



Sensor Fusion for Semantic Segmentation of Urban Scenes

Richard Zhang¹ Stefan A. Candra¹ Kai Vetter^{1,2} Avideh Zakhor¹

¹Department of Electrical Engineering and Computer Science, UC Berkeley

²Department of Nuclear Engineering, UC Berkeley

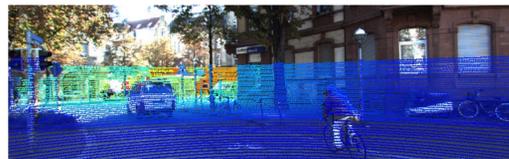


Introduction

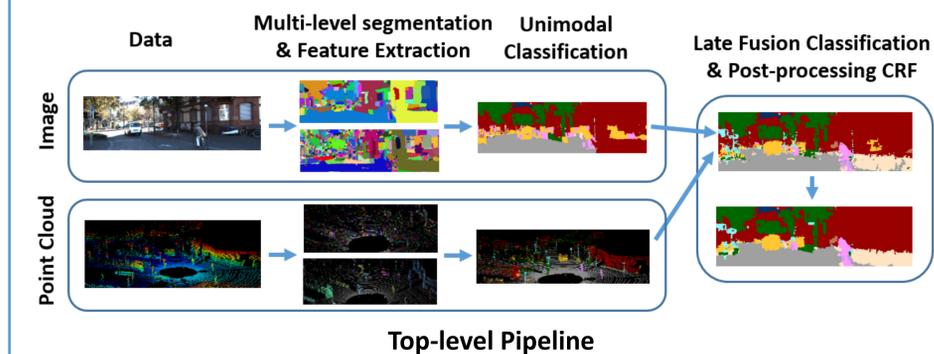
Goal: effectively fuse information from multiple modalities to obtain semantic information

Contributions:

- information from multiple scales considered
- late fusion used to maximally leverage training data
- post-processing CRF used
- validated on KITTI data [1] with augmented labels; performance increase obtained over state-of-the-art method [3]



building ■ sky ■ road ■ vegetation ■ sidewalk ■ car ■ pedestrian ■ cyclist ■ sign/pole ■ fence ■

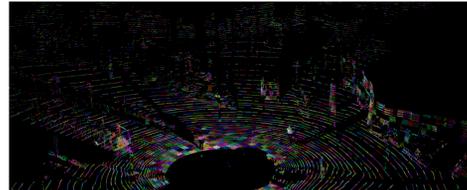


Segmentation

- Multiple segmentations to consider cues from varying scales of information in classification
- Image: hierarchical segmentation [2] extracted
- Point cloud: 0.5 m supervoxels and connected component segmentation

Feature extraction

- Inference performed on low level segments
- Low level segments associated with high-level segments
- Feature vectors of low level segments augmented with associated high level segment
- High dimensional features extracted for low level segments



Features Extracted

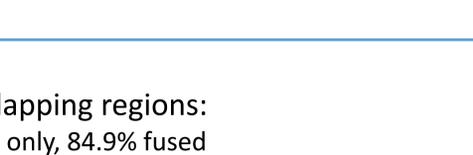
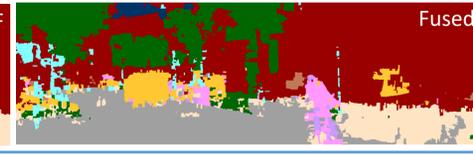
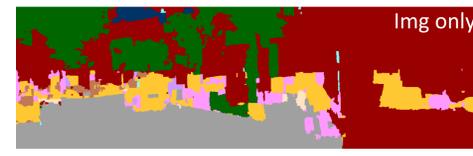
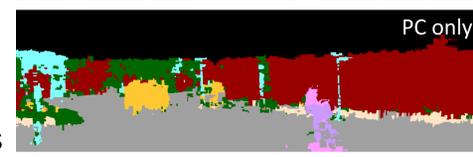
Type	Name	Dim	Low	High
Size	Length proxy - λ_1	1	✓	✓
	Area proxy - $\sqrt{\lambda_1 \lambda_2}$	1	✓	✓
	Volume proxy - $\sqrt[3]{\lambda_1 \lambda_2 \lambda_3}$	1	✓	✓
	Scatter - λ_3/λ_1	1	✓	✓
Shape	Planarity - $(\lambda_2 - \lambda_3)/\lambda_1$	1	✓	✓
	Linearity - $(\lambda_1 - \lambda_2)/\lambda_1$	1	✓	✓
	Verticalness - v_{1z}	1	✓	✓
Position	$z - z_{\text{groundplane}} - \min, \text{mean}, \text{max}$	3	✓	✓
	Horizontalness - $\sqrt{1 - v_{1z}^2}$	1	✓	✓
High-dim	Spin image BoW	1000	✓	✓
	Area	1	✓	✓
	Equivalent Diameter	1	✓	✓
Size/Shape	Major/minor axes	2	✓	✓
	Orientation	1	✓	✓
Position	$(x, y) - \min, \text{mean}, \text{max}$	6	✓	✓
	superpixel mask (8x8)	64	✓	✓
Color	rgb+lab (mean, std)	6	✓	✓
	rgb+lab (histogram)	48	✓	✓
Contextual	SIFT BoW	400	✓	✓
	contextual rgb+lab (mean, std)	6	✓	✓
	contextual rgb+lab (histogram)	48	✓	✓
	contextual SIFT BoW	400	✓	✓

Point cloud supervoxel features

Image superpixel features

Classification & Fusion

- Random Forest (RF) classifier used for each modality separately
- Weights for samples of rare classes artificially boosted
- For overlapping region, fusion classifier evaluated on output pmfs of unimodality classifications
- Pmfs serve as *compact* and *descriptive* mid-level features
- Post-processing pairwise CRF



Late-fusion Results

- Performance increases. On overlapping regions:
 - Pixel-wise: 68.1% pc only, 77.8% img only, 84.9% fused
 - Class-wise: 41.4% pc only, 52.1% img only, 65.2% fused

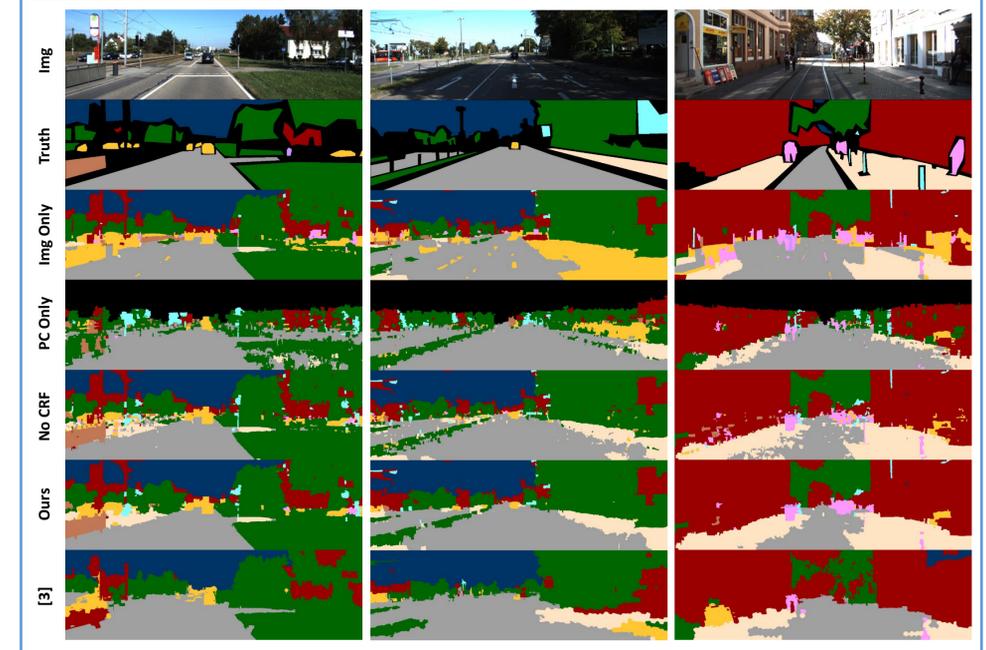
Examples

- *sidewalk* more likely to be classified correctly vs road only after fusion
- modes of failure can be found during fusion e.g. looks *building*-like in point cloud and *road*-like in image → actually a fence

	building	sky	road	vegetation	sidewalk	car	pedestrian	cyclist	signage	fence
building	.75	.19	.01	.03	.01	.01	.01	.01	.01	.01
sky	.07	.63	.01	.01	.01	.01	.01	.01	.01	.01
road	.01	.01	.99	.01	.01	.01	.01	.01	.01	.01
vegetation	.14	.14	.66	.01	.03	.01	.01	.01	.01	.01
sidewalk	.01	.83	.03	.13	.01	.01	.01	.01	.01	.01
car	.14	.12	.13	.54	.01	.01	.01	.01	.01	.01
pedestrian	.16	.09	.16	.02	.12	.44	.01	.01	.01	.01
cyclist	.09	.28	.10	.02	.03	.34	.14	.01	.01	.01
signage	.33	.12	.21	.01	.01	.02	.30	.01	.01	.01
fence	.18	.08	.54	.01	.01	.20	.01	.01	.01	.01

Performance on overlapping region

Qualitative Results



Conclusions

- Dataset: 252 images (140 training, 112 testing) from 8 sequences
- multiscale information provides strong cues for classifier
- late fusion greatly boosts performance
- performance increase over current state-of-the-art [3]
- *stuff* classes well discriminated

Path Forward

- Add 2D+3D object detectors to increase performance on *things*
- Enforce consistency across *temporal* and *3D spatial* dims
- Integrate with reconstruction algorithms
- Extend algorithm to additional modalities, e.g. infrared and hyperspectral, and validate

	glob	class	bldg	sky	road	veg
Cadena <i>et al.</i> [3]	84.1%	52.4%	92.5%	95.7%	92.5%	86.3%
Ours (image only)	83.5%	53.3%	87.5%	92.5%	94.5%	92.5%
Ours (late fused)	88.0%	64.8%	93.5%	92.5%	91.2%	92.0%
Ours (CRF)	89.3%	65.4%	95.0%	92.6%	92.6%	92.8%

	side	car	ped	cycl	sgn	fnc
Cadena <i>et al.</i> [3]	51.5%	67.9%	28.6%	4.0%	2.5%	2.3%
Ours (image only)	34.5%	71.4%	49.0%	3.6%	4.1%	3.3%
Ours (late fused)	69.7%	76.5%	63.7%	10.0%	16.6%	42.2%
Ours (CRF)	73.3%	78.7%	65.1%	7.3%	13.8%	43.2%

	building	sky	road	vegetation	sidewalk	car	pedestrian	cyclist	signage	fence
building	.95	.01	.02	.01	.01	.01	.01	.01	.01	.01
sky	.07	.93	.01	.01	.01	.01	.01	.01	.01	.01
road	.01	.01	.93	.01	.01	.01	.01	.01	.01	.01
vegetation	.05	.01	.01	.93	.01	.01	.01	.01	.01	.01
sidewalk	.01	.23	.02	.73	.01	.01	.01	.01	.01	.01
car	.15	.01	.01	.01	.79	.02	.01	.01	.01	.01
pedestrian	.15	.01	.02	.04	.12	.65	.01	.01	.01	.01
cyclist	.02	.04	.03	.07	.03	.73	.07	.01	.01	.01
signage	.71	.01	.07	.03	.01	.03	.14	.01	.01	.01
fence	.11	.05	.35	.02	.02	.01	.43	.01	.01	.01

References

- [1] Geiger, et al. Vision meets robotics: The KITTI Dataset. IJRR 2013.
- [2] Arbelaez. Multiscale combinatorial grouping. CVPR 2014.
- [3] Cadena and Kosecka. Semantic segmentation with heterogeneous sensor coverages. ICRA 2014.