
ECCV'20
ONLINE
23-28 AUGUST 2020
16TH EUROPEAN CONFERENCE ON
COMPUTER VISION
WWW.ECCV2020.EU

Carnegie Mellon University
The Robotics Institute

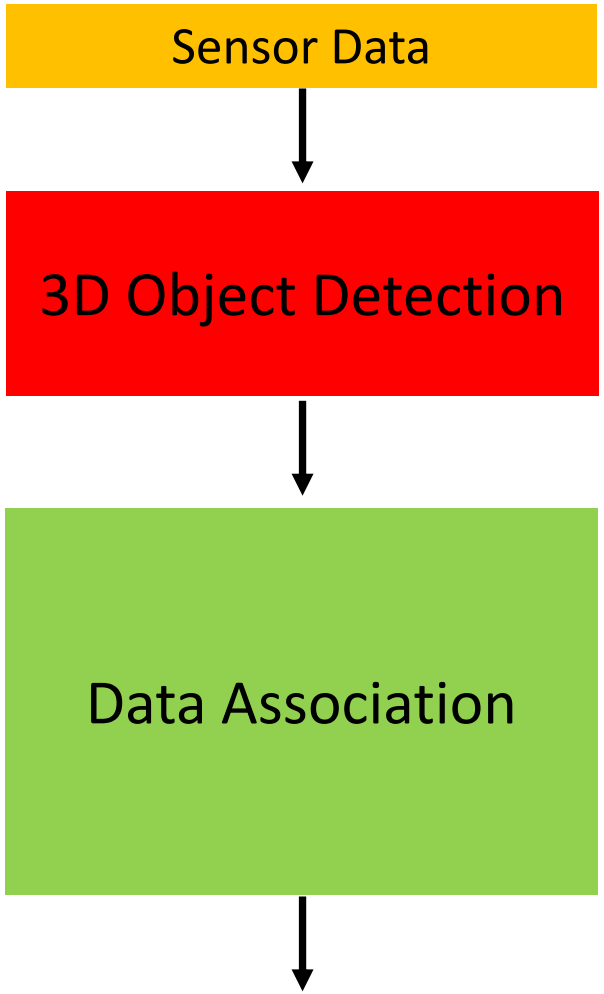
Graph Neural Network for 3D Multi-Object Tracking

Xinshuo Weng, Yongxin Wang, Yunze Man, Kris Kitani

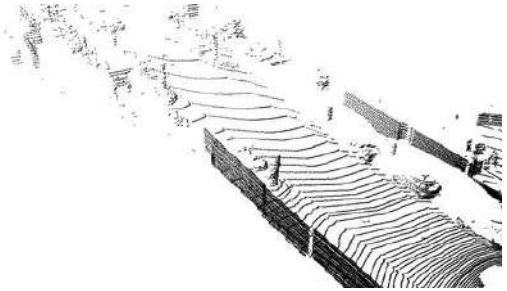
Robotics Institute, Carnegie Mellon University

European Conference on Computer Vision (ECCV) Workshops

Standard 3D MOT Pipeline



Standard 3D MOT Pipeline



LiDAR point clouds

Sensor Data

3D Object Detection

Data Association

Object trajectories

RGB frames

Standard 3D MOT Pipeline

Sensor Data

↓

3D Object Detection

↓

Data Association

↓

Object trajectories

Detection results

Standard 3D MOT Pipeline

Sensor Data

3D Object Detection

Feature Extraction

New Detections

Past Tracklets

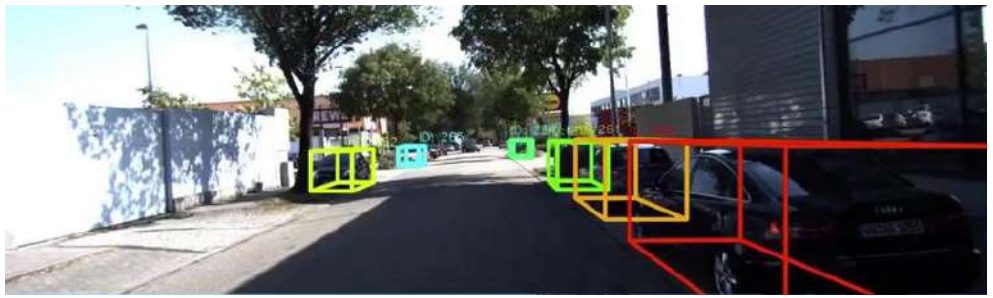
$$\begin{bmatrix} 0.004 & 0.987 & 0. & \cdots & 0.002 \\ 0.992 & 0.001 & 0. & \cdots & 0. \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0. & 0. & 1. & \cdots & 0. \end{bmatrix}$$

Affinity matrix

Bipartite Matching

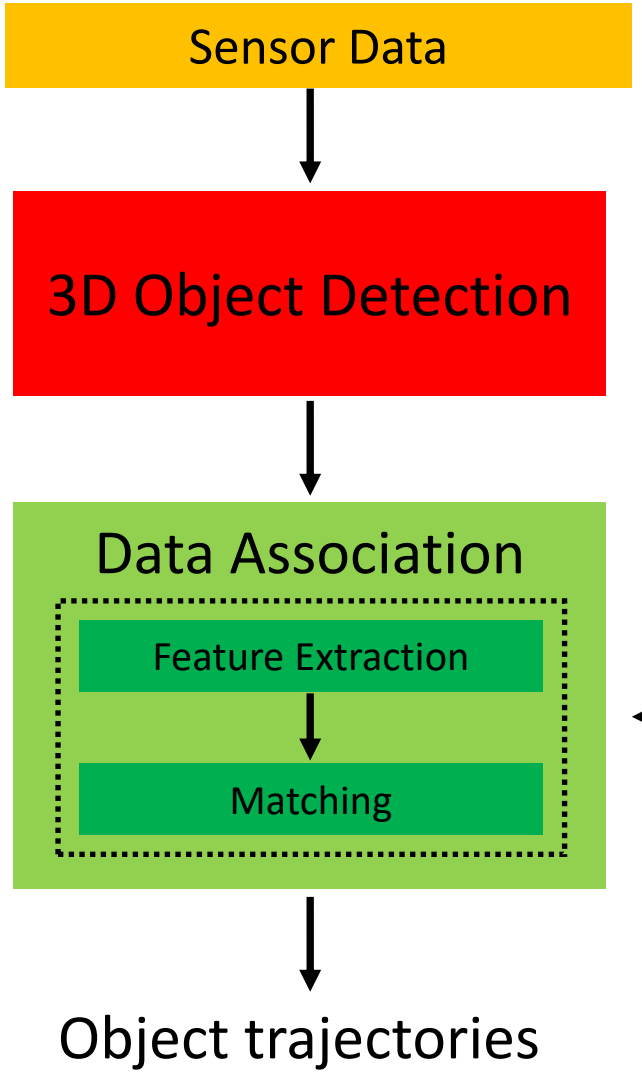
Data Association

Object trajectories



3D MOT results

Limitation of the Prior Work



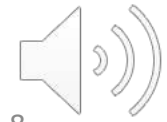
Limitation

1. Feature representation does not take into account contexts of other objects

2. Feature representation does not fully utilize information from multiple modalities that is complementary

Our Contributions

1. A novel feature interaction mechanism to encode contexts via object interaction
2. A 2D-3D joint feature extractor to learn multi-modal features that are complementary



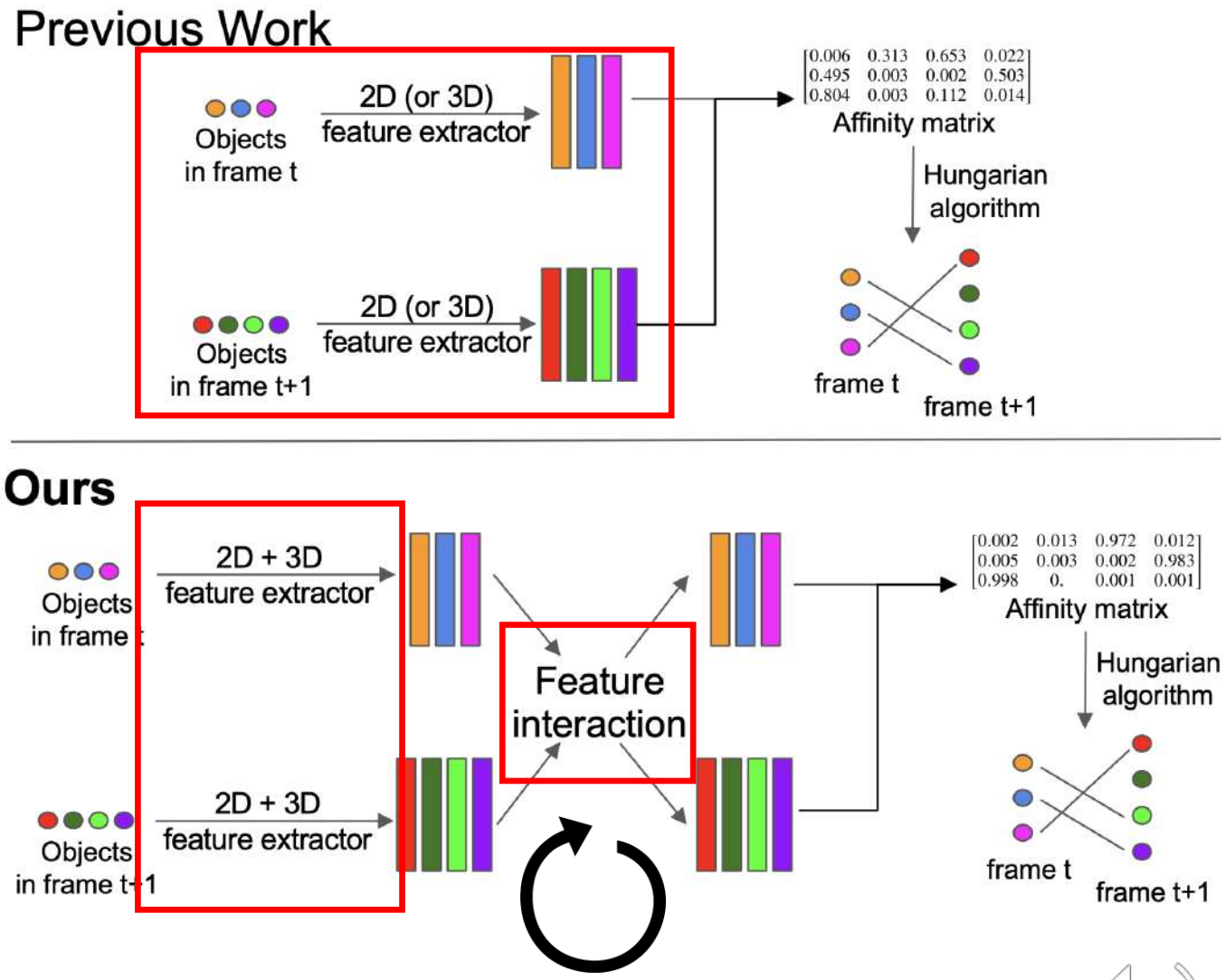
Our Contributions

Prior work

- Feature extraction is independent of each object
- Employs features from one modality (2D or 3D)

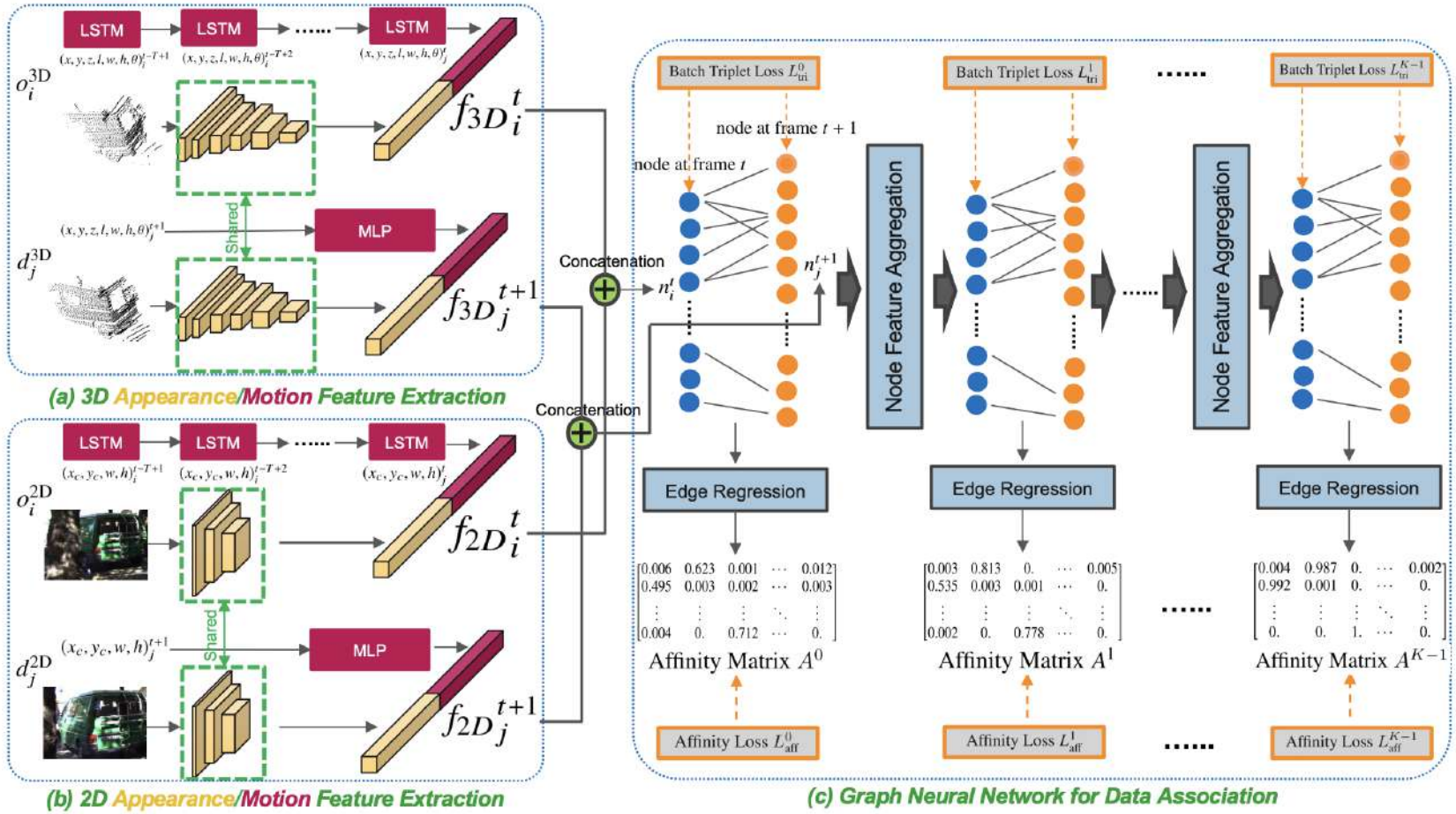
Our Approach

- A joint feature extractor to learn multi-modal features
- A novel feature interaction mechanism to iteratively encode context and improve discriminative feature learning



Our Approach

- (a) Obtain the appearance / motion features from the 3D space
- (b) Obtain the appearance / motion features from the 2D space
- (c) Learn discriminative object features by encoding context through object feature interaction



Ablation Study

Improve Feature Learning for 3D MOT

- Is encoding the multi-modal features really useful?

Feature Extractor	**sAMOTA** (%) ↑	AMOTA (%) ↑	AMOTP (%) ↑	MOTA (%) ↑
2D A	88.31	41.62	76.22	79.42
2D M	64.24	23.95	61.13	54.88
3D A	88.27	41.55	76.29	77.38
3D M	88.57	41.62	76.22	81.84

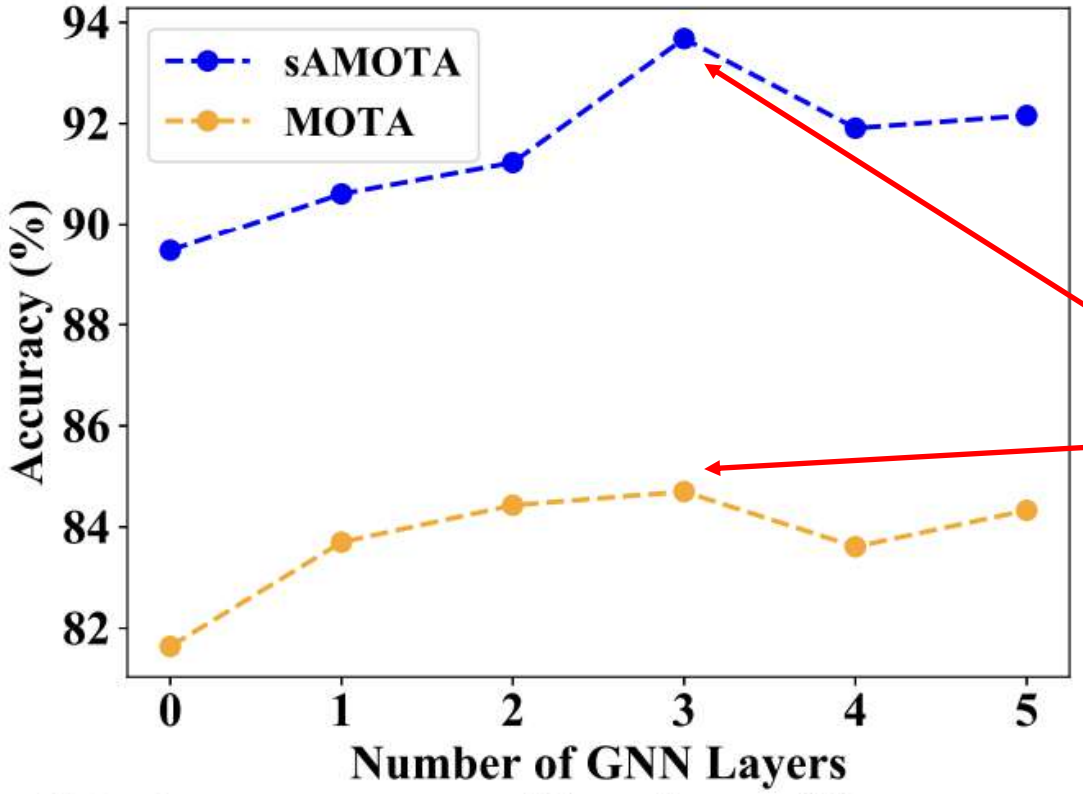
A: appearance feature, M: motion feature

Use feature from single modality

Use feature from multiple modalities: Performance increased!

Improve Feature Learning for 3D MOT

- Is feature interaction using GNNs useful to 3D MOT?

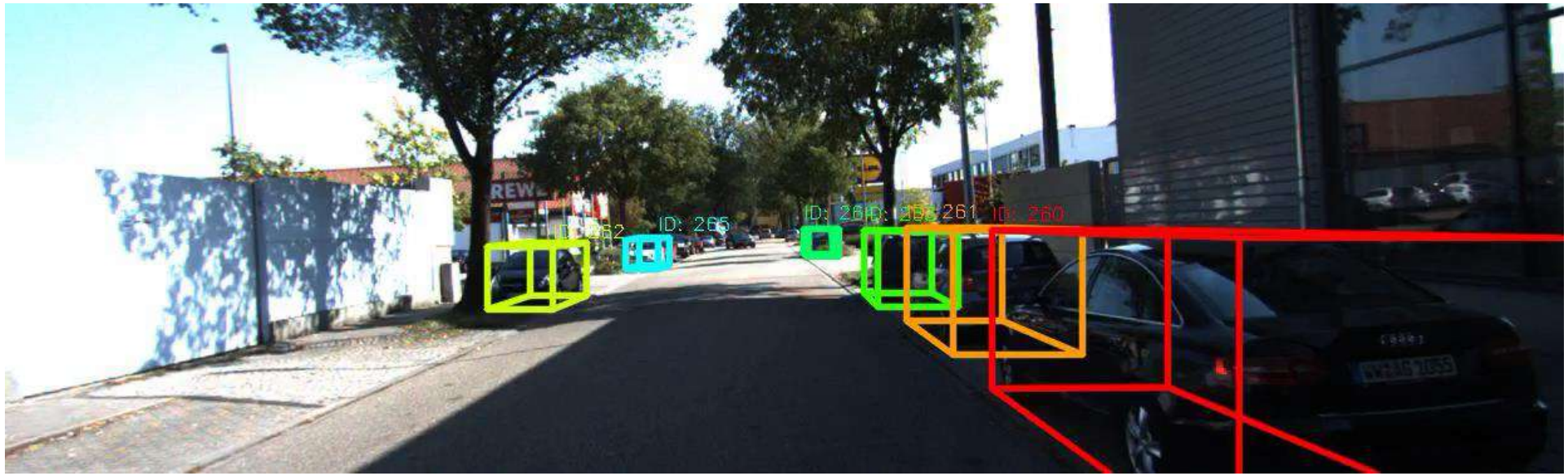


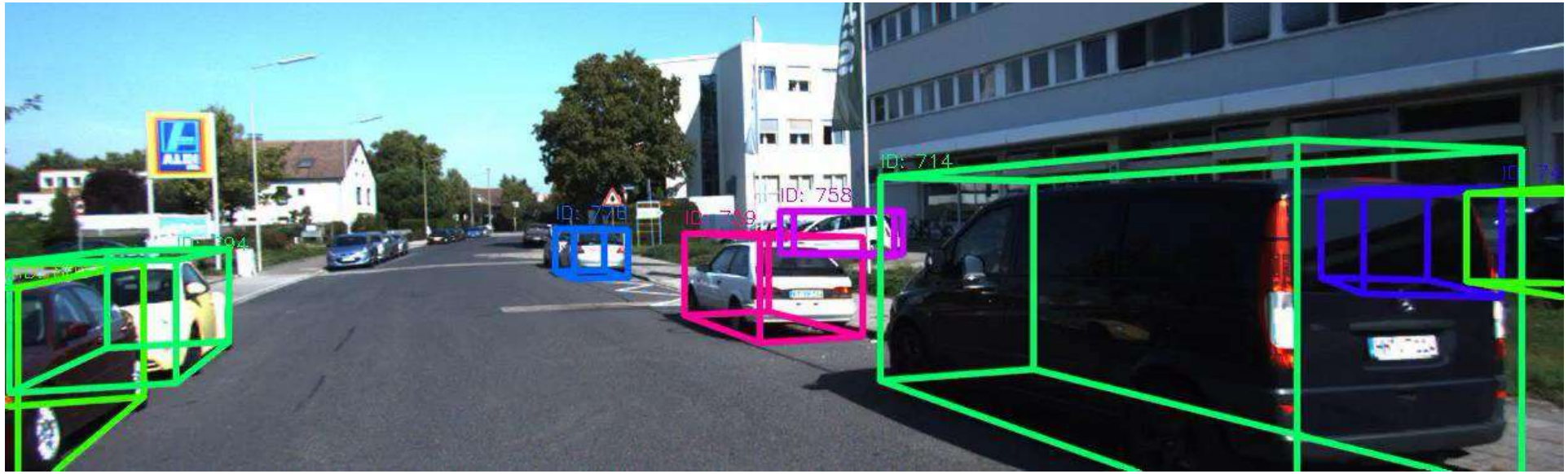
(b) Accuracy v.s. Number of Layers

Performance largely increased with GNN layers = 3 v.s. 0 !

Qualitative Results

Qualitative Results





Carnegie Mellon University
The Robotics Institute

Graph Neural Network for 3D Multi-Object Tracking

Xinshuo Weng, Yongxin Wang, Yunze Man, Kris Kitani

Robotics Institute, Carnegie Mellon University

European Conference on Computer Vision (ECCV) Workshops