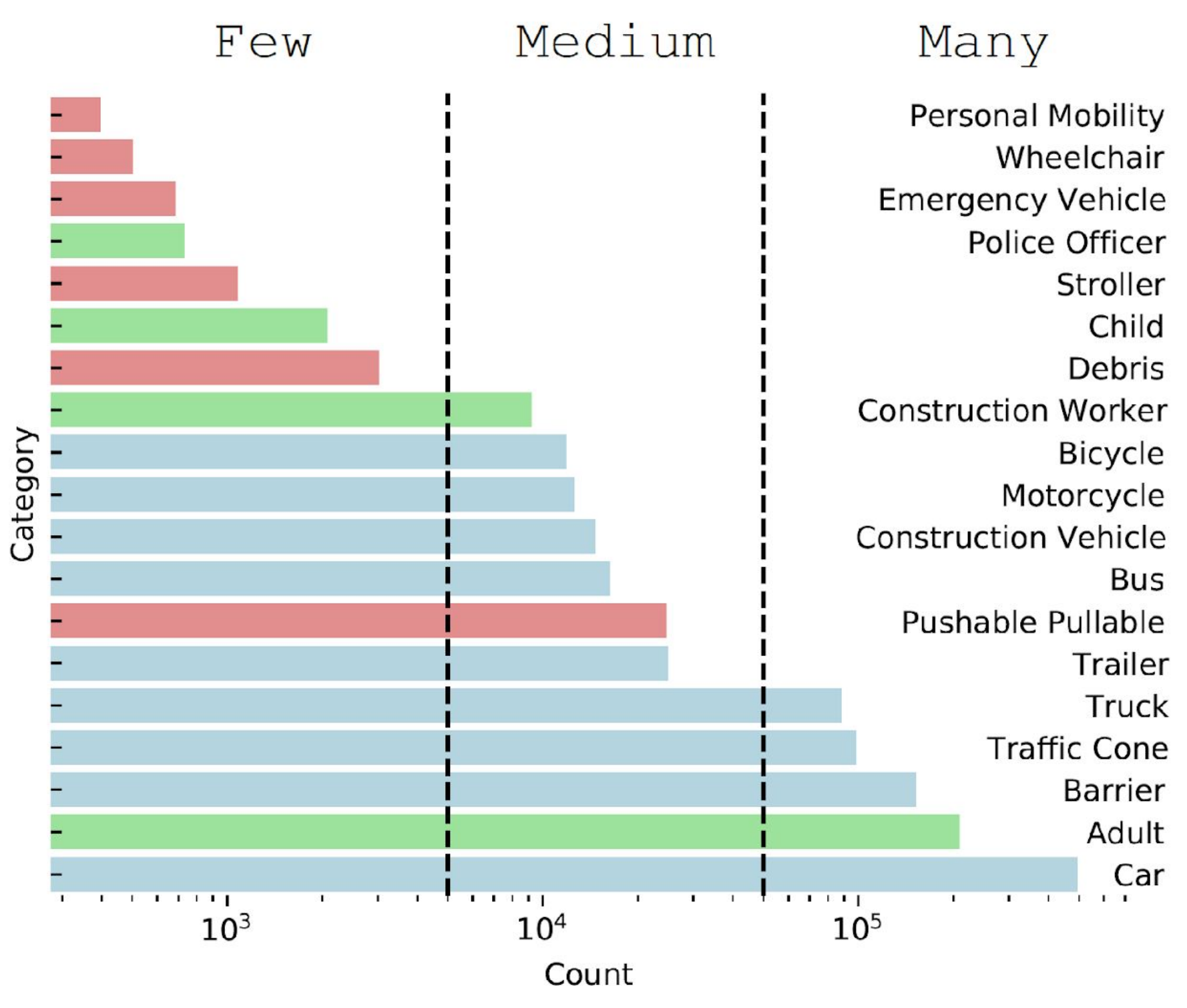
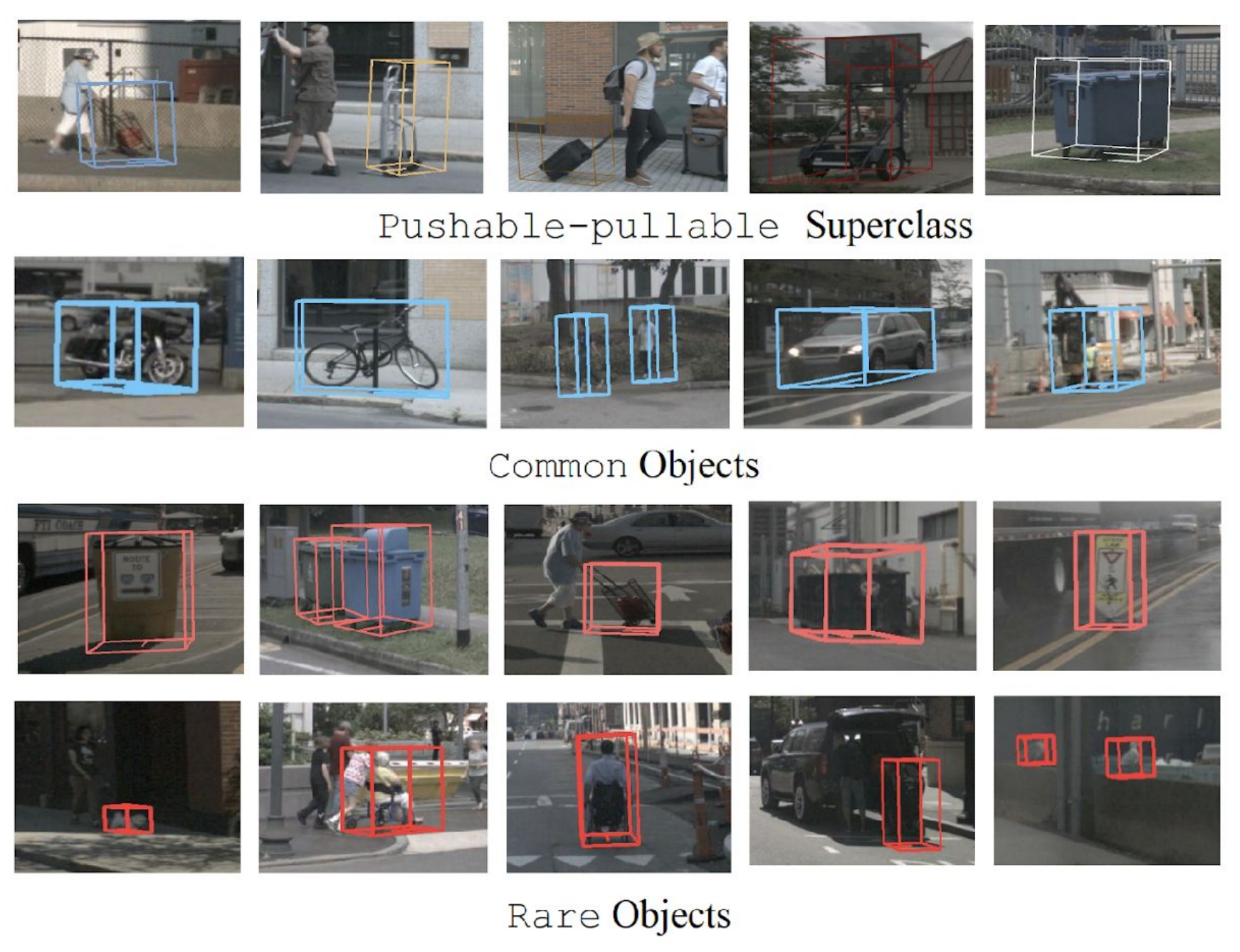
Towards Long-Tailed 3D Detection

Neehar Peri, Achal Dave, Shu Kong, Deva Ramanan

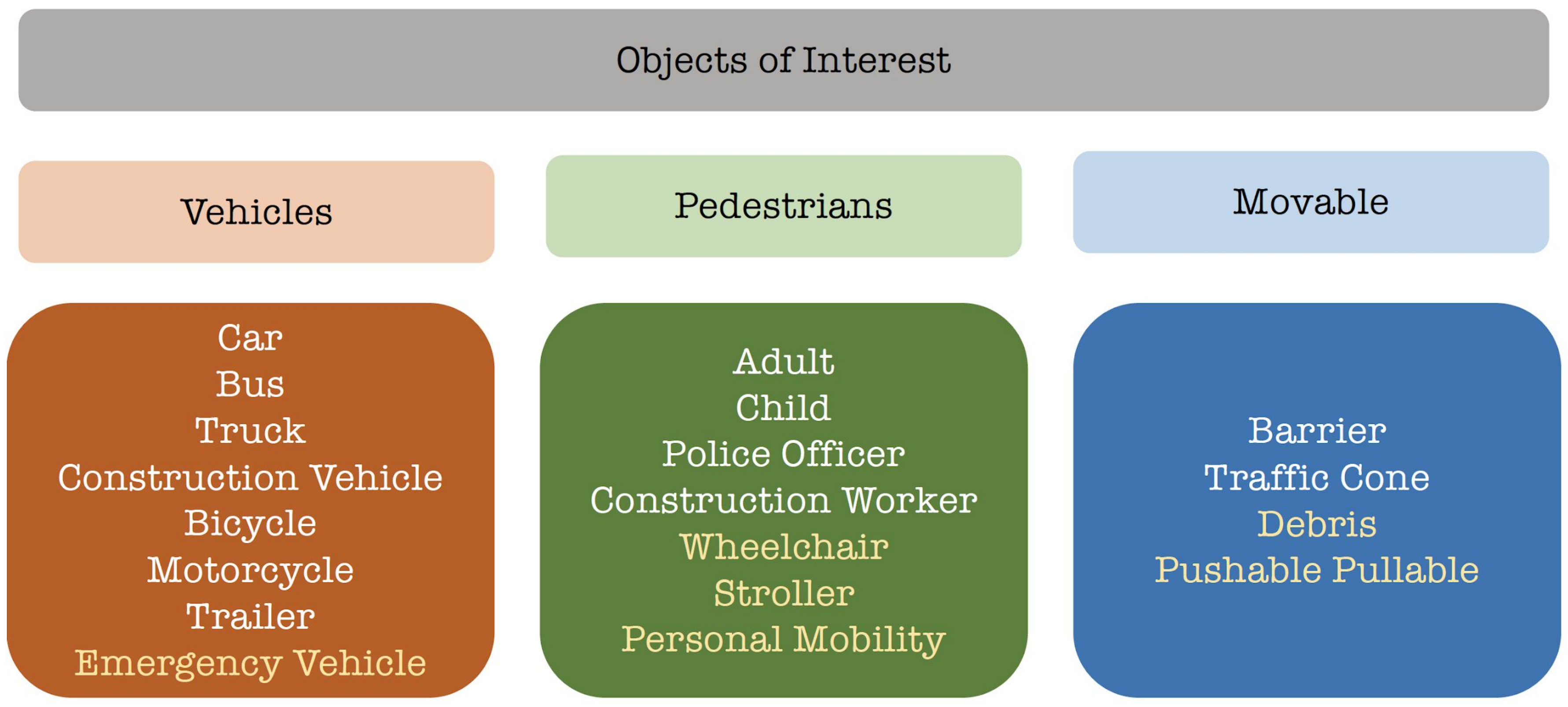
Long Tailed 3D Detection





- Standard benchmarks ignore `rare` classes (e.g. `stroller`)
- Vulnerable classes (e.g. `child` and `construction worker`) are grouped into the `pedestrian` superclass

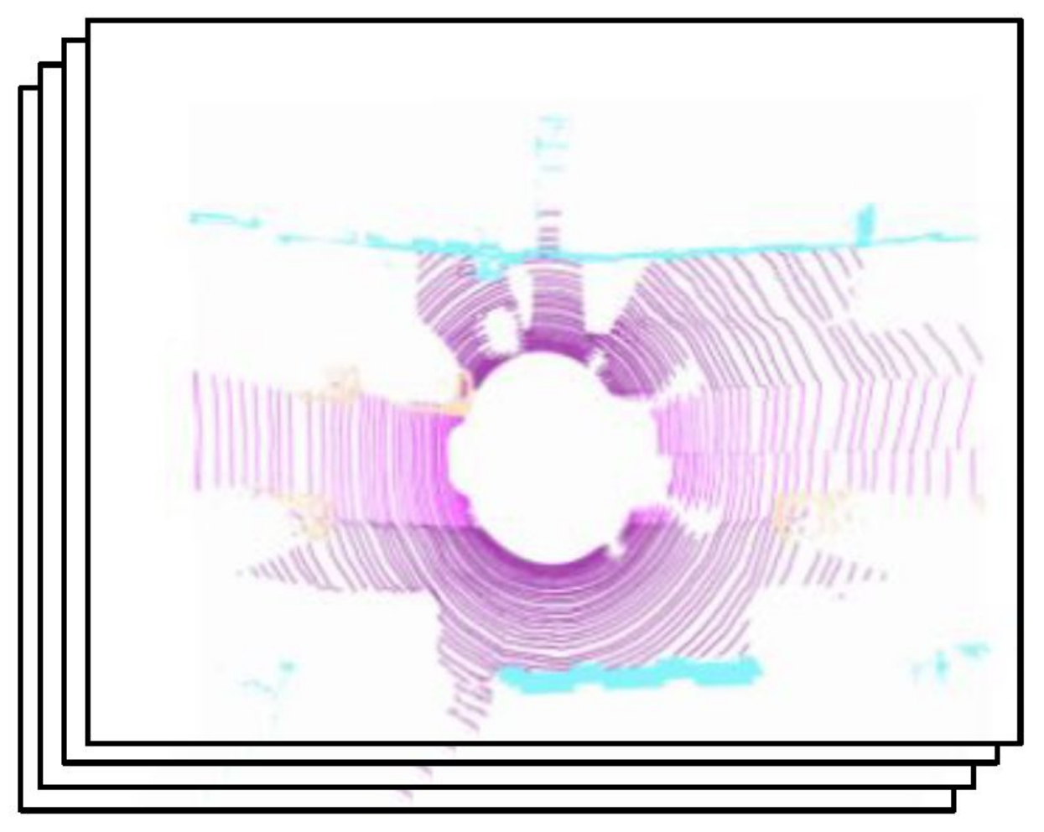
Semantic Hierarchy



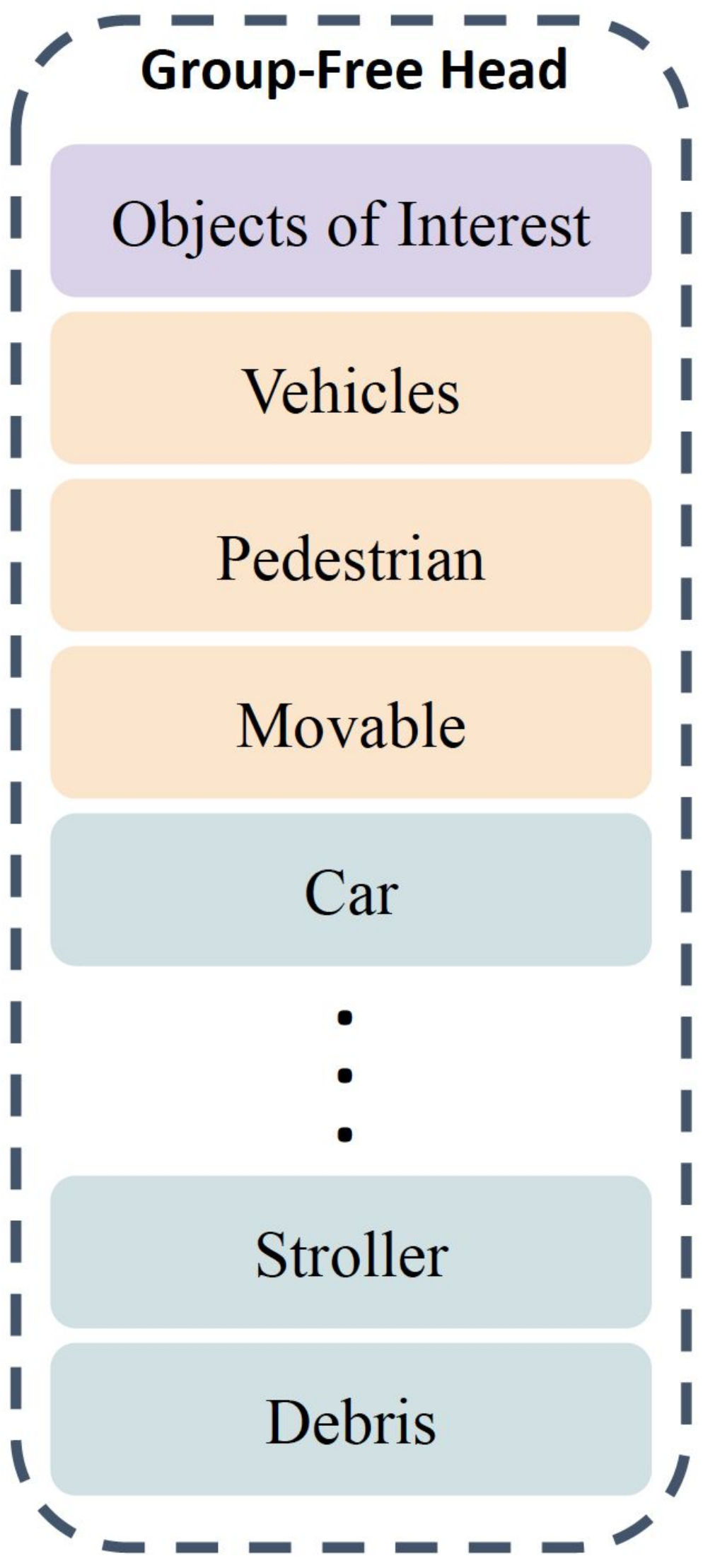
- nuScenes organizes all classes with a semantic hierarchy
- Hierarchical structure has been historically ignored, leading to missed opportunities for innovation

Group-Free Head Architecture

- Simplified architecture makes it easier to add new, diverse classes



Voxel Backbone



- Detector can predict non-exclusive classes from the semantic hierarchy (e.g. `object`, `vehicle`, `car`)

Hierarchical AP

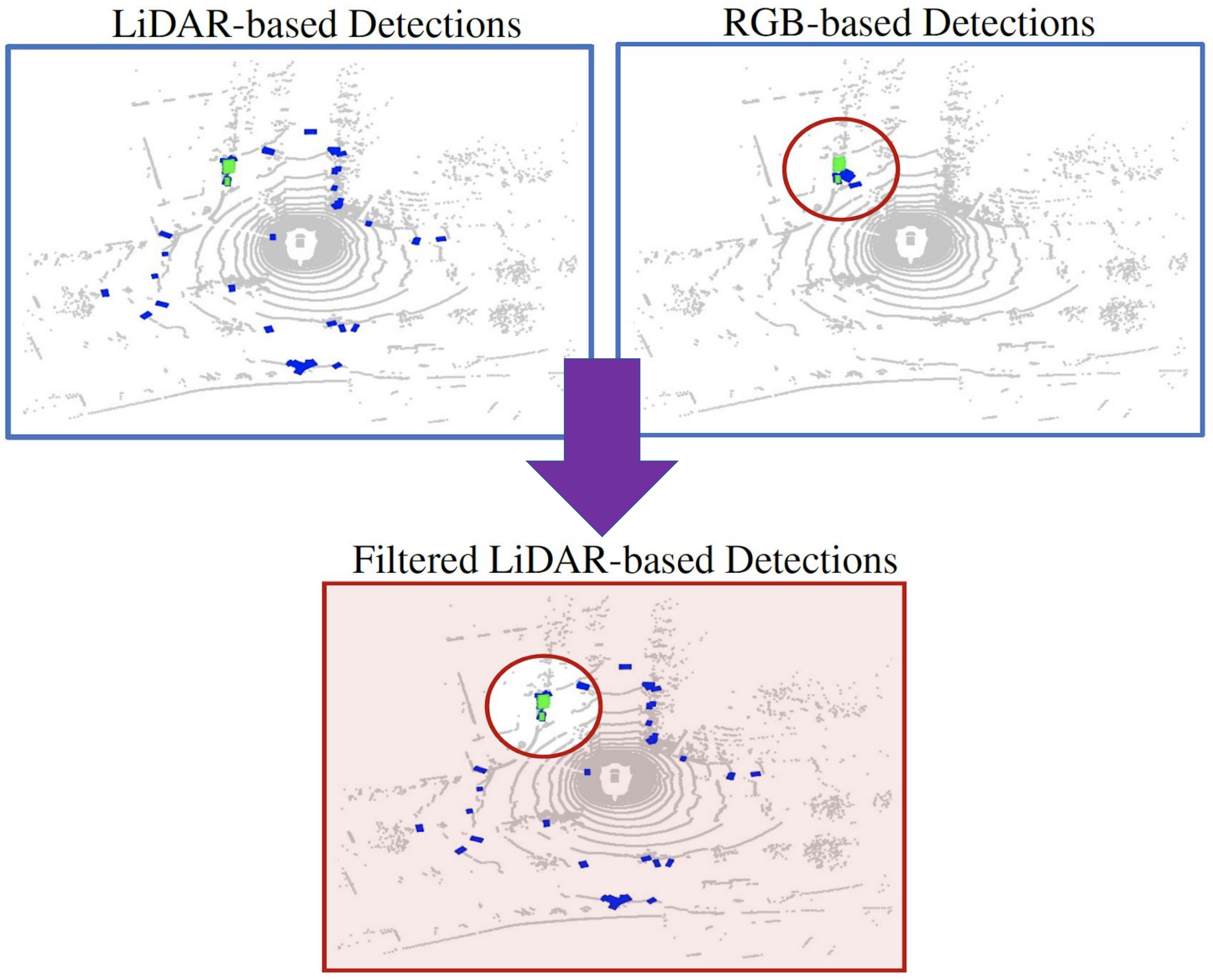
`LCA = 0`
- Identical to standard AP metric

`LCA = 1`
- Partial credit for mistaking sibling classes (i.e. mistaking `child` for `adult`)

`LCA = 2`
- Partial credit for mistaking any two classes (i.e. mistaking `child` for `traffic-cone`)

Method	mAP_H	Car	Adult	Truck	CV	Bicycle	MC	Child	CW	Stroller	PP
CenterPoint	LCA=0	86.5	84.0	53.9	23.5	47.2	60.2	0.1	20.2	3.6	32.2
	LCA=1	87.3	84.7	59.5	25.2	48.8	61.7	0.1	26.4	3.8	32.4
	LCA=2	87.3	85.0	59.6	25.3	49.5	62.1	0.1	27.2	4.0	32.9
CenterPoint *w/ Hierarchy*	LCA=0	88.6	86.9	63.4	25.7	50.2	63.2	0.1	25.3	8.7	36.8
	LCA=1	89.5	87.6	72.4	27.5	52.2	65.2	0.1	32.4	9.4	37.0
	LCA=2	89.6	88.0	72.5	27.7	53.2	65.7	0.1	34.0	9.8	37.6
CenterPoint *w/ Hier. & Filtering*	LCA=0	88.5	86.6	63.4	29.0	58.5	68.2	5.3	35.8	31.6	39.3
	LCA=1	89.4	87.4	72.4	31.3	61.2	69.7	15.2	52.0	37.7	39.4
	LCA=2	89.5	87.7	72.5	31.5	62.3	69.9	16.9	56.3	38.8	39.8
TransFusion	LCA=0	84.4	84.5	58.5	15.1	44.9	57.2	1.0	15.1	3.2	19.6
	LCA=1	85.5	85.7	67.4	21.8	46.7	59.1	1.6	21.8	3.7	19.8
	LCA=2	85.5	86.1	67.5	22.6	47.7	59.9	1.7	22.6	4.2	20.4
TransFusion *w/ Camera*	LCA=0	84.4	84.2	58.4	24.5	46.7	60.8	3.1	21.6	13.3	25.3
	LCA=1	86.0	85.4	67.3	26.3	50.1	63.5	14.4	34.7	20.6	25.6
	LCA=2	86.0	85.9	67.4	26.8	52.2	65.1	15.2	36.1	22.8	26.4
TransFusion *w/ Cam. & Filtering*	LCA=0	84.4	84.2	58.4	25.3	52.3	62.8	4.0	27.5	14.7	27.3
	LCA=1	86.0	85.4	67.3	26.6	55.7	64.0	25.1	46.7	24.3	27.4
	LCA=2	86.0	85.9	67.4	27.0	56.9	64.3	25.8	48.6	28.3	27.9

Multimodal Fusion



- LiDAR-only detectors are accurate w.r.t 3D localization and yield high recall (though classification is poor)
- RGB-only detectors are accurate w.r.t recognition (though 3D localization is poor)

Keep LiDAR-based detections that are nearby (i.e. within *m* meters) RGB-based detections. Discard all other detections.

Experimental Results

Method	Multimodal	Many	Medium	Few	All
FCOS3D (RGB-only) [40]		39.0	23.3	2.9	20.9
PointPillars (LiDAR-only) [8]		64.2	28.4	3.4	30.0
+ Hierarchy		**66.4**	30.4	2.9	31.2
w/ Data Aug.		54.4	24.2	1.8	25.1
w/ Filtering	✓	66.2	**41.0**	**4.4**	**35.8**
CBGS (LiDAR-only) [9]		47.2	10.4	**0.1**	17.2
+ Hierarchy		**49.5**	11.1	**0.1**	18.1
w/ Data Aug.		49.9	17.1	**0.1**	20.6
w/ Filtering	✓	48.0	**20.3**	**0.1**	**21.5**
CenterPoint (LiDAR-only) [13]		73.7	41.3	3.0	37.5
+ Hierarchy		**77.1**	45.1	4.3	40.4
w/ Data Aug.		73.8	44.5	7.4	40.3
w/ Filtering	✓	**77.1**	**49.0**	**9.4**	**43.6**
MVP [22]	✓	65.6	31.6	1.5	31.0
+ Hierarchy	✓	67.0	33.0	0.1	31.6
w/ Data Aug.	✓	65.9	35.8	0.1	32.5
w/ Filtering	✓	**67.1**	**39.2**	**1.6**	**34.4**
TransFusion [17]		68.5	**42.8**	8.4	38.5
+ Camera	✓	**73.9**	41.2	**9.8**	39.8
w/ Data Aug.	✓	73.4	40.9	8.2	39.0
w/ Filtering	✓	**73.9**	42.5	9.1	**40.1**

Full paper & code are available at github.com/neeharperi/LT3D