Università di Genova | DIPARTIMENTO INFORMATICA, BIOINGEGNERIA, ROBOTICA E INGEGNERIA DEI SISTEMI

**Computer Science Workshop**
*PhD program in Computer Science and Systems Engineering*

# Time-to-Label: Temporal Consistency for Self-Supervised Monocular 3D Object Detection

Issa Mouawad, Nikolas Brasch, Fabian Manhardt, Federico Tombari, Francesca Odone

## Motivations

- 3D perception is essential for several applications (Autonomous Driving, Robotics,...)
- The annotation process for 3D tasks is expensive and labor-intensive
- Self-supervised learning proved beneficial to reduce the amount of supervision for several other visual tasks

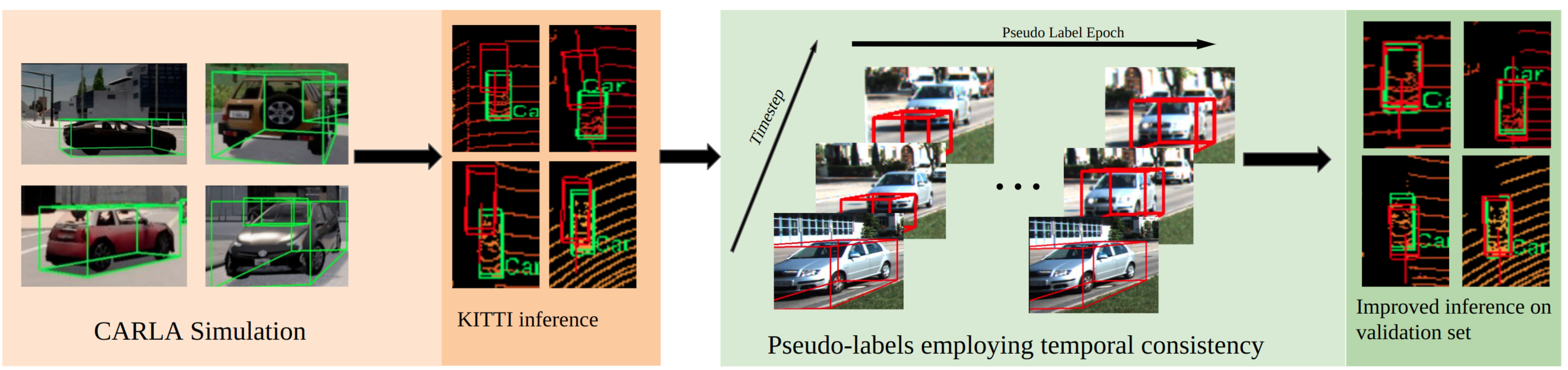
## Objective

Training a **monocular** 3D object detector without access to manually generated labels

## Contributions

1. A self-supervised framework to address 3D object detection without labels
2. A self-supervised loss that harnesses temporal and geometric prior in video sequences
3. Achieving state-of-the-art results on unsupervised 3D object detection.

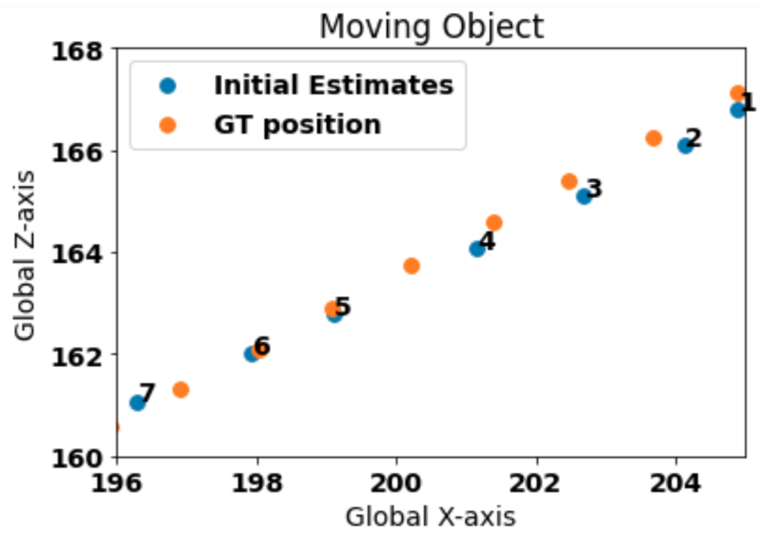
## Method Overview

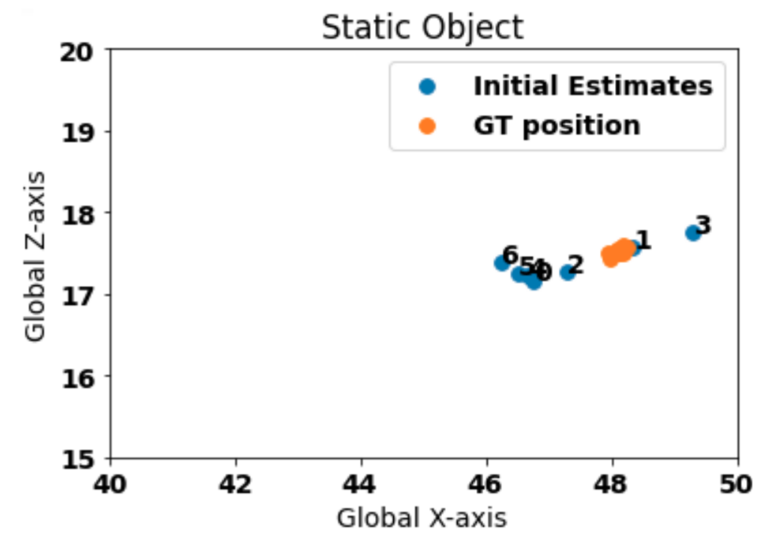



1. 3D monocular object detector is trained on synthetic data
2. The detector is used to generate initial estimates on the real-images dataset
3. The initial estimates are refined using geometry priors and our novel self-supervised loss
4. The resulting estimates are used as pseudo-labels to finetune the detector

### Temporal Prior

We use the trajectories we recover, in addition to the ego-motion from on-board sensors, to classify the motion state of objects to: **Static** and **Moving** objects.
Using the trajectory and the motion status, we derive, at each time step, the temporally consistent translation and rotation: $t_i^{temporal}$, $yaw_i^{temporal}$




### Self-supervised Loss

We use the temporal prior established on objects motion to further regularize the refined translation and rotation:

$$\mathcal{L}_{temporal} = \lambda_t \left\| t_i - t_i^{temporal} \right\|_2^2 + \lambda_r \left\| yaw_i - yaw_i^{temporal} \right\|_2^2 \tag{1}$$

Additionally, the **raw lidar** available during training is used to establish alignment between the predicted pose and the observed geometry using Chamfer distance:

$$\mathcal{L}_{CD} = \sum_{x \in \tilde{P}} \min_{y \in \hat{P}_{lidar}} ||x-y||_2^2 + \sum_{y \in \hat{P}_{lidar}} \min_{x \in \tilde{P}} ||x-y||_2^2, \tag{2}$$

## Experimental Analysis

### 1- Pseudo-labels Quality

We generate high-quality pseudo-labels compared to other similar methods [2] vspace0.2cm

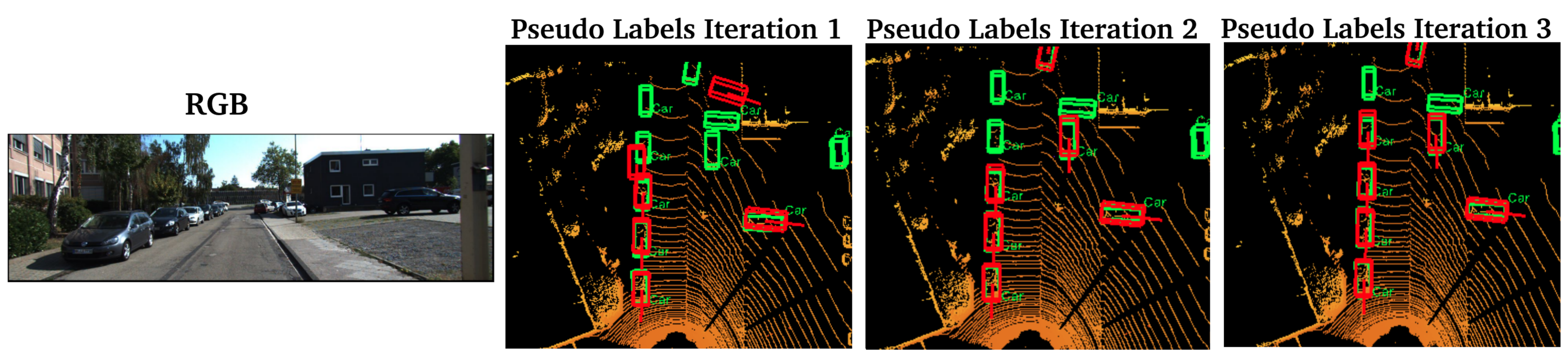
| Iteration | AP 2D % | | | AP BEV % | | |
|---|---|---|---|---|---|---|
| | Easy | Mod | Hard | Easy | Mod | Hard |
| 1 | 84.5 | 63.2 | 56.0 | 66.7 | 45.0 | 37.9 |
| 2 | 91.5 | 67.3 | 57.6 | 87.2 | 60.5 | 50.8 |
| 3 | 91.9 | 69.8 | 60.1 | **89.9** | **63.1** | **53.4** |
| Autolabeling [2] | Ground truth boxes | | | 77.8 | 59.7 | N/A |

### 2- Evaluation on KITTI Validation Set

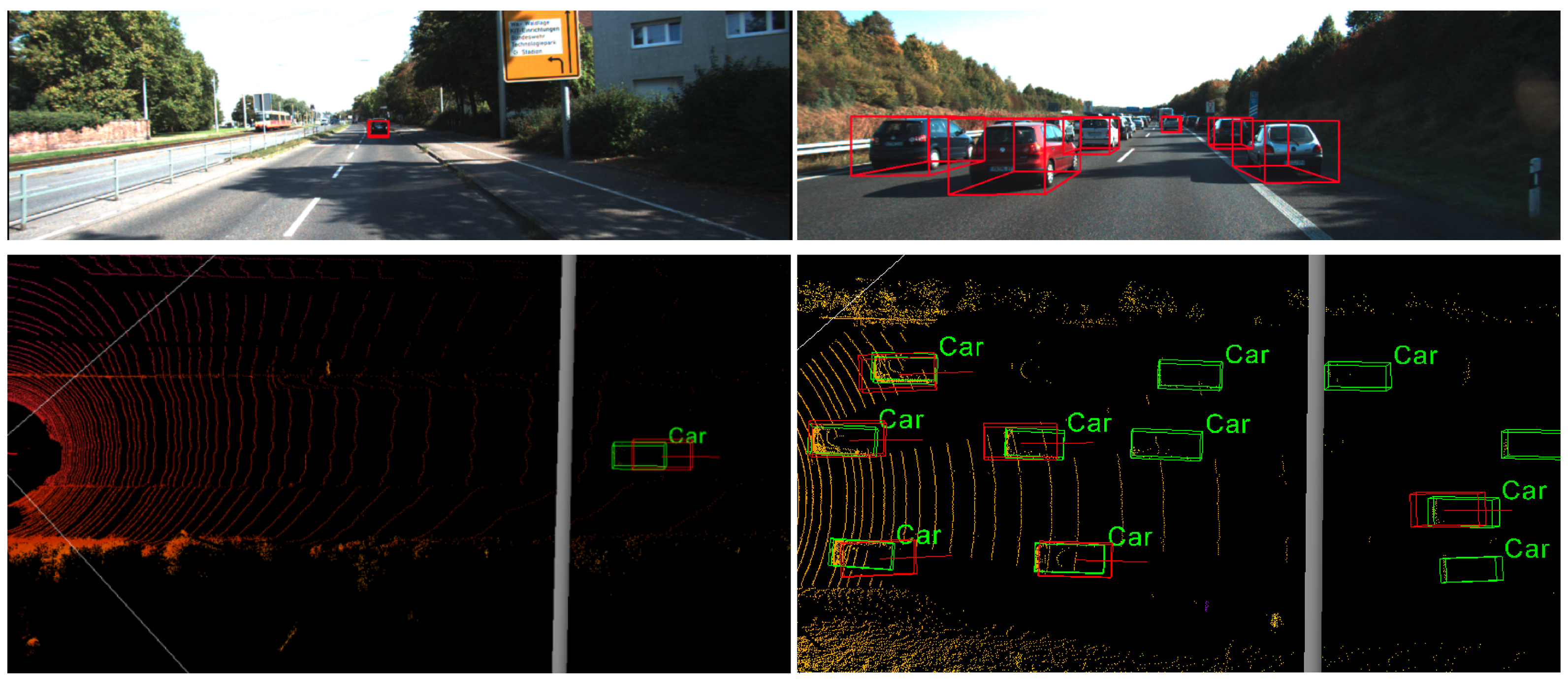
We finetune the detector with the generated pseudo-labels, and outperform other unsupervised methods on unseen validation set

| Method | Images | $AP_{BEV}$ / $AP_{3D}$ ($AP_{R11}$@ 0.5 IoU) Easy | Mod | Hard |
|---|---|---|---|---|
| Supervised | | | | |
| Deep3DBBox | trainsplit | 30.02/27.04 | 23.77/20.55 | 18.83/15.88 |
| Mono3D | trainsplit | 30.50/25.19 | 22.39/18.20 | 19.16/15.52 |
| M3D-RPN | trainsplit | 55.37/48.96 | 42.49/39.57 | 35.29/33.01 |
| LPCG-M3D-RPN [1] | trainsplit | 67.66/61.75 | **52.27**/49.51 | 46.65/**44.70** |
| MonoFlex [3] | trainsplit | **68.62**/**65.33** | 51.61/**49.54** | **49.73**/43.04 |
| Unsupervised | | | | |
| MonoDIS- SDFLabel [2] | trainsplit | 51.10/32.90 | 34.50/22.10 | - |
| **Ours** w/ MonoFlex | trainsplit | 52.43/36.71 | 37.55/26.74 | 31.21/22.09 |
| MonoDR | - | 51.13/45.76 | 37.29/32.31 | 30.20/26.19 |
| LPCG-M3D-RPN[1] | Raw data | 52.06/47.58 | 35.37/29.06 | 28.61/26.58 |
| **Ours** w/ MonoFlex | Raw data | 63.94/51.90 | 42.29/33.24 | 35.31/30.39 |

Pseudo-Labels




Inference on validation set



## Contacts

**Issa Mouawad**
issa.mouawad@dibris.unige.it

## Main References

[1] Liang Peng, Fei Liu, Zhengxu Yu, Senbo Yan, Dan Deng, and Deng Cai. Lidar point cloud guided monocular 3d object detection. *arXiv preprint arXiv:2104.09035*, 2021.

[2] Sergey Zakharov, Wadim Kehl, Arjun Bhargava, and Adrien Gaidon. Autolabeling 3d objects with differentiable rendering of sdf shape priors. In *CVPR*, 2020.

[3] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *CVPR*, 2021.