GNN3DMOT: Graph Neural Network for 3D Multi-Object Tracking with 2D-3D Multi-Feature Learning

Xinshuo Weng, Yongxin Wang, Yunze Man, Kris Kitani

Robotics Institute, Carnegie Mellon University

Motivation

- 3D multi-object tracking (MOT) is crucial to the perception of autonomous systems

Assistive robot

Sports

Autonomous driving

THE ROBOTICS INSTITUTE

Goal: Tracking-by-Detection

- Associate the detections across frames
- Leverage information from the sensor data
- Learn discriminative features to differentiate objects with different identities

3D object detection

3D multi-object tracking

… …

Sensor data

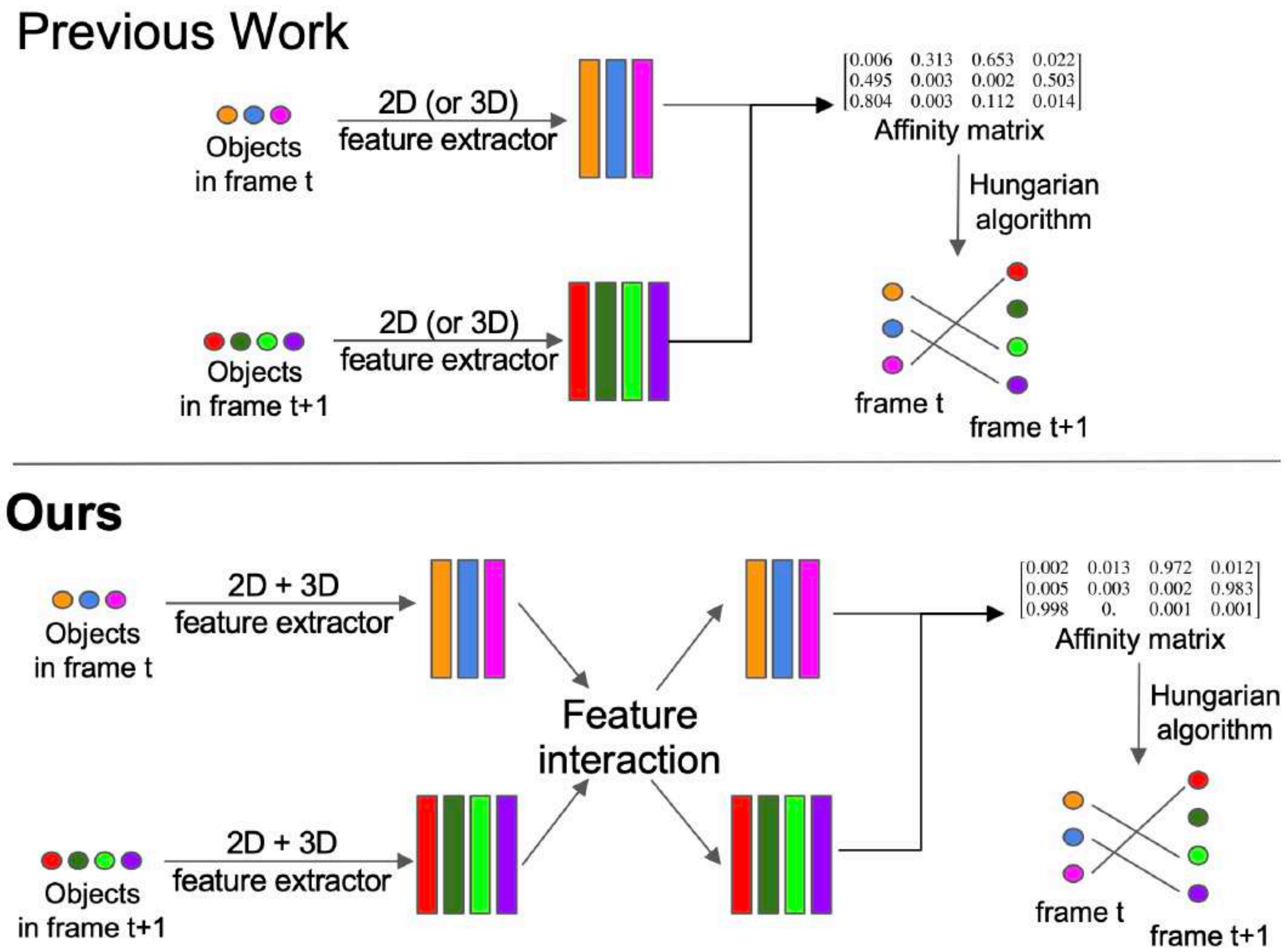
Limitation of the Prior Work

Prior work

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

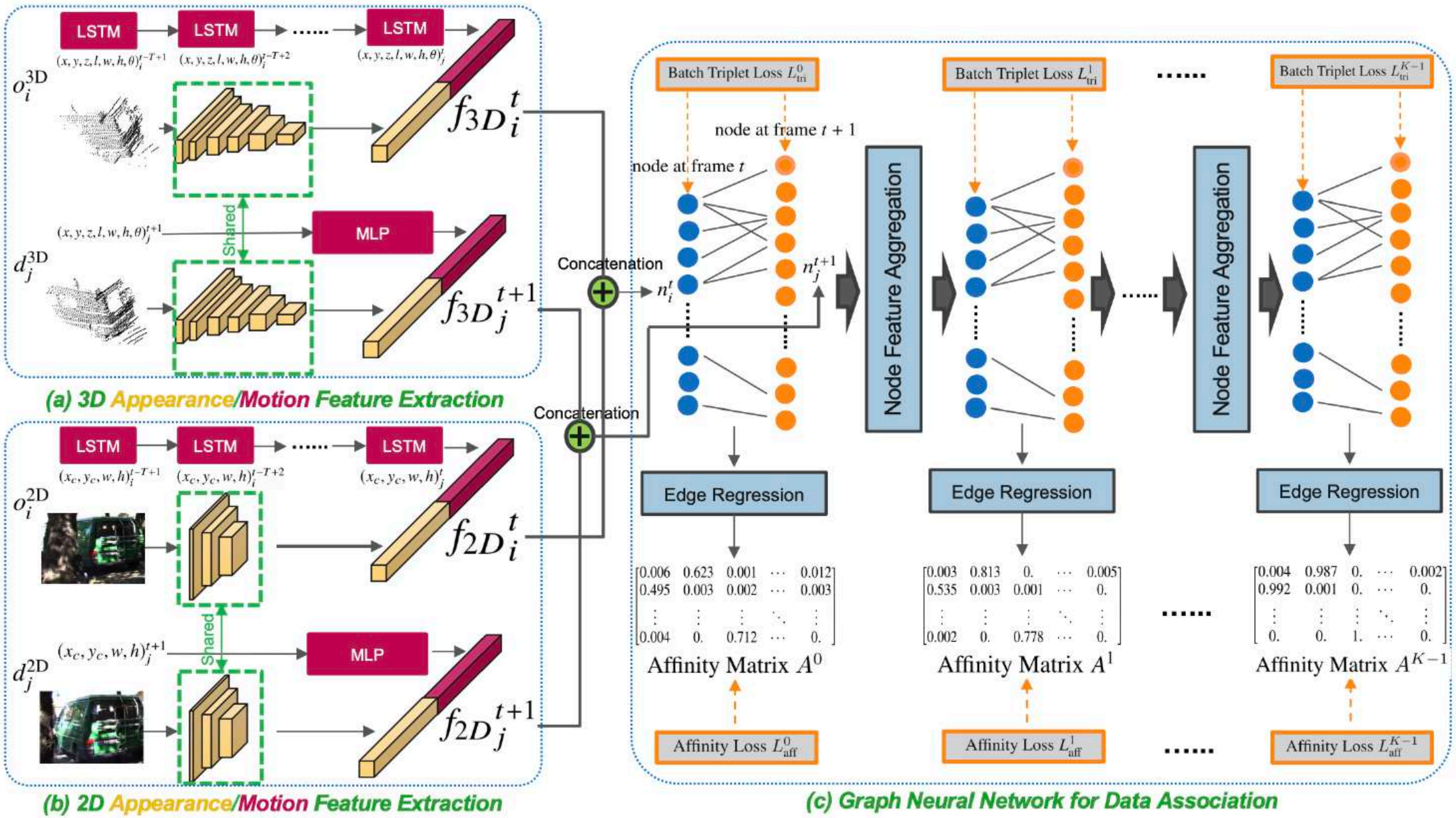
Our Approach

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



Our Approach

- (a, b) Obtain the appearance / motion features from both 2D images and 3D point cloud
- (c) Learn discriminative object features through interaction



Quantitative Results

- State-of-the-art performance in 3D MOT
- Competitive performance in 2D MOT by projecting 3D MOT results to 2D space

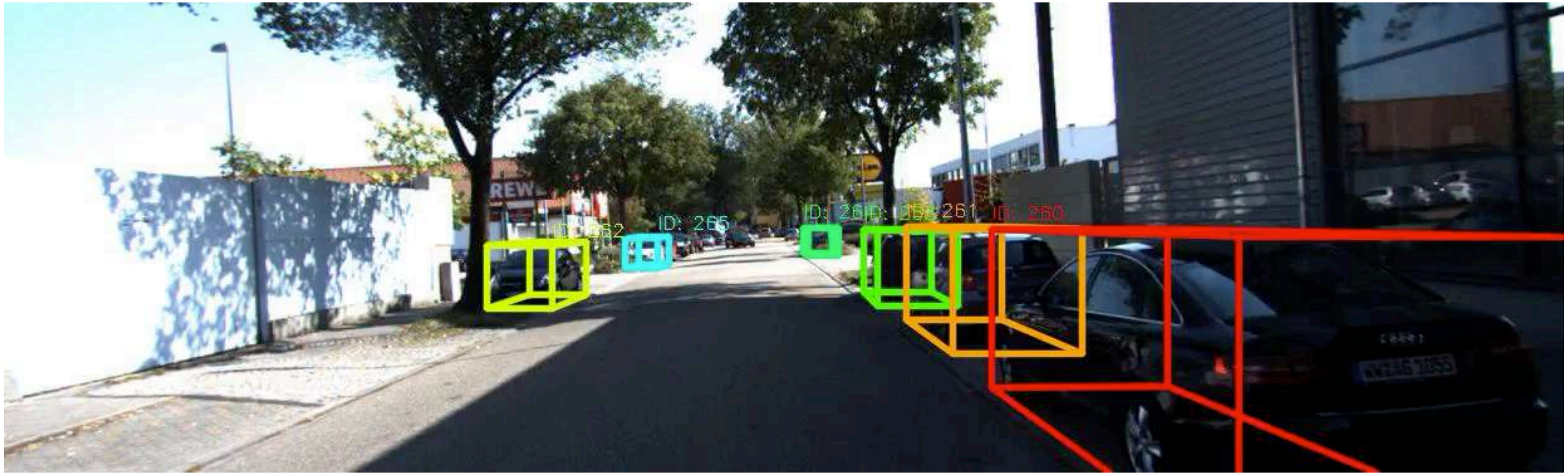
Table 1. Quantitative comparison on KITTI-Car val set. The evaluation is conducted in **3D** space using [45] 3D MOT evaluation tool.

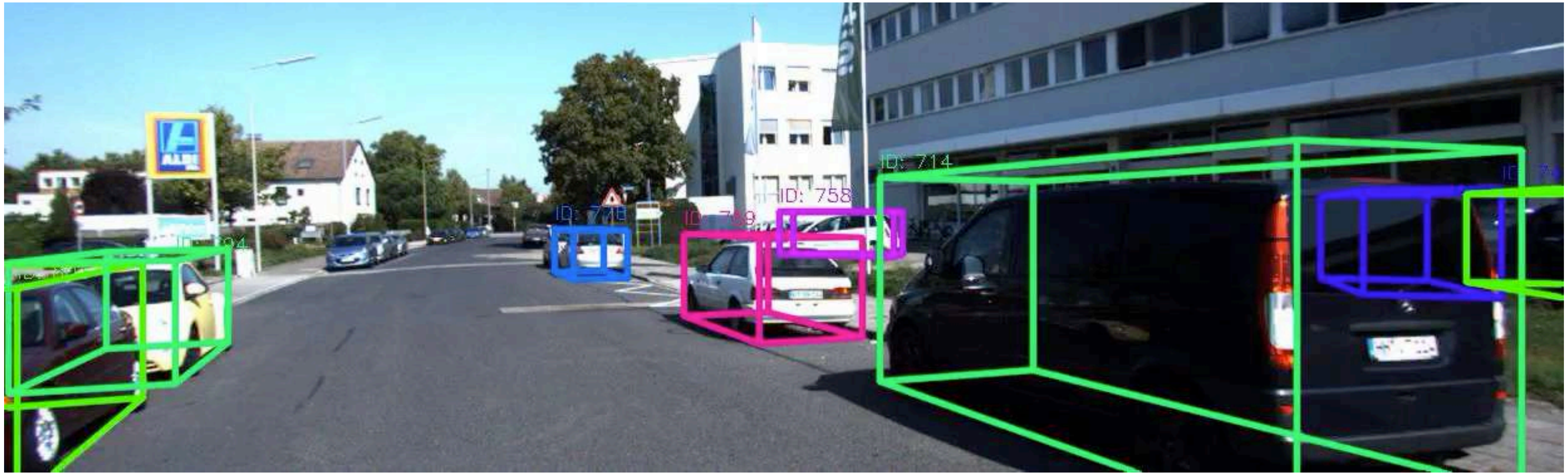
Method	Input Data	**sAMOTA** (%) ↑	AMOTA (%) ↑	AMOTP (%) ↑	MOTA (%) ↑	MOTP (%) ↑	IDS ↓	FRAG ↓
mmMOT [56] (ICCV'19)	2D + 3D	70.61	33.08	72.45	74.07	78.16	10	125
FANTrack [2] (IV'19)	2D + 3D	82.97	40.03	75.01	74.30	75.24	35	202
AB3DMOT[45] (arXiv'19)	3D	91.78	44.26	77.41	83.35	78.43	**0**	15
Ours	2D + 3D	**93.68**	**45.27**	**78.10**	**84.70**	**79.03**	**0**	**10**

Table 2. Quantitative comparison on KITTI-Car test set. The evaluation is conducted in **2D** space using KITTI 2D MOT evaluation tool.

Method	Input Data	**MOTA** (%) ↑	MOTP (%) ↑	MT (%) ↑	ML (%) ↓	IDS ↓	FRAG ↓	FPS ↑
CIWT [24] (ICRA'17)	2D	75.39	79.25	49.85	10.31	165	660	2.8
FANTrack [2] (IV'19)	2D + 3D	77.72	82.32	62.61	8.76	150	812	25.0 (GPU)
AB3DMOT[45] (arXiv'19)	3D	83.84	85.24	66.92	11.38	**9**	**224**	**214.7**
BeyondPixels [30] (ICRA'18)	2D	84.24	85.73	73.23	2.77	468	944	3.33
3DT [16] (ICCV'19)	2D	84.52	**85.64**	73.38	**2.77**	377	847	33.3 (GPU)
mmMOT [56] (ICCV'19)	2D + 3D	84.77	85.21	73.23	**2.77**	284	753	4.8 (GPU)
MASS [17] (IEEE Access'19)	2D	**85.04**	85.53	**74.31**	**2.77**	301	744	100.0
Ours	2D + 3D	80.40	85.05	70.77	11.08	113	265	5.2 (GPU)
Ours + 2D detections from [27]	2D	82.24	84.05	64.92	6.00	142	416	5.1 (GPU)

Qualitative Results





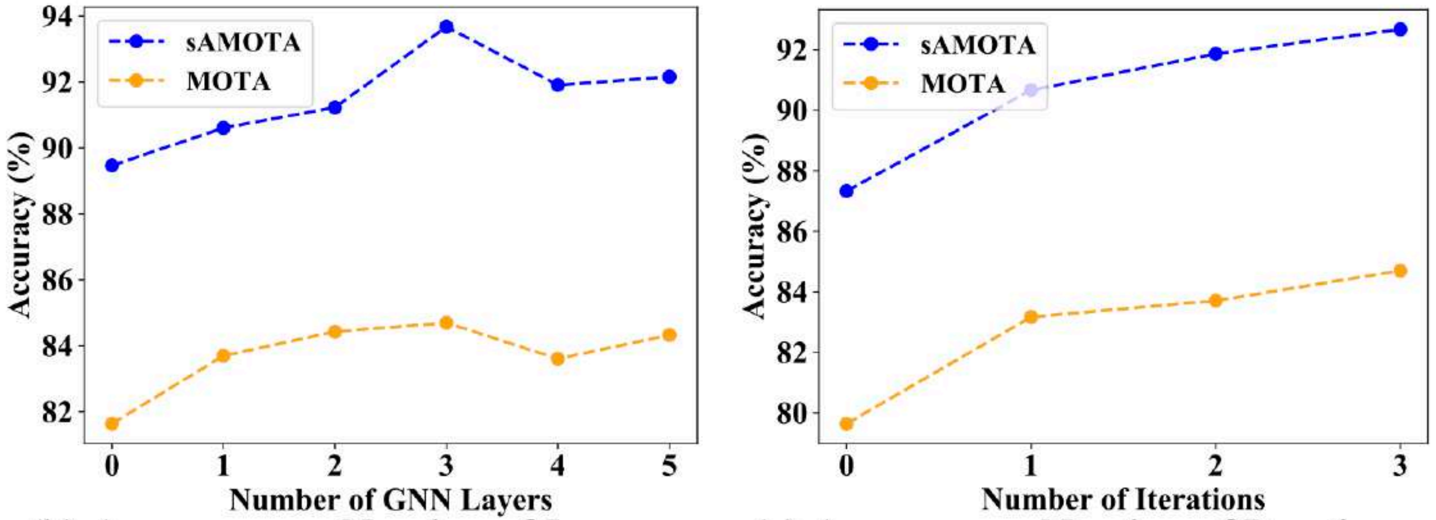
THE ROBOTICS INSTITUTE

Ablation Study on the Graph Neural Network

- Ablation on different graph networks

Node Aggregation	sAMOTA (%) ↑	AMOTA (%) ↑	AMOTP (%) ↑	MOTA (%) ↑
Type 1	75.61	32.84	65.81	67.43
Type 2 (SAGEConv [13])	87.81	41.06	76.29	77.22
Type 2 (GCN [18])	89.78	43.37	78.06	80.67
Type 2 (GraphConv [23])	91.15	44.78	77.93	82.31
Type 2 (GATConv [38])	91.66	44.57	77.99	82.37
Type 2 (AGNNConv [36])	91.88	44.95	78.00	84.32
Type 3 (EdgeConv [43])	92.17	44.65	77.98	83.73
Type 4 (**Ours**)	**93.68**	**45.27**	**78.10**	**84.70**

- Ablation on different number of graph layers



(b) Accuracy v.s. Number of Layers

(c) Accuracy vs Number of Iterations

Ablation Study on the 2D-3D Multi-Feature Learning

- Combining features from different modalities improves 3D MOT performance

Table 4. **Effect of Joint Feature Extractor**. Results are evaluated on KITTI-Car val set using **3D** MOT evaluation tool. Appearance and motion features are denoted as “A” and “M” respectively.

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
2D+3D A	89.39	42.55	76.24	83.02
2D+3D M	91.75	44.75	78.05	84.54
2D M+A	90.56	44.39	**78.20**	83.15
3D M+A	91.30	44.31	78.16	84.06
2D+3D M+A (**Ours**)	**93.68**	**45.27**	78.10	**84.70**

Thank You!