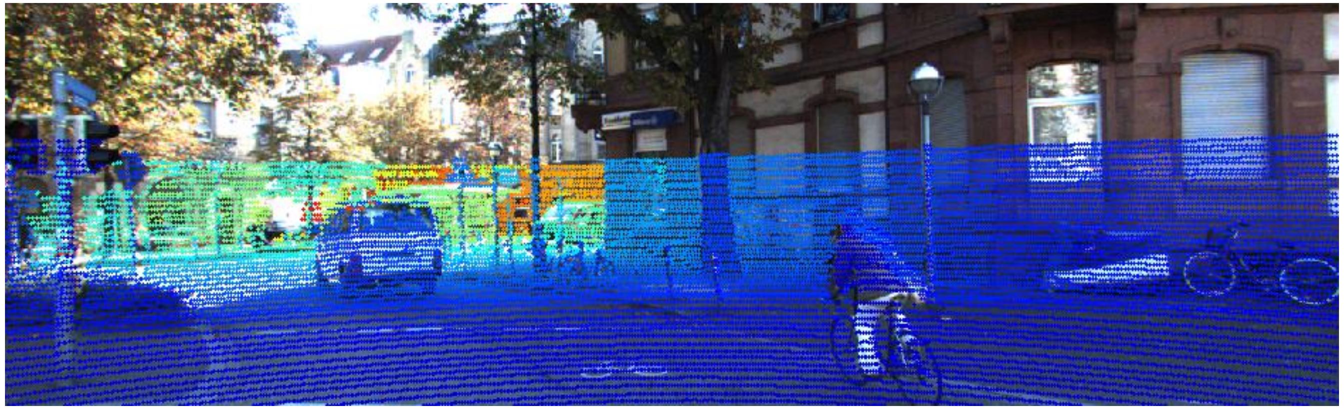
Sensor Fusion for Semantic Segmentation of Urban Scenes

Richard Zhang[1], Stefan A. Candra[1], Kai Vetter[12], Avideh Zakhor[1]
[1]University of California, Berkeley
[2]Lawrence Berkeley National Laboratory

Presented at ICRA, May 2015

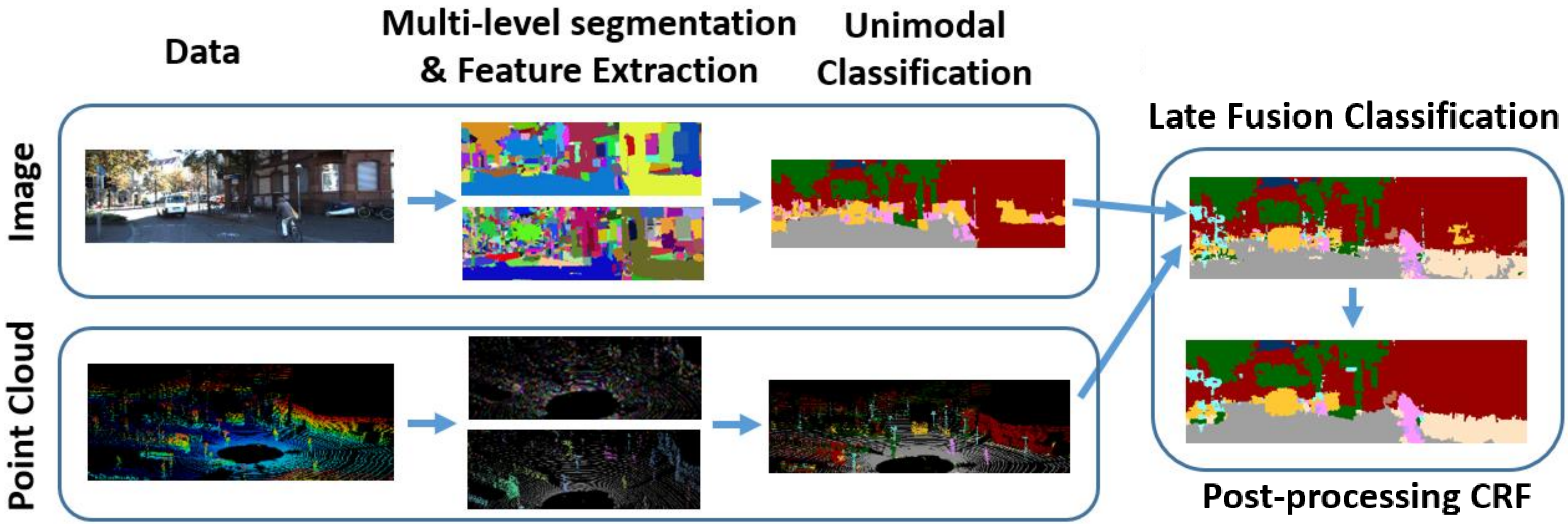
Goal

- Extract semantic information by fusing multiple modalities
  - Image
  - LiDAR scan
- Challenges
  - Incorporate information from multiple scales
  - Fuse information from multiple modalities with non-overlapping sensor coverage



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

Top Level Pipeline



Previous Work

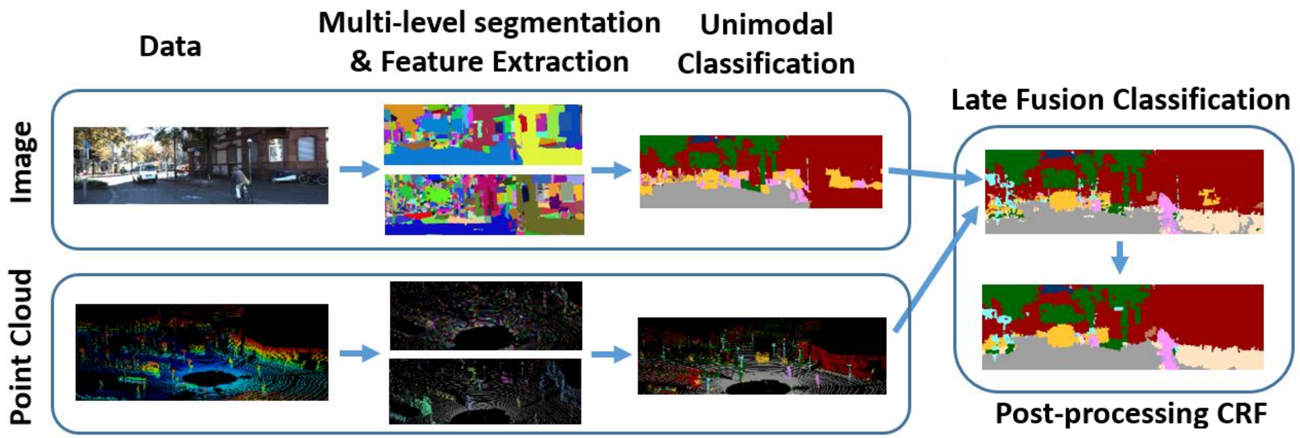
	Segmentation	Image Features	Point cloud features	Fusion
Sengupta et al. ICRA 2013	none	Descriptive	none	majority voting of image classifications in 3D
He and Upcroft. IROS 2013	Single segmentation image only	Descriptive	none	

Previous Work

	Segmentation	Image Features	Point cloud features	Fusion
Sengupta et al. ICRA 2013	none	Descriptive	none	majority voting of image classifications in 3D
He and Upcroft. IROS 2013	Single segmentation image only	Descriptive	none	
Cadena and Košecká. ICRA 2014.	Single segmentation image only	Simple	Simple	Early fusion

Previous Work

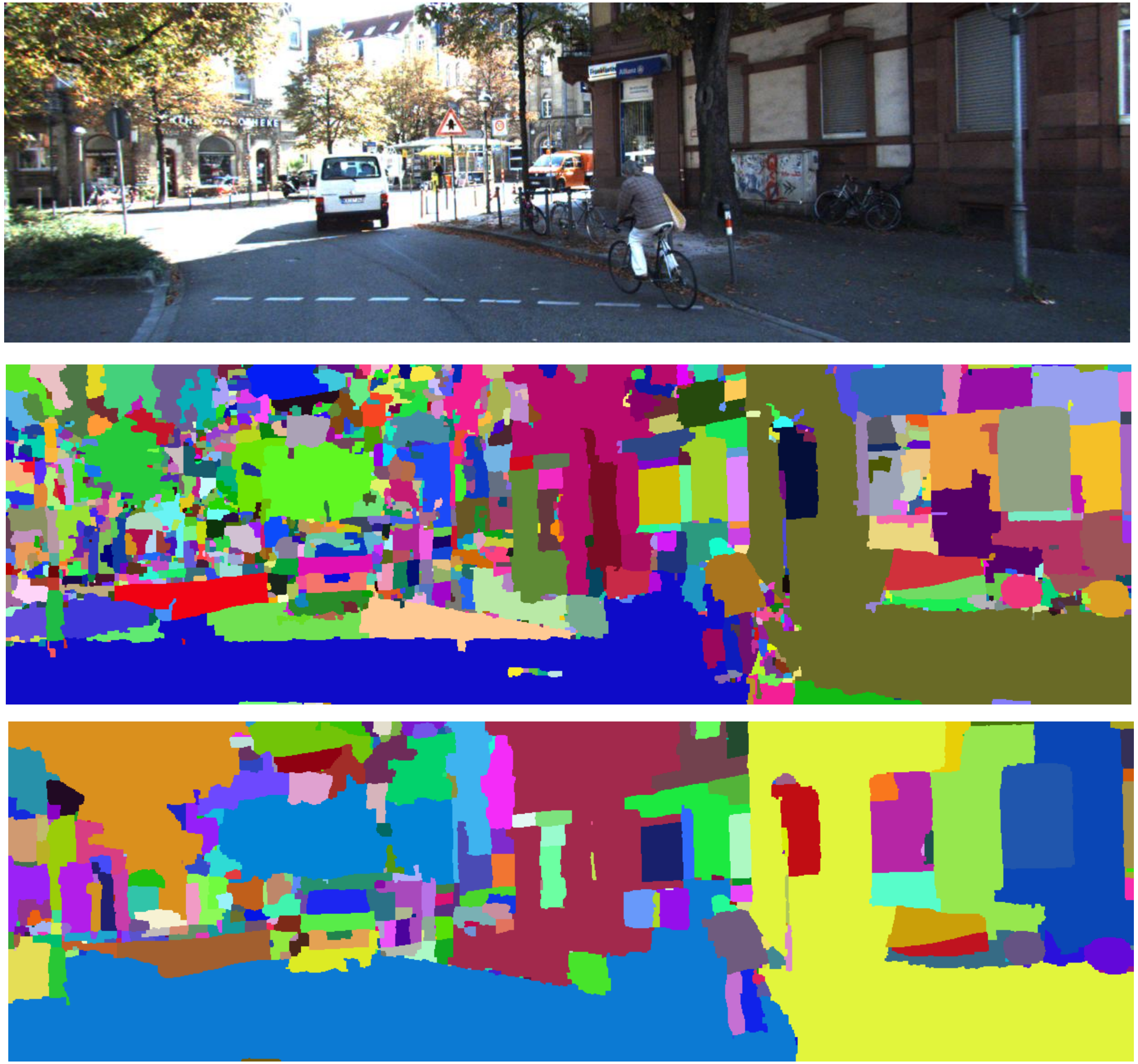
	Segmentation	Image Features	Point cloud features	Fusion
Sengupta et al. ICRA 2013	none	Descriptive	none	majority voting of image classifications in 3D
He and Upcroft. IROS 2013	Single segmentation image only	Descriptive	none	
Cadena and Košecká. ICRA 2014.	Single segmentation image only	Simple	Simple	Early fusion
Ours	***multiple*** segmentations ***both*** domains	***Descriptive***	***Descriptive***	***Late fusion***



Segmentation

- Treating each pixel/point separately is computationally intensive
- Under-segmentation & Over-segmentation errors

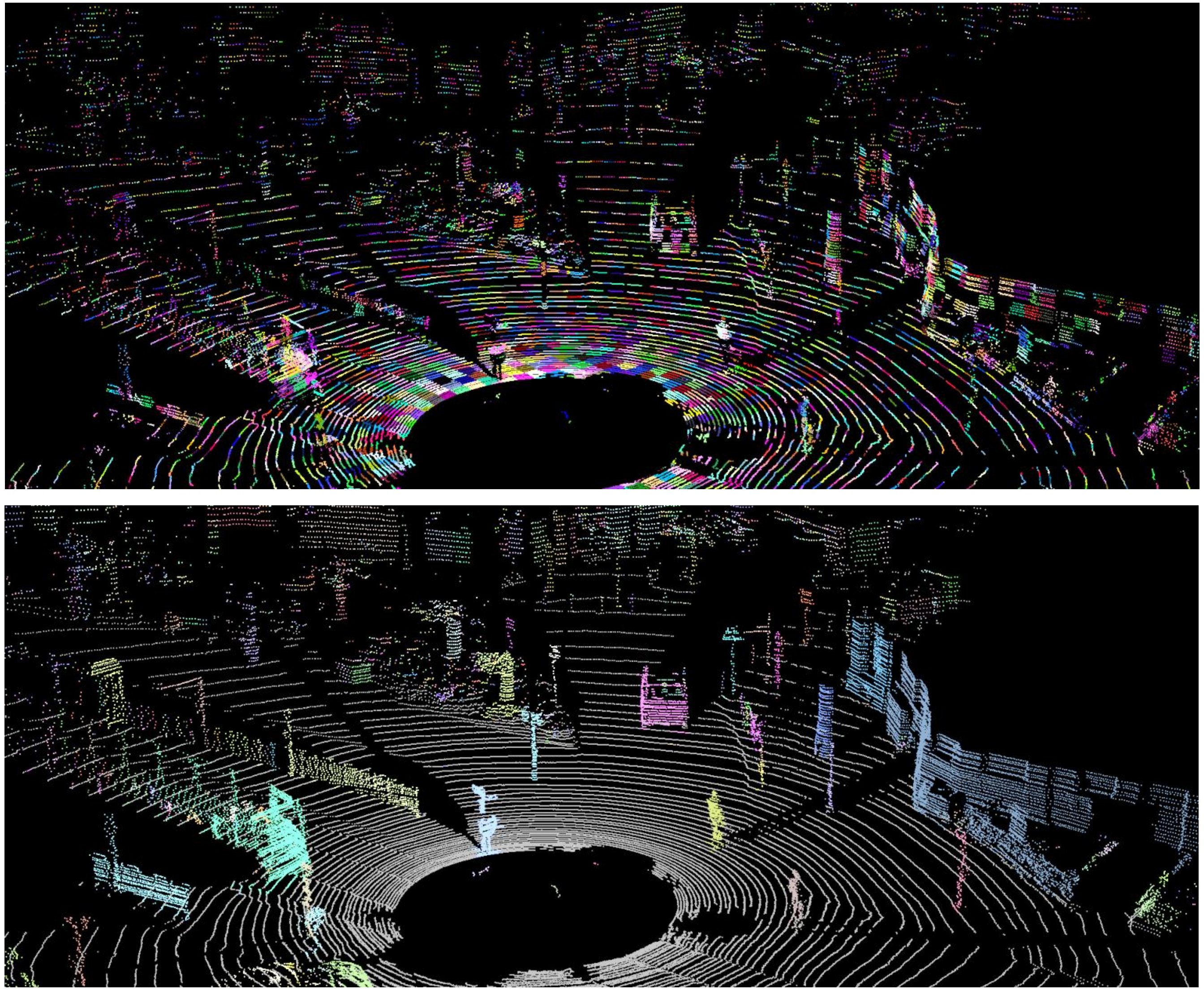
➔ *Multiple* segmentations performed to get *superpixels/supervoxels*



Segmentation

- Treating each pixel/point separately is computationally intensive
- Under-segmentation & Over-segmentation errors

➔ *Multiple* segmentations performed to get *superpixels/supervoxels*



Feature Extraction

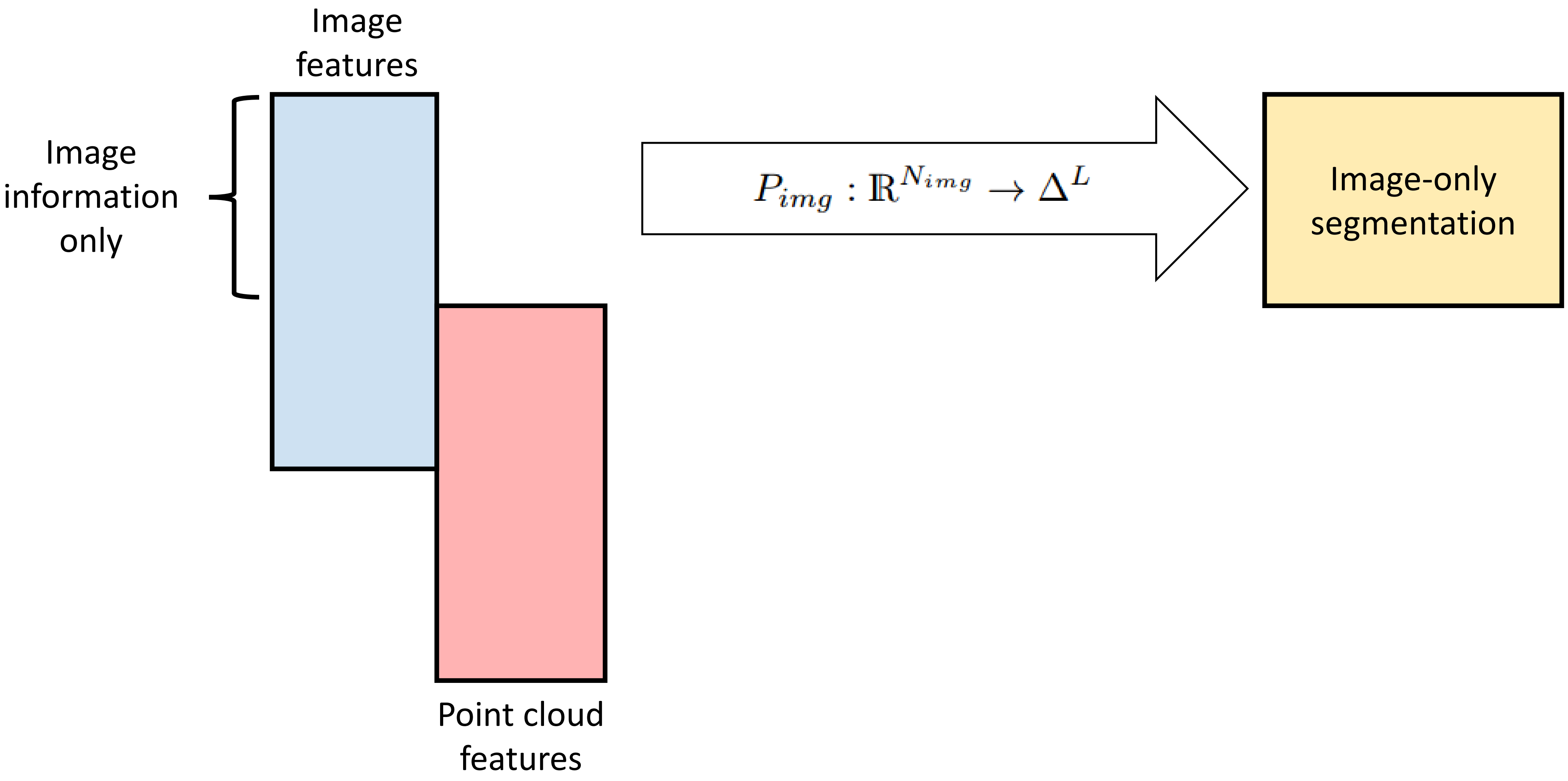
Type	Name	Dim	Low	High
Size/Shape	Area	1	✓	✓
	Equivalent Diameter	1	✓	✓
	Major/minor axes	2	✓	✓
	Orientation	1	✓	✓
	Eccentricity	1	✓	✓
Position	(x, y) - min, mean, max	6	✓	✓
	superpixel mask (8x8)	64	✓	✓
Color	rgb+lab (mean, std)	6	✓	✓
	rgb+lab (histogram)	48	✓	✓
High-dim	SIFT BoW	400	✓	
Contextual	contextual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

Image superpixel features

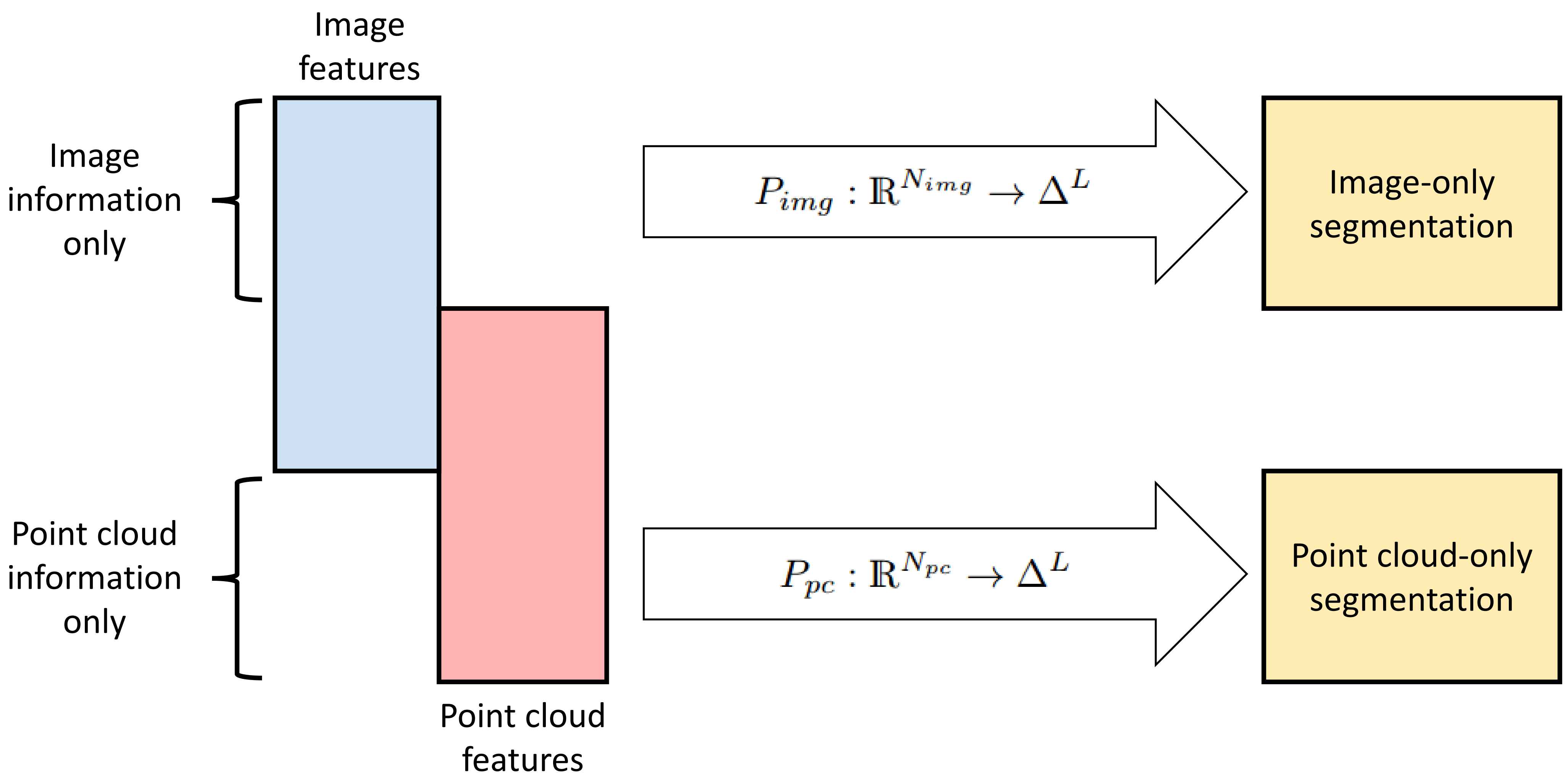
Type	Name	Dim	Low	High
Size	Length proxy - λ_1	1	✓	✓
	Area proxy - $\sqrt{\lambda_1 \lambda_2}$	1	✓	✓
	Volume proxy - $\sqrt[3]{\lambda_1 \lambda_2 \lambda_3}$	1	✓	✓
Shape	Scatter - λ_3 / Λ	1	✓	✓
	Planarity - $(\lambda_2 - \lambda_3)/\Lambda$	1	✓	✓
	Linearity - $(\lambda_1 - \lambda_2)/\Lambda$	1	✓	✓
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

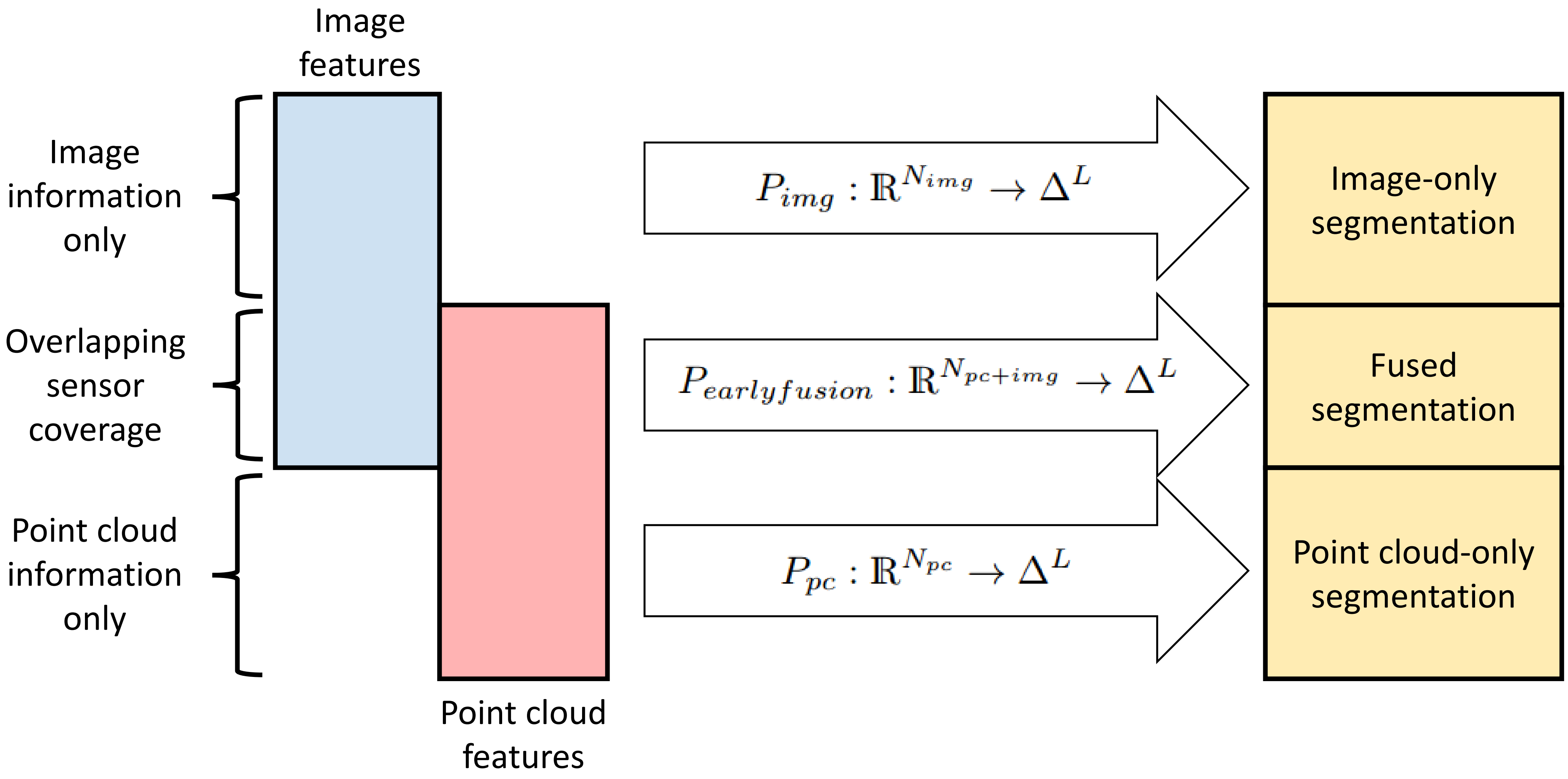
Early Fusion



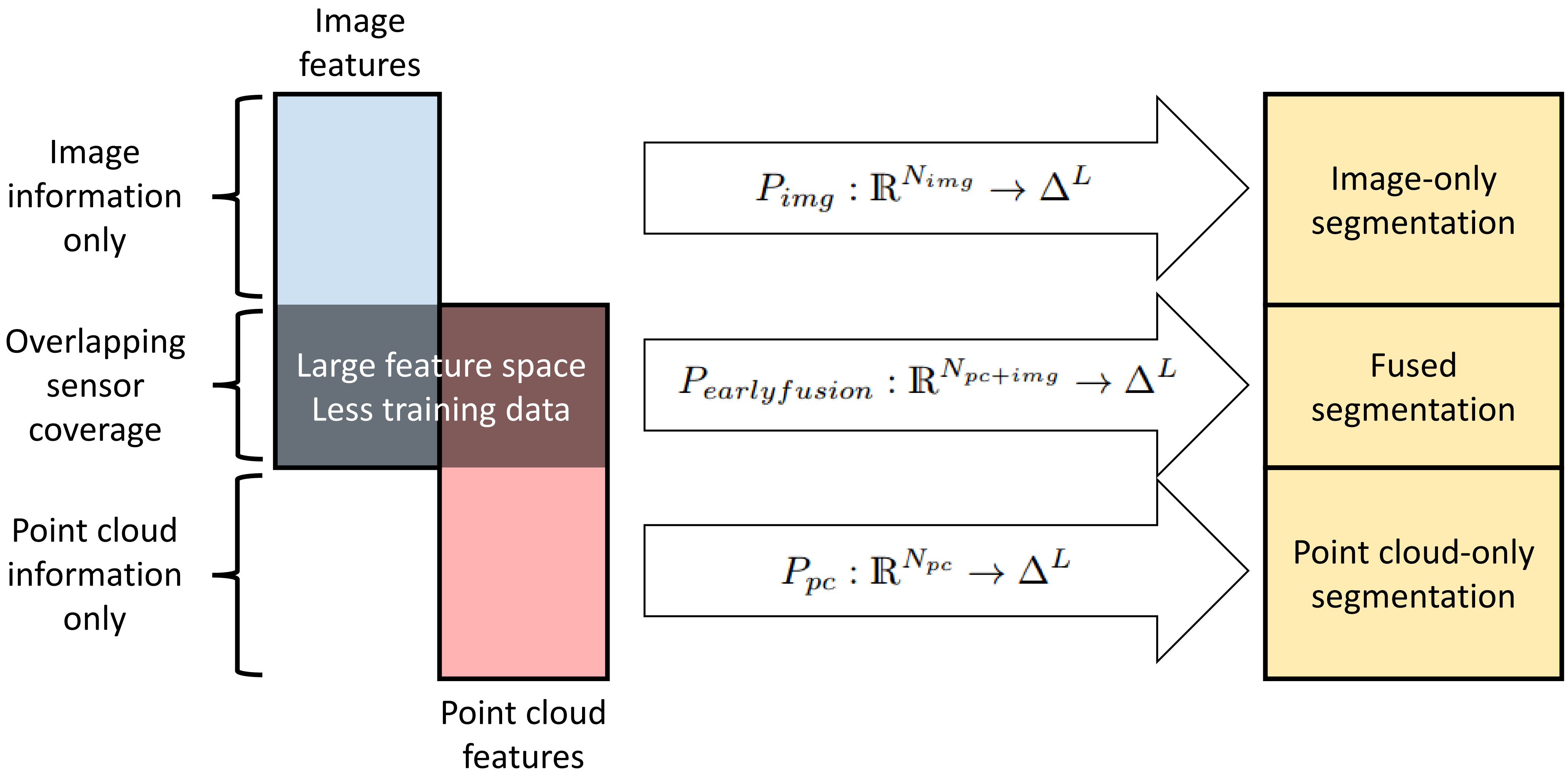
Early Fusion



Early Fusion



Early Fusion



Late Fusion

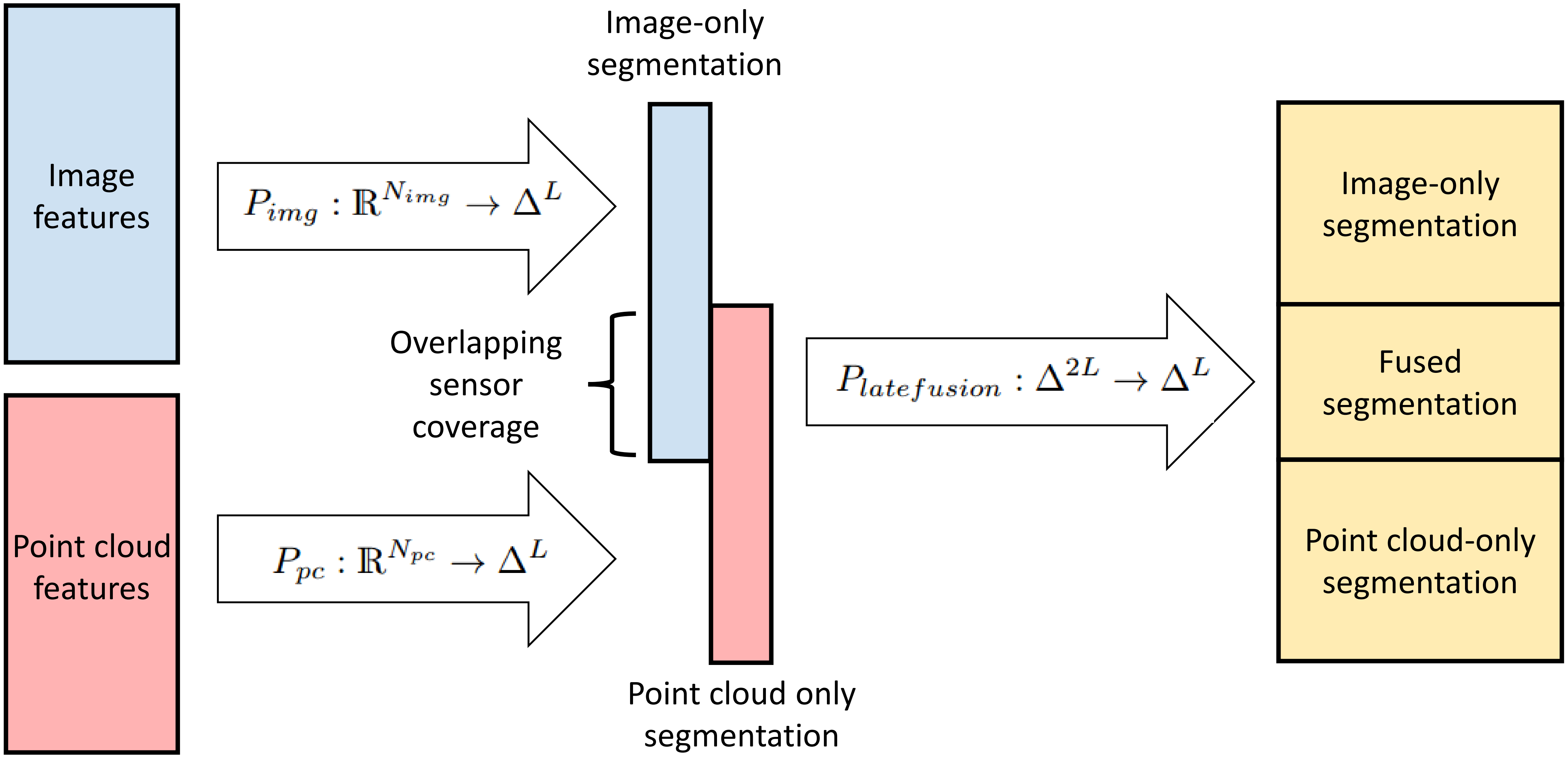


Image features
Point cloud features
$P_{img} : \mathbb{R}^{N_{img}} \rightarrow \Delta^L$
$P_{pc} : \mathbb{R}^{N_{pc}} \rightarrow \Delta^L$
Image-only segmentation
Overlapping sensor coverage
Point cloud only segmentation
$P_{latefusion} : \Delta^{2L} \rightarrow \Delta^L$
Image-only segmentation
Fused segmentation
Point cloud-only segmentation

Late Fusion

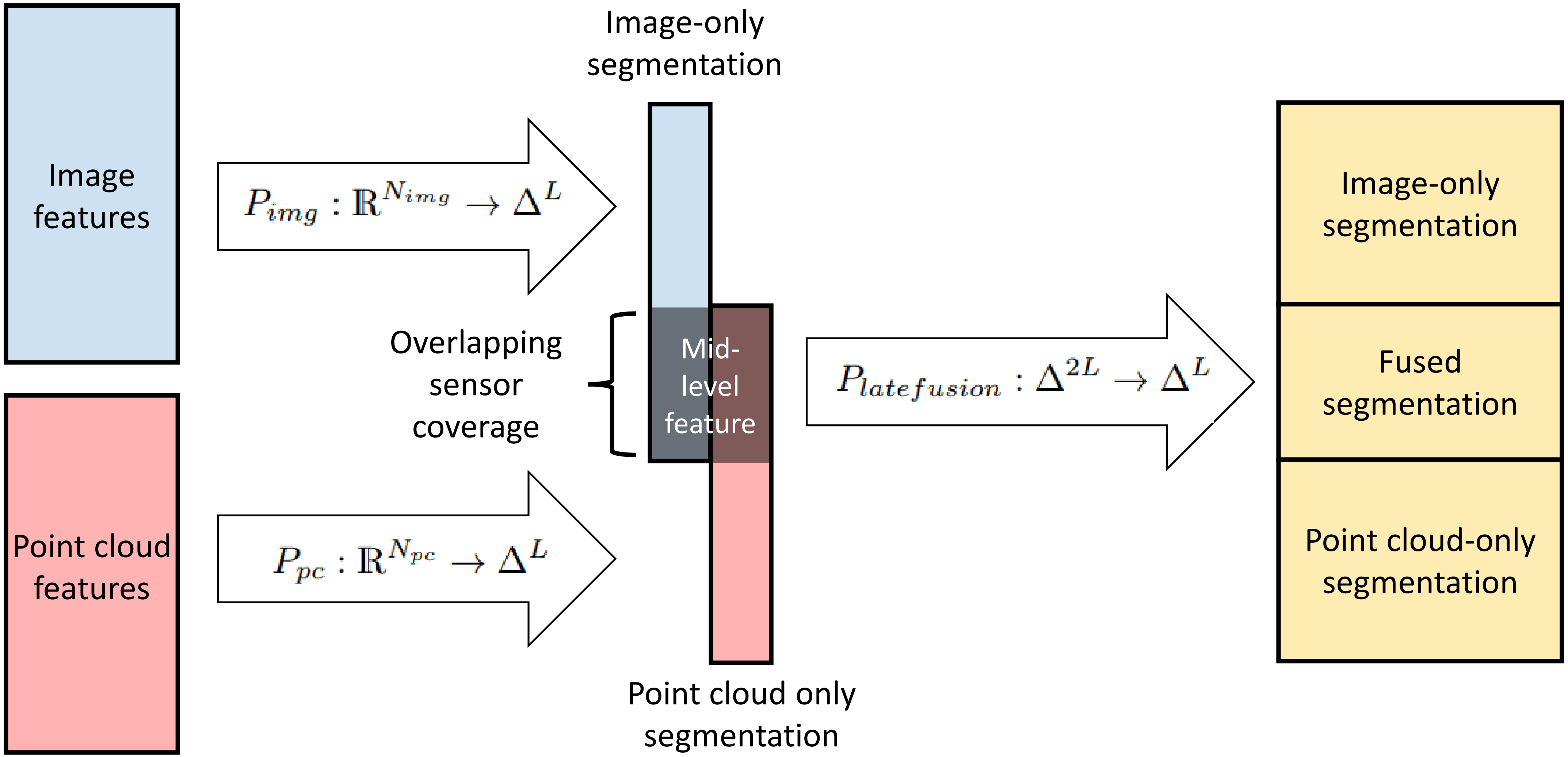
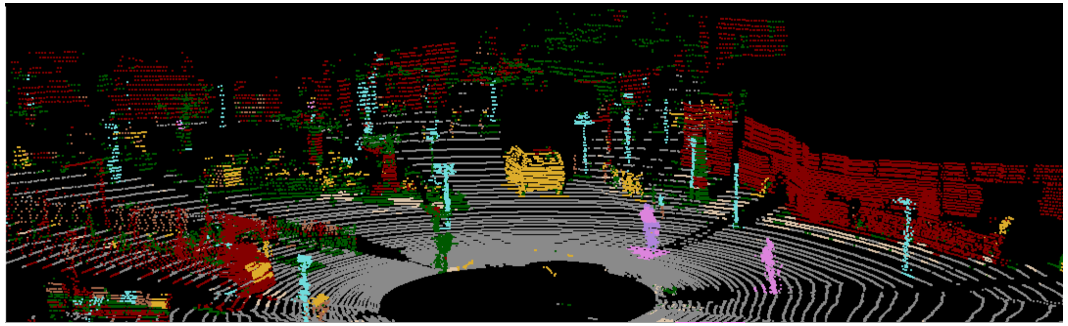


Image features
Point cloud features
$P_{img} : \mathbb{R}^{N_{img}} \rightarrow \Delta^L$
$P_{pc} : \mathbb{R}^{N_{pc}} \rightarrow \Delta^L$
Image-only segmentation
Overlapping sensor coverage
Mid-level feature
Point cloud only segmentation
$P_{latefusion} : \Delta^{2L} \rightarrow \Delta^L$
Image-only segmentation
Fused segmentation
Point cloud-only segmentation

Fusion

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



Point cloud unimodal segmentation

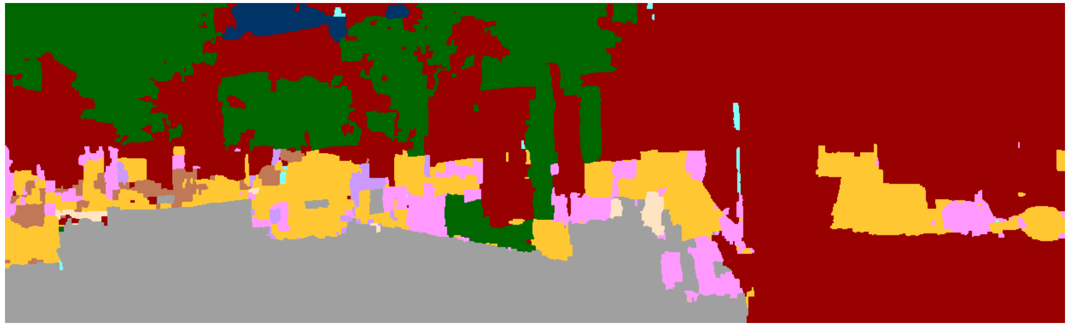


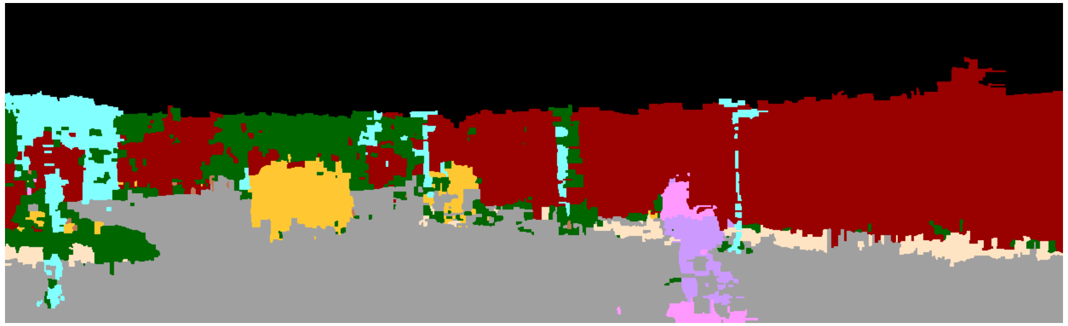
Image unimodal segmentation

Ground Truth

Fusion

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

Ground Truth



Point cloud unimodal segmentation
(projected onto image)

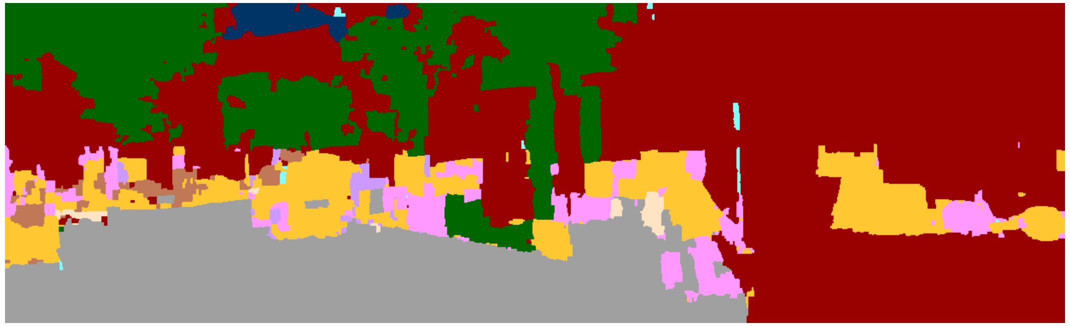
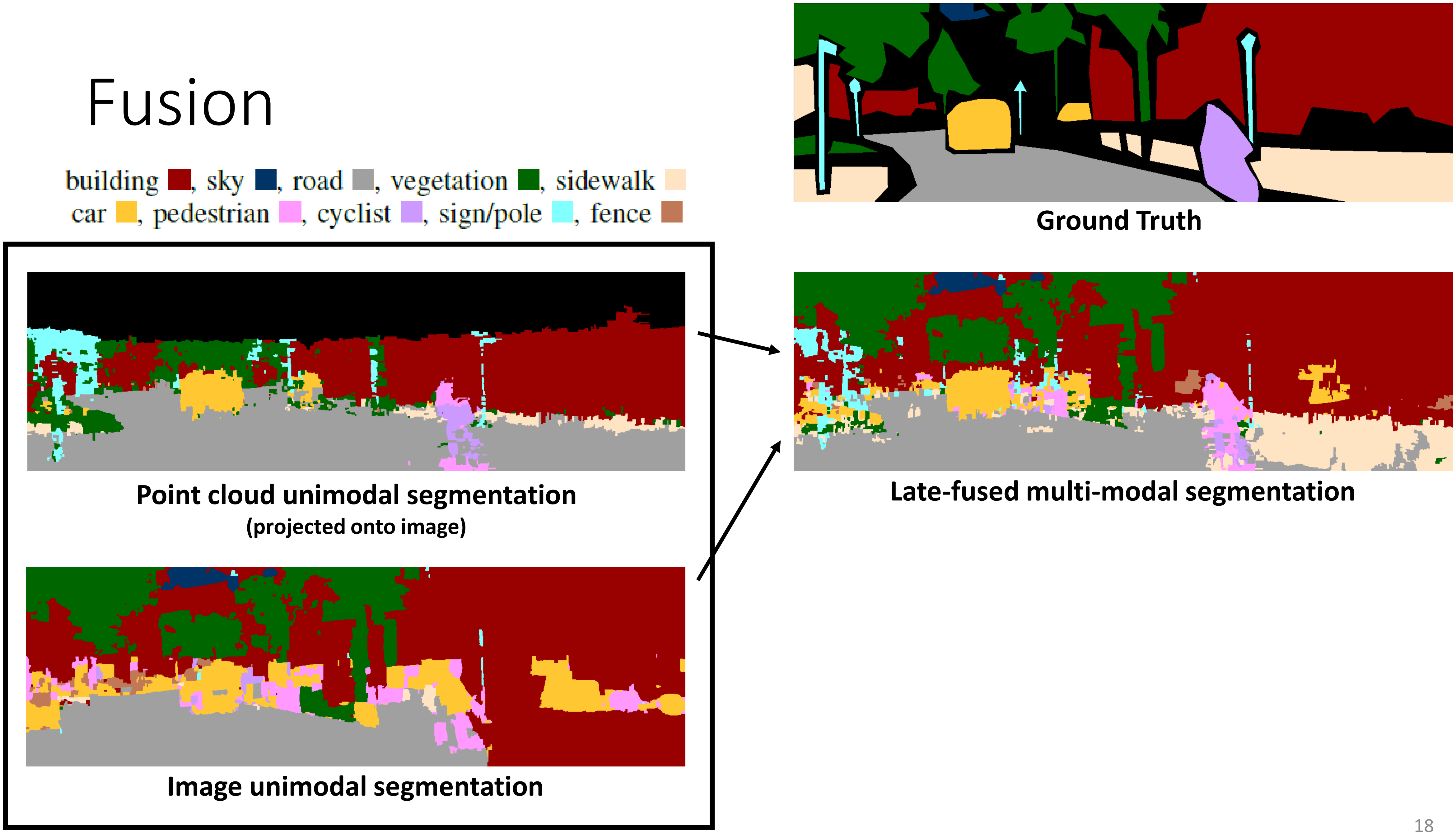


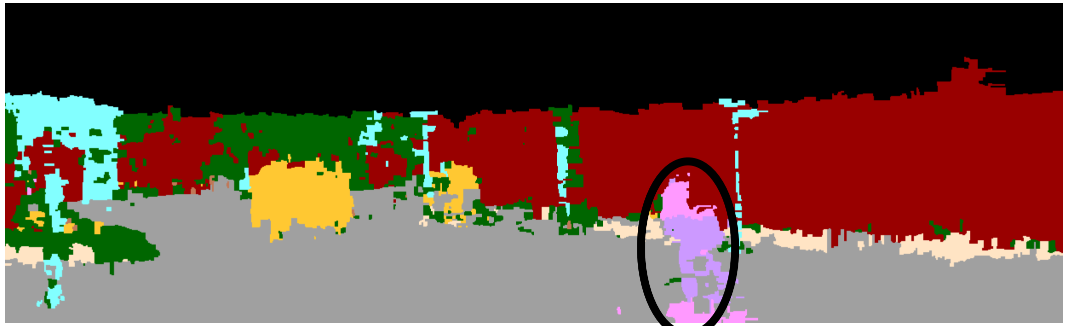
Image unimodal segmentation

Fusion


building, sky, road, vegetation, sidewalk
car, pedestrian, cyclist, sign/pole, fence
Ground Truth
Point cloud unimodal segmentation
(projected onto image)
Late-fused multi-modal segmentation
Image unimodal segmentation

Fusion

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



Point cloud unimodal segmentation
(projected onto image)

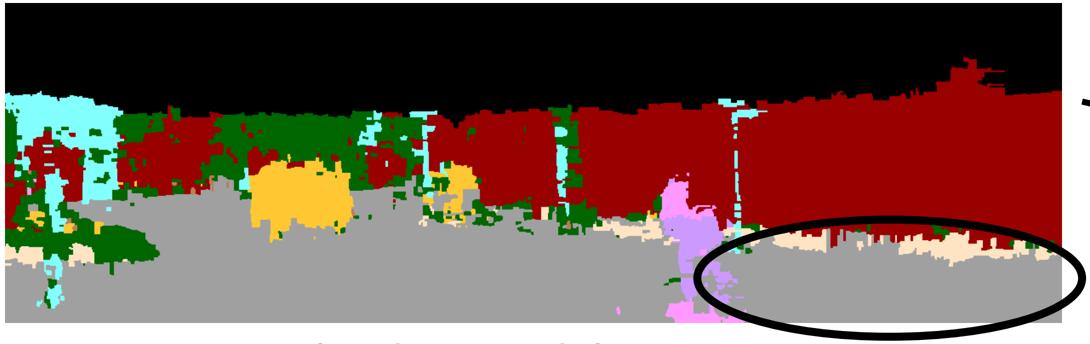
Image unimodal segmentation

Ground Truth

Late-fused multi-modal segmentation

Fusion

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



Point cloud unimodal segmentation
(projected onto image)

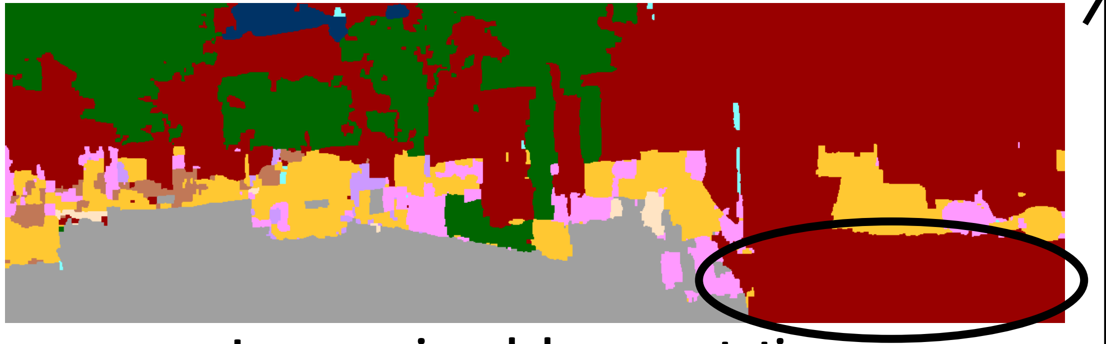


Image unimodal segmentation

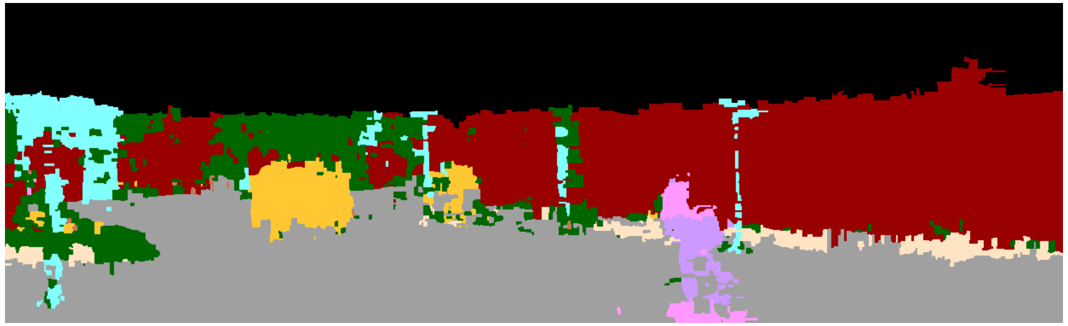
Ground Truth

Late-fused multi-modal segmentation

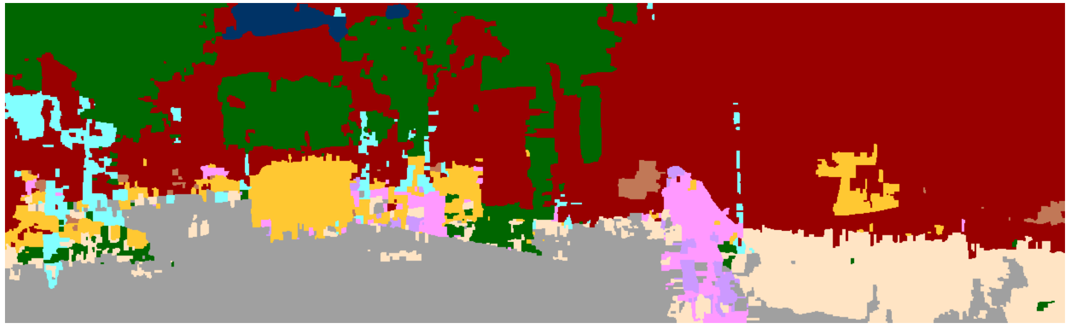
Fusion

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

Ground Truth



Point cloud unimodal segmentation
(projected onto image)



Late-fused multi-modal segmentation

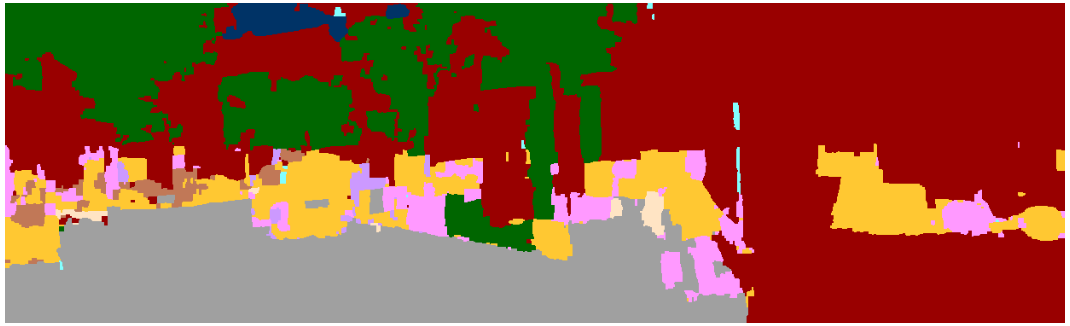
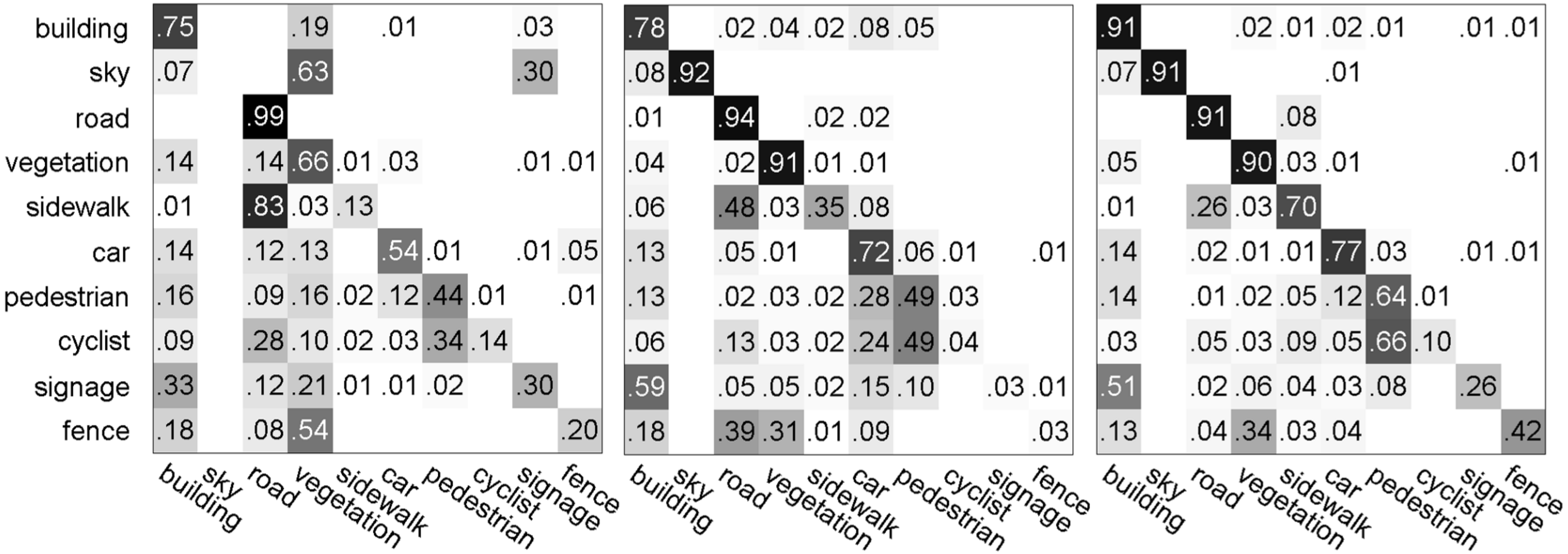


Image unimodal segmentation

Fused segmentation & Post-processing
Conditional Random Field (CRF)

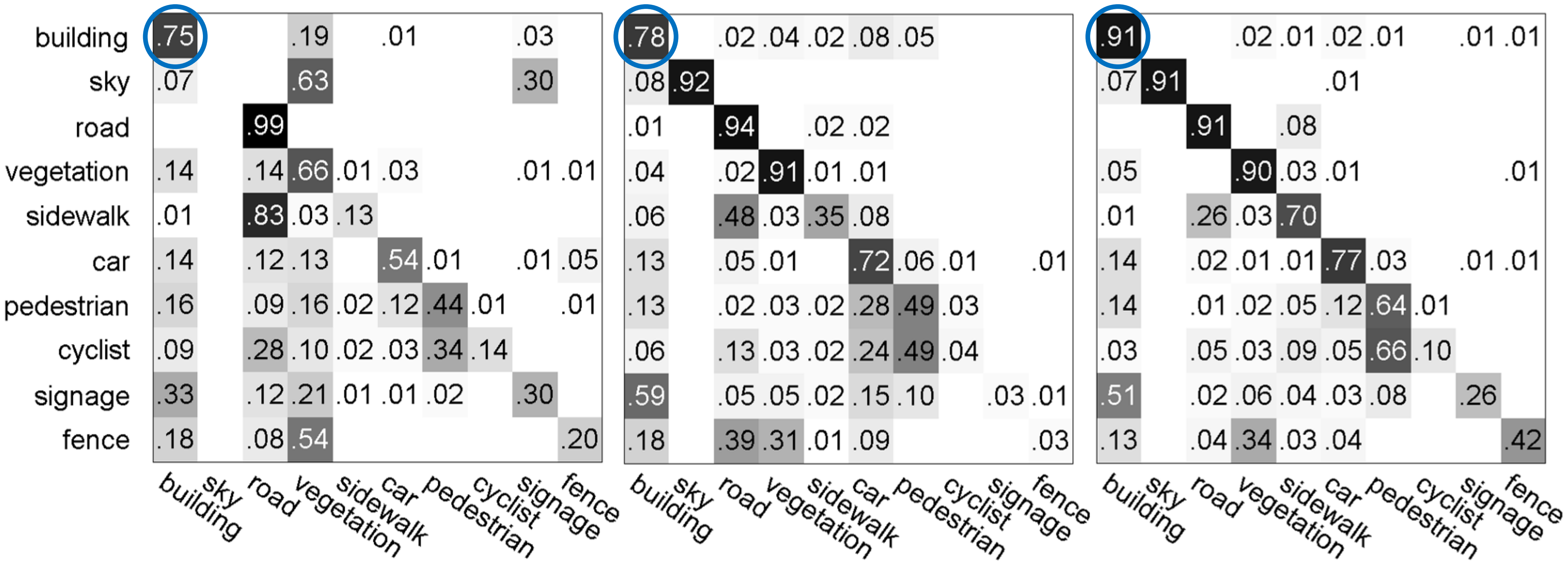
Fusion



PC only **Image only** **Fused**

Performance on overlapping region

Fusion



PC only **Image only** **Fused**

Performance on overlapping region

Fusion

	building	sky	road	vegetation	sidewalk	car	pedestrian	cyclist	signage	fence
building	.75			.19		.01			.03	
sky	.07			.63					.30	
road			.99							
vegetation	.14		.14	.66	.01	.03			.01	.01
sidewalk	.01		.83	.03	.13					
car	.14		.12	.13		.54	.01		.01	.05
pedestrian	.16		.09	.16	.02	.12	.44	.01		.01
cyclist	.09		.28	.10	.02	.03	.34	.14		
signage	.33		.12	.21	.01	.01	.02		.30	
fence	.18		.08	.54						.20

PC only

	building	sky	road	vegetation	sidewalk	car	pedestrian	cyclist	signage	fence
building	.78		.02	.04	.02	.08	.05			
sky	.08	.92								
road	.01		.94		.02	.02				
vegetation	.04		.02	.91	.01	.01				
sidewalk	.06		.48	.03	.35	.08				
car	.13		.05	.01		.72	.06	.01		.01
pedestrian	.13		.02	.03	.02	.28	.49	.03		
cyclist	.06		.13	.03	.02	.24	.49	.04		
signage	.59		.05	.05	.02	.15	.10		.03	.01
fence	.18		.39	.31	.01	.09				.03

Image only

	building	sky	road	vegetation	sidewalk	car	pedestrian	cyclist	signage	fence
building	.91			.02	.01	.02	.01		.01	.01
sky	.07	.91				.01				
road			.91		.08					
vegetation	.05			.90	.03	.01				.01
sidewalk	.01		.26	.03	.70					
car	.14		.02	.01	.01	.77	.03		.01	.01
pedestrian	.14		.01	.02	.05	.12	.64	.01		
cyclist	.03		.05	.03	.09	.05	.66	.10		
signage	.51		.02	.06	.04	.03	.08		.26	
fence	.13		.04	.34	.03	.04				.42

Fused

Performance on overlapping region

Fusion

	building	sky	road	vegetation	sidewalk	car	pedestrian	cyclist	signage	fence
building	.75			.19		.01			.03	
sky	.07			.63					.30	
road			.99							
vegetation	.14		.14	.66	.01	.03			.01	.01
sidewalk	.01		.83	.03	.13					
car	.14		.12	.13		.54	.01		.01	.05
pedestrian	.16		.09	.16	.02	.12	.44	.01		.01
cyclist	.09		.28	.10	.02	.03	.34	.14		
signage	.33		.12	.21	.01	.01	.02		.30	
fence	.18		.08	.54						.20

PC only

	building	sky	road	vegetation	sidewalk	car	pedestrian	cyclist	signage	fence
building	.78		.02	.04	.02	.08	.05			
sky	.08	.92								
road	.01		.94		.02	.02				
vegetation	.04		.02	.91	.01	.01				
sidewalk	.06		.48	.03	.35	.08				
car	.13		.05	.01		.72	.06	.01		.01
pedestrian	.13		.02	.03	.02	.28	.49	.03		
cyclist	.06		.13	.03	.02	.24	.49	.04		
signage	.59		.05	.05	.02	.15	.10		.03	.01
fence	.18		.39	.31	.01	.09				.03

Image only

	building	sky	road	vegetation	sidewalk	car	pedestrian	cyclist	signage	fence
building	.91			.02	.01	.02	.01		.01	.01
sky	.07	.91				.01				
road			.91		.08					
vegetation	.05			.90	.03	.01				.01
sidewalk	.01		.26	.03	.70					
car	.14		.02	.01	.01	.77	.03		.01	.01
pedestrian	.14		.01	.02	.05	.12	.64	.01		
cyclist	.03		.05	.03	.09	.05	.66	.10		
signage	.51		.02	.06	.04	.03	.08		.26	
fence	.13		.04	.34	.03	.04				.42

Fused

Performance on overlapping region

Results

- Quantitative Results
  - Train and test on KITTI[1] dataset, augmented with additional annotations
  - 252 images across 8 sequences
    - 140 images for training, 112 for testing

	glob	class	bldg	sky	road	veg	side	car	ped	cycl	sgn	fnc
Cadena *et al.* [3]	84.1%	52.4%	92.5%	**95.7%**	92.5%	86.3%	51.5%	67.9%	28.6%	4.0%	2.5%	2.3%

[1]Geiger, et al. KITTI. CVPR 12

Results

- Quantitative Results
  - Train and test on KITTI[1] dataset, augmented with additional annotations
  - 252 images across 8 sequences
    - 140 images for training, 112 for testing

	glob	class	bldg	sky	road	veg	side	car	ped	cycl	sgn	fnc
Cadena *et al.* [3]	84.1%	52.4%	92.5%	**95.7%**	92.5%	86.3%	51.5%	67.9%	28.6%	4.0%	2.5%	2.3%
Ours (image only)	83.5%	53.3%	87.5%	92.5%	94.5%	92.5%	34.5%	71.4%	49.0%	3.6%	4.1%	3.3%

[1]Geiger, et al. KITTI. CVPR 12

Results

- Quantitative Results
  - Train and test on KITTI[1] dataset, augmented with additional annotations
  - 252 images across 8 sequences
    - 140 images for training, 112 for testing

	glob	class	bldg	sky	road	veg	side	car	ped	cycl	sgn	fnc
Cadena *et al.* [3]	84.1%	52.4%	92.5%	**95.7%**	92.5%	86.3%	51.5%	67.9%	28.6%	4.0%	2.5%	2.3%
Ours (image only)	83.5%	53.3%	87.5%	92.5%	94.5%	92.5%	34.5%	71.4%	49.0%	3.6%	4.1%	3.3%
Ours (late fused)	88.0%	64.8%	93.5%	92.5%	91.2%	92.0%	69.7%	76.5%	63.7%	**10.0%**	**16.6%**	42.2%

[1]Geiger, et al. KITTI. CVPR 12

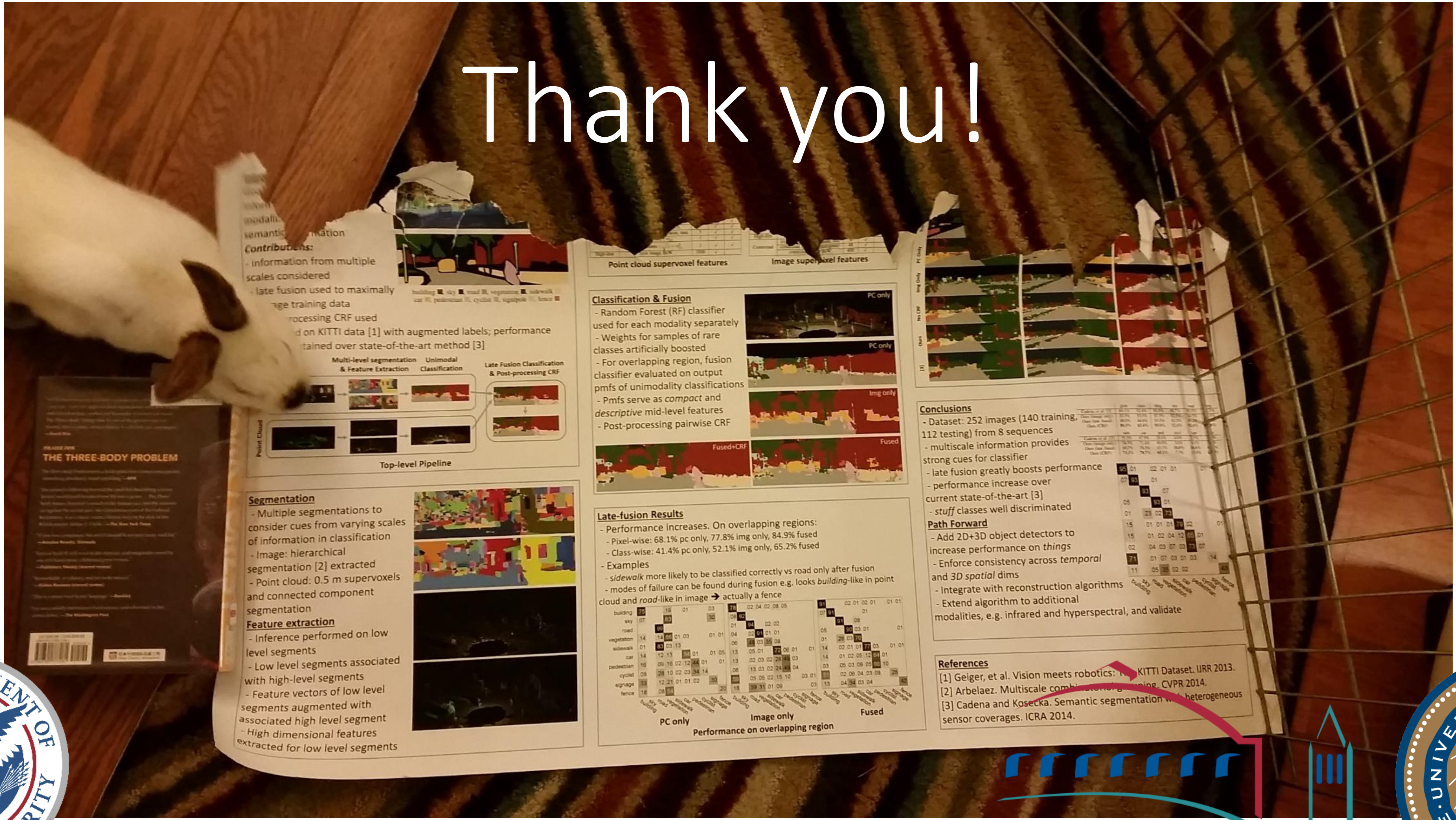
Results

- Quantitative Results
  - Train and test on KITTI[1] dataset, augmented with additional annotations
  - 252 images across 8 sequences
    - 140 images for training, 112 for testing

	glob	class	bldg	sky	road	veg	side	car	ped	cycl	sgn	fnc
Cadena *et al.* [3]	84.1%	52.4%	92.5%	**95.7%**	92.5%	86.3%	51.5%	67.9%	28.6%	4.0%	2.5%	2.3%
Ours (image only)	83.5%	53.3%	87.5%	92.5%	94.5%	92.5%	34.5%	71.4%	49.0%	3.6%	4.1%	3.3%
Ours (late fused)	88.0%	64.8%	93.5%	92.5%	91.2%	92.0%	69.7%	76.5%	63.7%	**10.0%**	**16.6%**	42.2%
Ours (CRF)	**89.3%**	**65.4%**	**95.0%**	92.6%	**92.6%**	**92.8%**	**73.3%**	**78.7%**	**65.1%**	7.3%	13.8%	**43.2%**

[1]Geiger, et al. KITTI. CVPR 12

Thank you!



Contributions:
- information from multiple scales considered
- late fusion used to maximally
processing CRF used
on KITTI data [1] with augmented labels; performance
tained over state-of-the-art method [3]
Multi-level segmentation & Feature Extraction
Unimodal Classification
Late Fusion Classification & Post-processing CRF
Point Cloud
Top-level Pipeline
Point cloud supervoxel features
Image super pixel 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
PC only
Img only
Fused+CRF
Fused
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
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
PC only
Image only
Fused
Performance on overlapping region
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
References
[1] Geiger, et al. Vision meets robotics: KITTI Dataset. IJRR 2013.
[2] Arbelaez. Multiscale comb ning. CVPR 2014.
[3] Cadena and Kosecka. Semantic segmentation heterogeneous sensor coverages. ICRA 2014.
THE THREE-BODY PROBLEM