation

ScanNetV2	Validation Set			Hidden Test Set		
	mAP	@50%	@25%	mAP	@50%	@25%
SGPN [44]	-	11.3	22.2	4.9	14.3	39.0
3D-BEVIS [10]	-	-	-	11.7	24.8	40.1
3D-SIS [18]	-	18.7	35.7	16.1	38.2	55.8
GSPN [50]	19.3	37.8	53.4	15.8	30.6	54.4
3D-BoNet [49]	-	-	-	25.3	48.8	68.7
MTML [19]	20.3	40.2	55.4	28.2	54.9	73.1
3D-MPA (Ours)	**35.3**	**59.1**	**72.4**	**35.5**	**61.1**	**73.7**

3D Object Detection

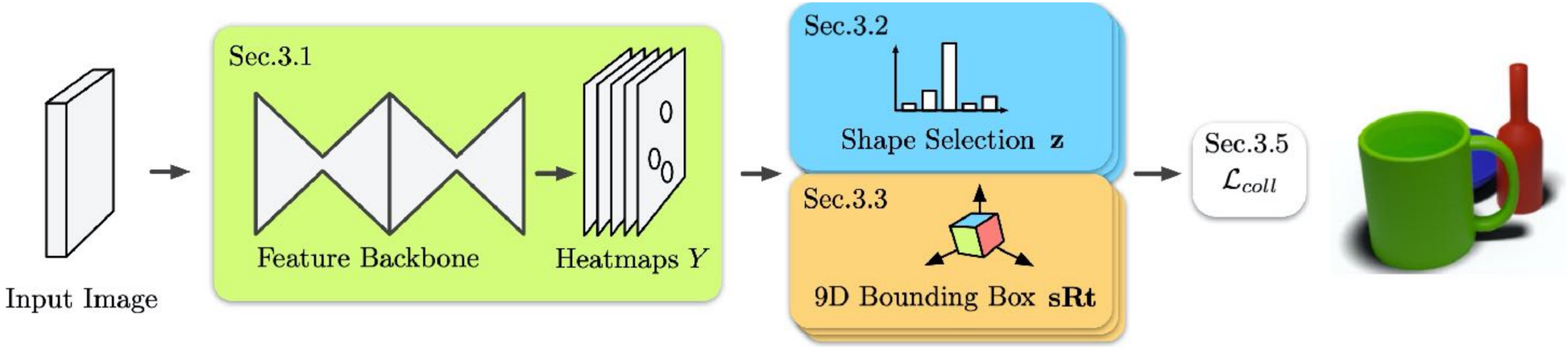
ScanNetV2	mAP@25%	mAP@50%
DSS [37]	15.2	6.8
MRCNN 2D-3D [17]	17.3	10.5
F-PointNet [30]	19.8	10.8
GSPN [50]	30.6	17.7
3D-SIS [18]	40.2	22.5
VoteNet [29]	58.6	33.5
3D-MPA (Ours)	**64.2**	**49.2**

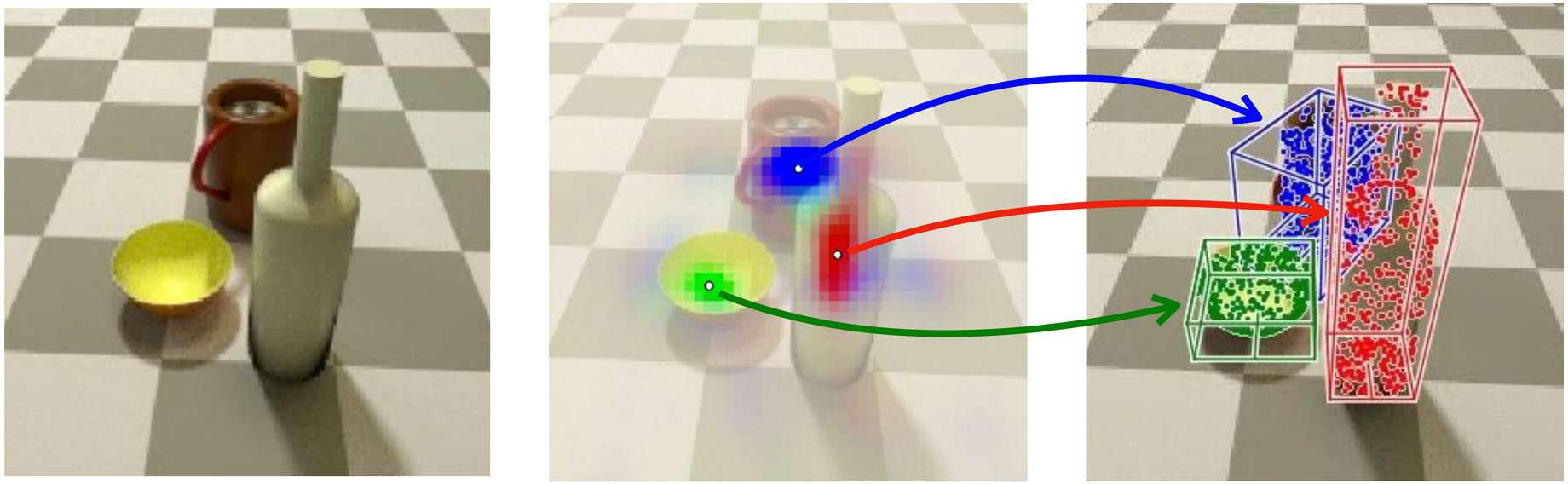
Ablation Study

3D Instance Segmentation (ScanNetV2 val.)	mAP@50%
① Proposals + NMS	47.5
② Agg. Props. (proposal positions)	52.4 (+4.9)
③ Agg. Props. (embedding features)	56.7 (+9.2)
④ Agg. Props. (geometric features)	57.8 (+10.3)
⑤ Agg. Props. (geometric features + GCN)	**59.1** (+11.6)

3D-MPA: Multi Proposal Aggregation for 3D Semantic Instance Segmentation
Engelmann Francis, Martin Bokeloh, Alireza Fathi, Leibe Bastian, Matthias Nießner, CVPR'20

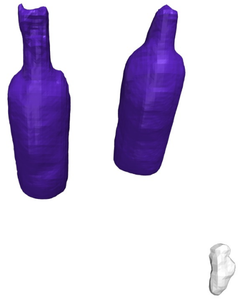
Last work: From Points to 3D Objects





Qualitative Results

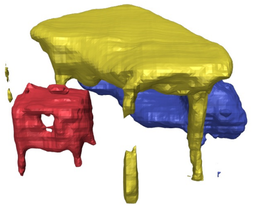
Input Image



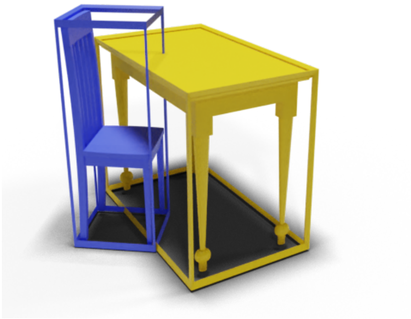
CoReNet

Ours

Input Image



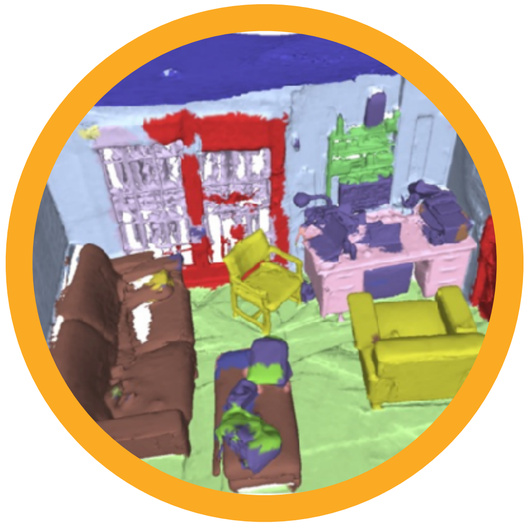
CoReNet



Ours

Summary

Shape and Pose Reconstruction



3D Semantic Segmentation

3D Instance Segmentation

Current Work / Next Work

Current work

- Self-supervised learning on 3D
- Analysis of data augmentation in 3D

Potential next directions

- High-level semantic reconstruction of scenes based on primitives
- Semantic Graph Reconstructions - how can humans interact with environments?
- Detecting and fixing dynamic objects in long-term scene reconstruction