Sensor Fusion for Semantic Segmentation of Urban Scenes

Richard Zhang[1] Stefan A. Candra[1] Kai Vetter[12] Avideh Zakhor[1]

[1]Department of Electrical Engineering and Computer Science, UC Berkeley

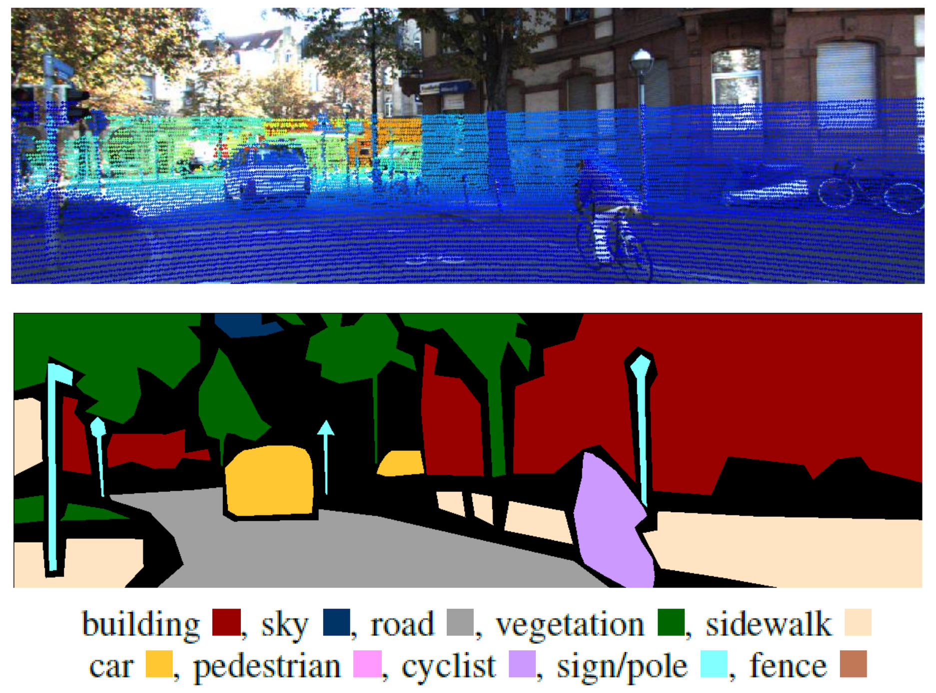
[2]Department of Nuclear Engineering, UC Berkeley

Introduction

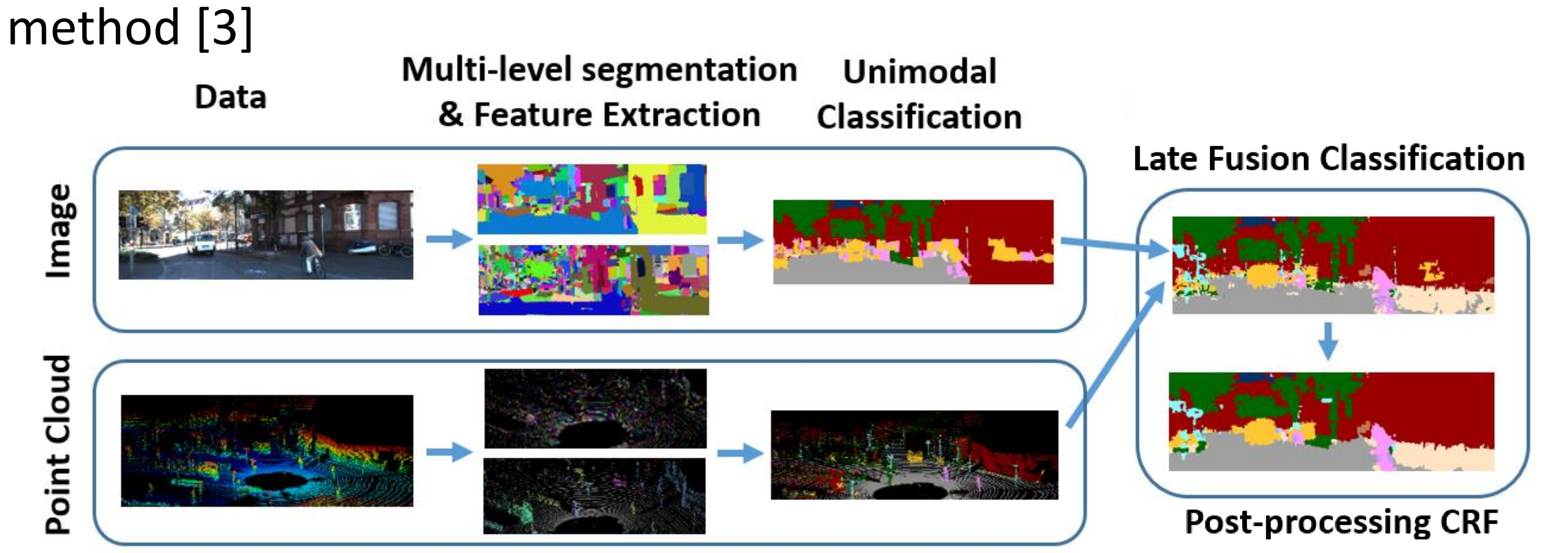
Goal: effectively fuse information from multiple modalities to obtain semantic information

Contributions:

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



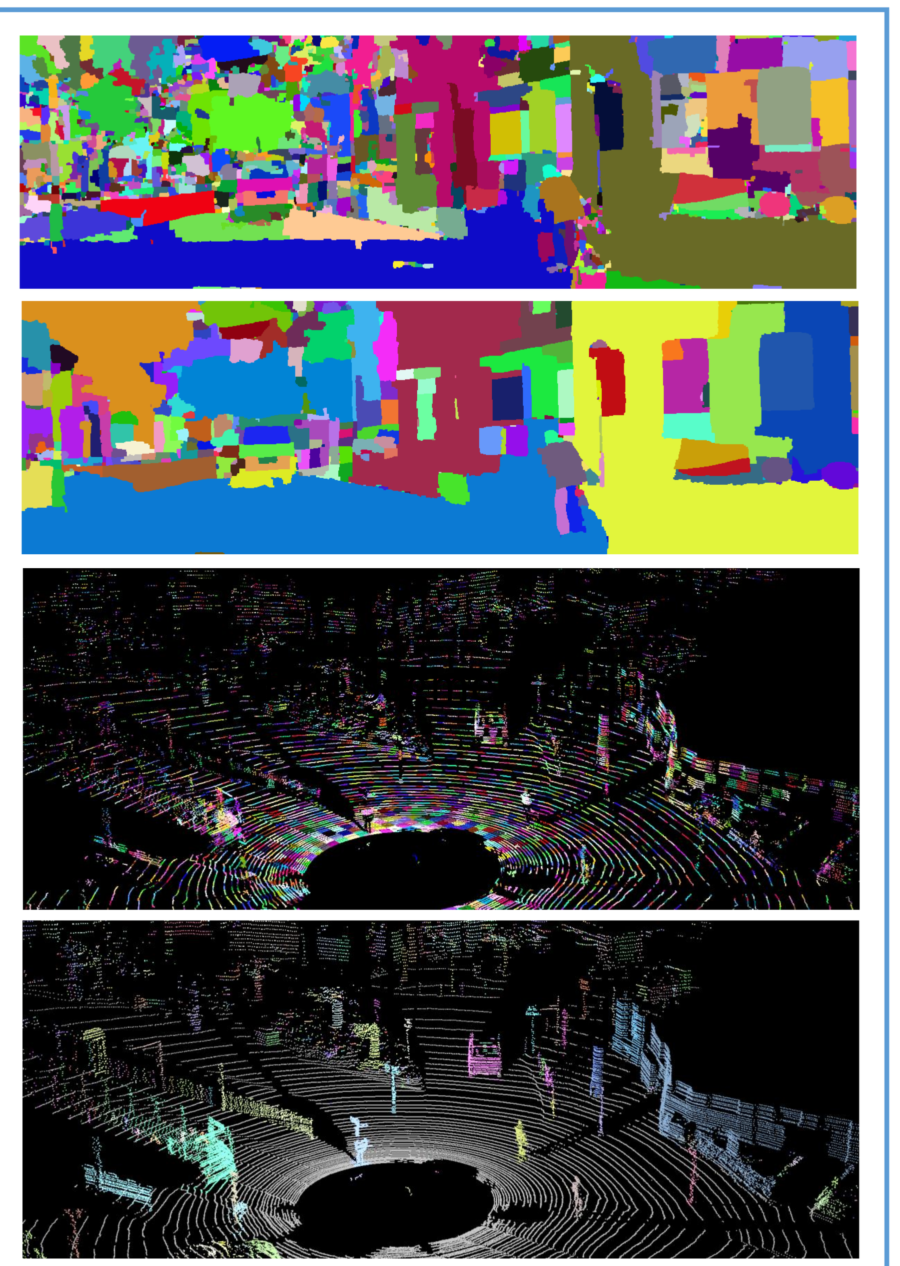
	Segmentation	Features	Fusion
[3]	Single segmentation image only	Simple	Early fusion
Proposed	***Multiple*** segmentations ***both*** domains	***Descriptive***	***Late fusion***



Top-level Pipeline

Multi-Level 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 small-scale segments
- Small-scale segments associated with large-scale segments
- Feature vectors of small-scale segments augmented with associated large-scale segment

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	✓	✓
Shape	Scatter - λ_3/Λ	1	✓	✓
	Planarity - $(\lambda_2 - \lambda_3)/\Lambda$	1	✓	✓
	Linearity - $(\lambda_1 - \lambda_2)/\Lambda$	1	✓	✓
Position	$z - z_{gndplane}$ - min, mean, max	3	✓	✓
Orientation	Verticalness - v_{1z}	1	✓	✓
	Horizontalness - $\sqrt{1 - v_{1z}^2}$	1	✓	✓
High-dim	Spin image BoW	1000	✓	

Point cloud supervoxel features

Type	Name	Dim	Low	High
Size/Shape	Area	1	✓	✓
	Equivalent Diameter	1	✓	✓
	Major/minor axes	2	✓	✓
	Orientation	1	✓	✓
	Eccentricity	1	✓	✓
Position	(x, y) - min, mean, max	6	✓	✓
	superpixel mask (8x8)	64	✓	✓
Color	rgb+lab (mean, std)	6	✓	✓
	rgb+lab (histogram)	48	✓	✓
High-dim	SIFT BoW	400	✓	
Contextual	contextual rgb+lab (mean, std)	6	✓	
	contextual rgb+lab (histogram)	48	✓	
	contextual SIFT BoW	400	✓	

Image superpixel features

Classification & Late-Fusion

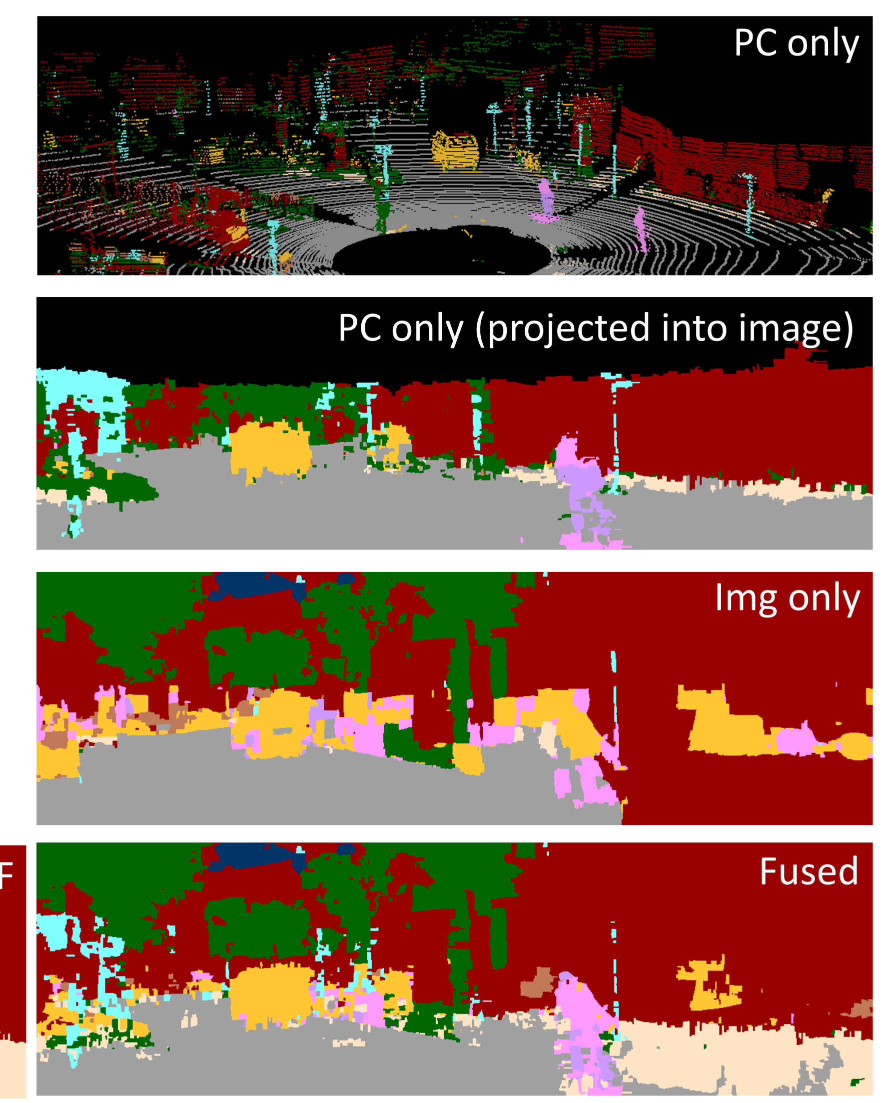
- Random Forest (RF) classifier used for each modality separately

$$P_{img} : \mathbb{R}^{N_{img}} \rightarrow \Delta^L \qquad P_{pc} : \mathbb{R}^{N_{pc}} \rightarrow \Delta^L$$

- For overlapping region, fusion classifier evaluated on output PMFs of unimodal classifications

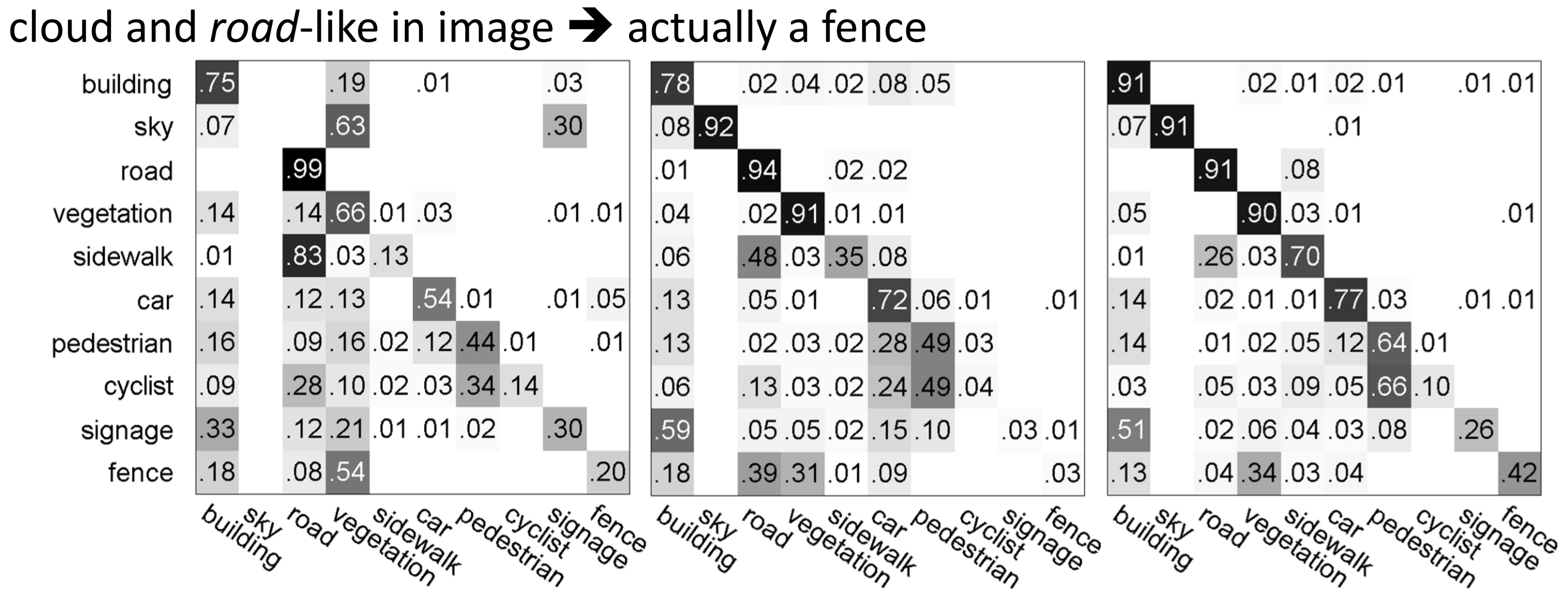
$$P_{latefusion} : \Delta^{2L} \rightarrow \Delta^L$$

- PMFs serve as *compact* and *descriptive* mid-level features
- Post-processing pairwise CRF provide spatial smoothing



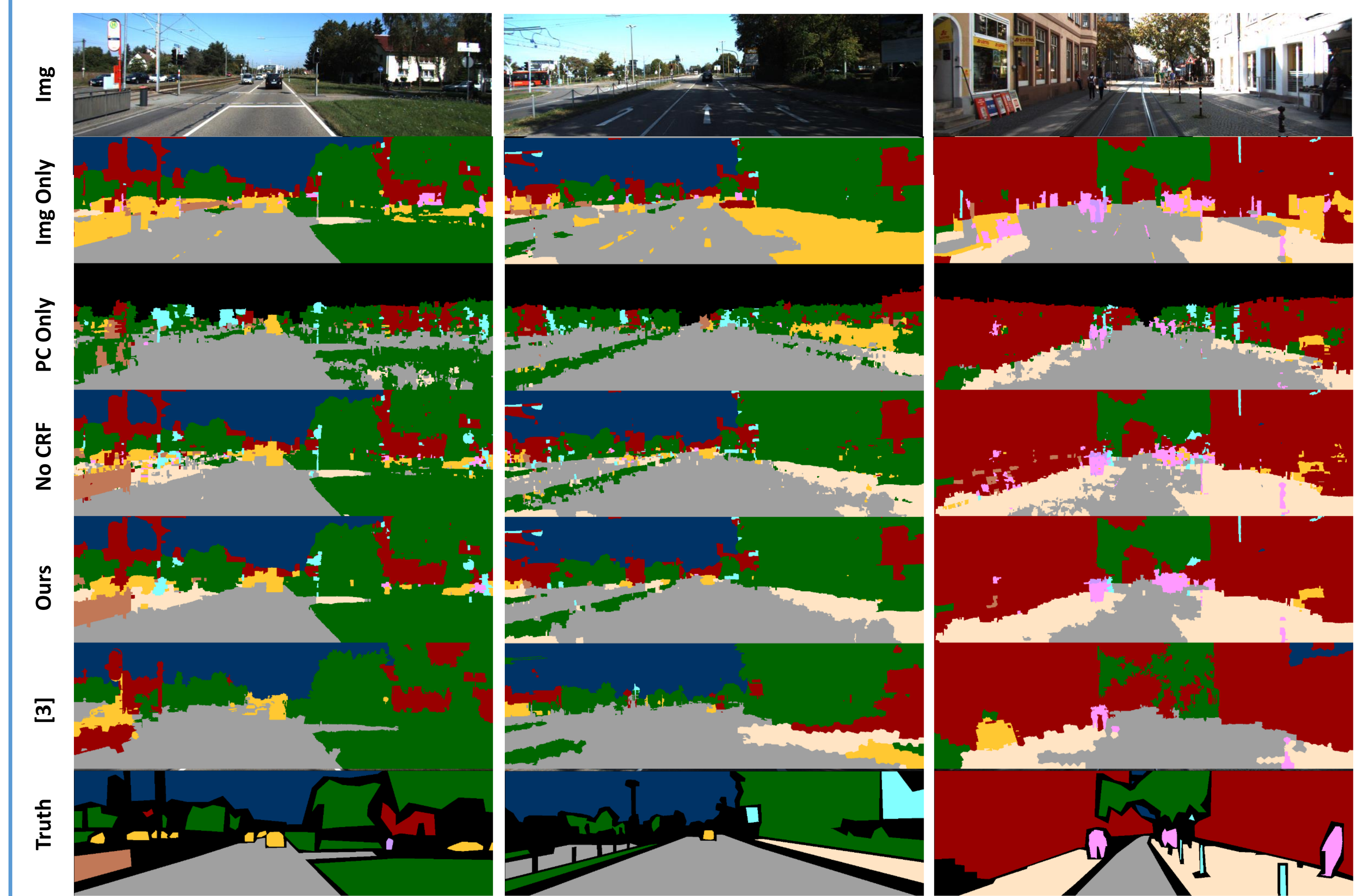
Late-fusion Results

- Fusion improves performance for 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



Performance on overlapping region

Qualitative Results

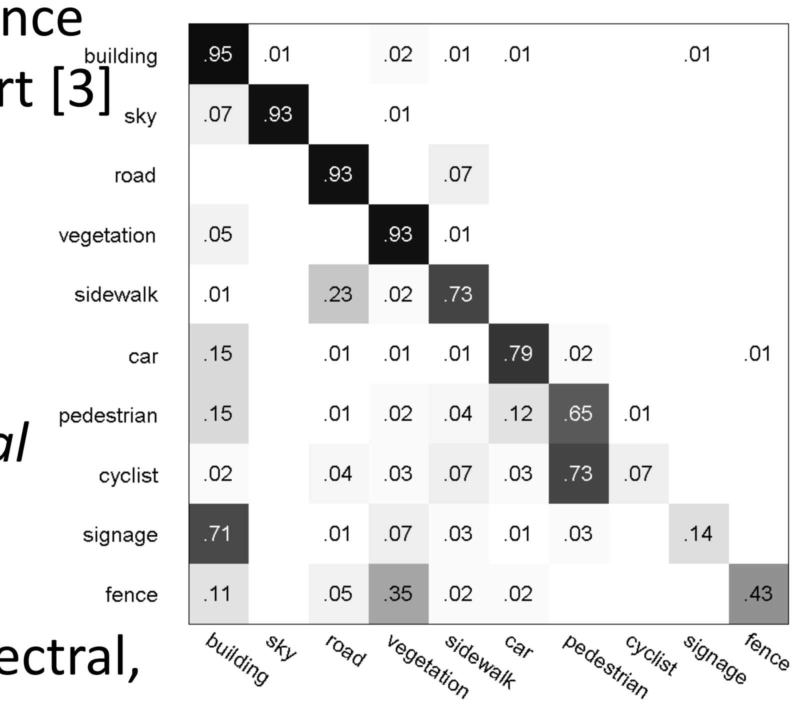


Conclusions

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

	glob	class	bldg	sky	road	veg
Cadena *et al.* [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 *et al.* [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%**



Path Forward

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

References

[1] Geiger, et al. Vision meets robotics: The KITTI Dataset. IJRR 2013.

[2] Arbelaez, et al. Multiscale combinatorial grouping. CVPR 2014.

[3] Cadena and Košecká. Semantic segmentation with heterogeneous sensor coverages. ICRA 2014.