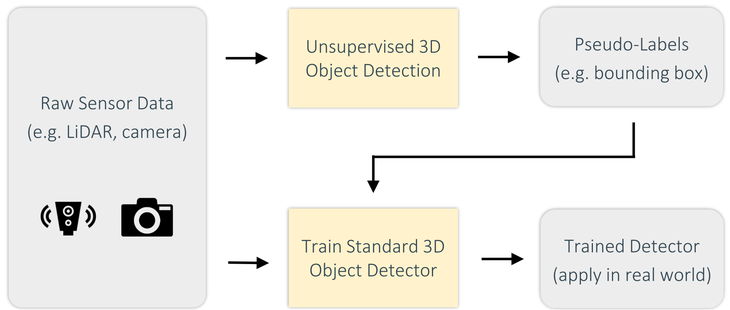
UNION: Unsupervised 3D Object Detection using Object Appearance-based Pseudo-Classes

Ted Lentsch, Holger Caesar, and Dariu M. Gavrila | Delft University of Technology

Unsupervised 3D object detection

- Goal: Discover mobile objects (e.g. vehicles, pedestrians, cyclists)
- Data: Unlabeled LiDAR point clouds and camera images (i.e. raw data)
- Task: Generate pseudo-labels and train standard 3D object detector



Contributions

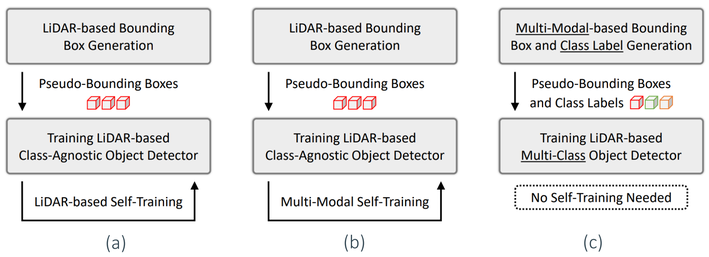
We propose UNION, which sets the new SOTA on nuScenes [1].

1) First to use camera, LiDAR, and temporal information *jointly*.
2) Reduce training complexity and time by *avoiding* iterative training protocols.
3) Extend 3D object discovery to *multi-class* 3D object detection.

Comparison with existing methods

In contrast to existing work, we use multi-modal data to generate pseudo-bounding boxes and labels for training detectors and we do not need self-training.

a) LiDAR 3D object discovery with LiDAR-based self-training [3][4]
b) LiDAR 3D object discovery with multi-modal self-training [7]
c) UNION: multi-modal multi-class 3D object discovery (ours)

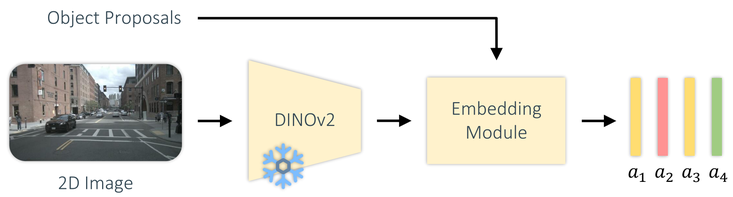


UNION pipeline

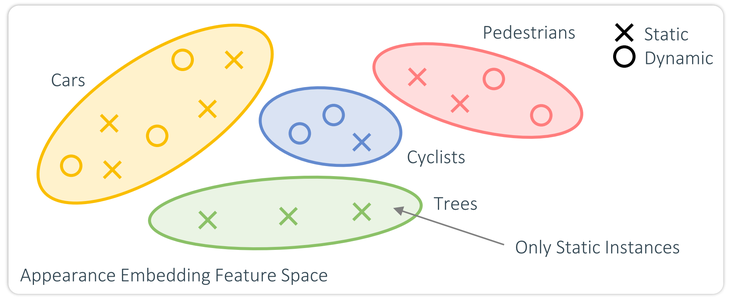
1) Generate 3D object proposals by clustering non-ground LiDAR points.

2) Estimate motion for each object proposal (static or dynamic).

3) Create appearance embedding for each object proposal using camera images.



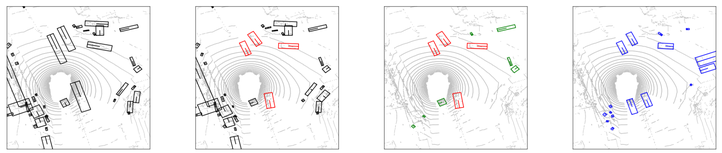
4) Cluster object proposals using their appearance embeddings.



5) Identify mobile clusters by selecting appearance clusters.
6) Train standard 3D object detector using pseudo-bounding boxes.

Qualitative results

Scene part of the nuScenes [1] training dataset.



Quantitative results

We increase AP by 178 % for unsupervised 3D object discovery on nuScenes [1].

Method	Labels	ST	AP ↑	NDS ↑	ATE ↓	ASE ↓	AOE ↓	AVE ↓
Supervised 1 %	Human	✗	27.8	26.3	0.456	0.309	1.302	1.307
Supervised 10 %	Human	✗	61.2	56.7	0.255	0.221	0.462	0.455
Supervised 100 %	Human	✗	76.5	68.7	0.209	0.198	0.241	0.305
HDBSCAN [2]	L	✗	13.8	15.9	**0.574**	0.522	1.601	1.531
OYSTER [3]	L	✓	9.1	11.5	0.784	0.521	1.514	-
LISO [4]	L	✓	10.9	13.9	0.750	**0.409**	1.062	-
UNION (ours)	L+C	✗	**38.4**	**31.2**	0.589	0.497	**0.874**	**0.836**

We test UNION with different image encoders.

Method	AP ↑	NDS ↑	ATE ↓	ASE ↓	AOE ↓	AVE ↓
DINOv2 ViT-L/14 w/ registers [5]	**38.4**	**31.2**	0.589	0.497	**0.874**	**0.836**
I-JEPA ViT-H/16 [6]	22.8	22.8	**0.561**	**0.486**	0.953	0.865

Conclusion

- We propose UNION for unsupervised 3D object detection.
- We are the first to use LiDAR, camera, and temporal information *jointly*.
- We set the new SOTA for unsupervised 3D object discovery.

References
[1] Caesar et al. (2020). nuScenes: A multimodal dataset for autonomous driving. In CVPR.
[2] McInnes et al. (2017). HDBSCAN: Hierarchical density based clustering. In JOSS.
[3] Zhang et al. (2023). Towards unsupervised object detection from LiDAR point clouds. In CVPR.
[4] Baur et al. (2024). LISO: Lidar-only self-supervised 3D object detection. In ECCV.
[5] Oquab et al. (2024). DINOv2: Learning robust visual features without supervision. In TMLR.
[6] Assran et al. (2023). Self-supervised learning from images with a joint-embedding predictive architecture. In CVPR.
[7] Wang et al. (2022). 4D unsupervised object discovery. In NeurIPS.

Acknowledgement:
This research has been conducted as part of the EVENTS project, which is funded by the European Union, under grant agreement No 101069614. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Paper GitHub