CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

Reconstruction-based 3D Perception

Jiawei He

Jul 16th, 2023

❖ Outline

Background & Previous Work

BA-Det

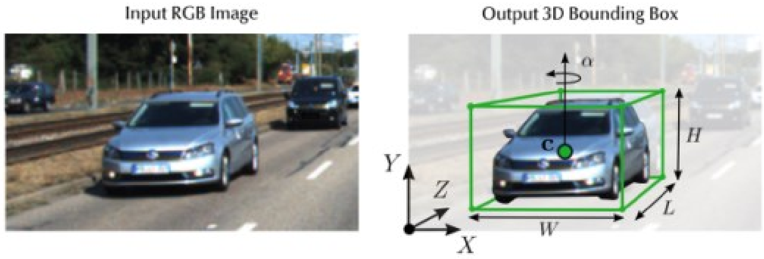
BA^2-Det

BA^2-Track

❖ Background

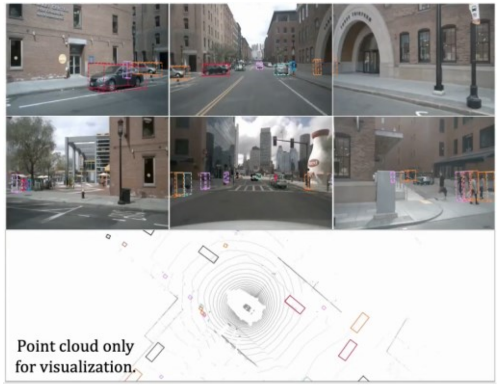
❑ 3D Perception from Images

3D object detection

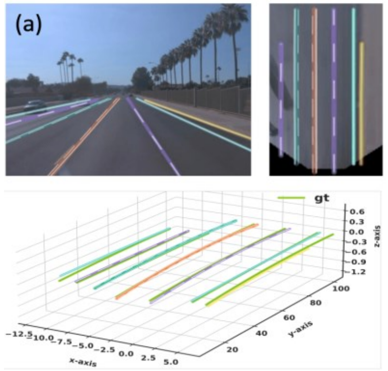


BEV semantic segmentation

3D multi-object tracking



3D lane detection

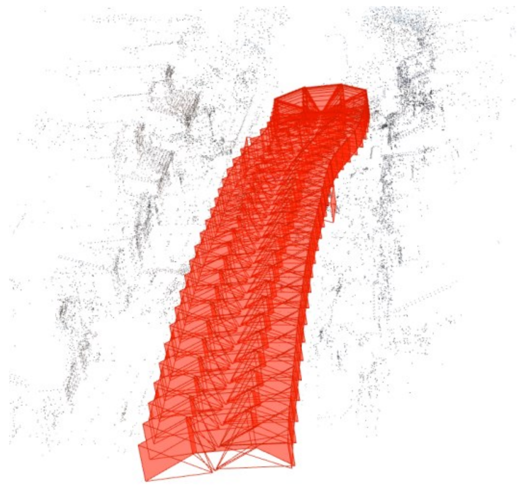


3D occupancy prediction

❖ Background

❑ Bundle Adjustment (BA)

- Widely used in scene reconstruction and SLAM
- Jointly optimizing camera pose and point cloud
- Non-linear least-squares problem



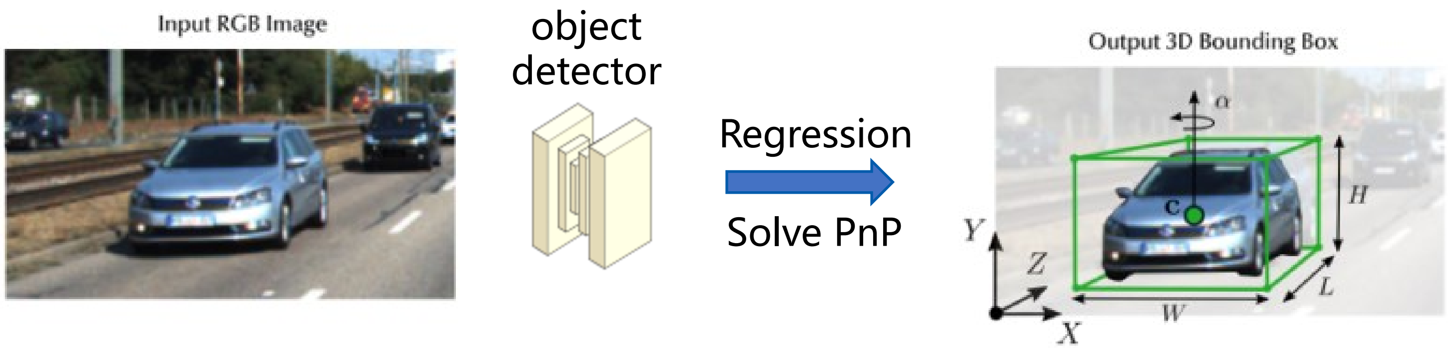
$$\underbrace{\{\bar{\mathbf{T}}^t_{gc}\}_{t=1}^{T}}_{\text{Optimizing Camera Pose}}, \underbrace{\{\bar{\mathbf{P}}_i\}_{i=1}^{m}}_{\text{Optimizing 3D Point Loc}} = \underset{\{\mathbf{T}^t_{gc}\}_{t=1}^{T}, \{\mathbf{P}_i\}_{i=1}^{m}}{\arg\min} \frac{1}{2}\sum_{i=1}^{m}\sum_{t=1}^{T} ||\mathbf{p}_i^t - \Pi(\mathbf{T}^t_{gc}, \mathbf{P}_i, \mathbf{K})||^2$$

2D images $\xrightarrow[\text{dimension lifting}]{\text{Bundle Adjustment}}$ 3D structure

- Providing constraints: **BA-Det**
- Providing supervision: **BA²-Det**, **BA²-Track**

❖ Previous Work

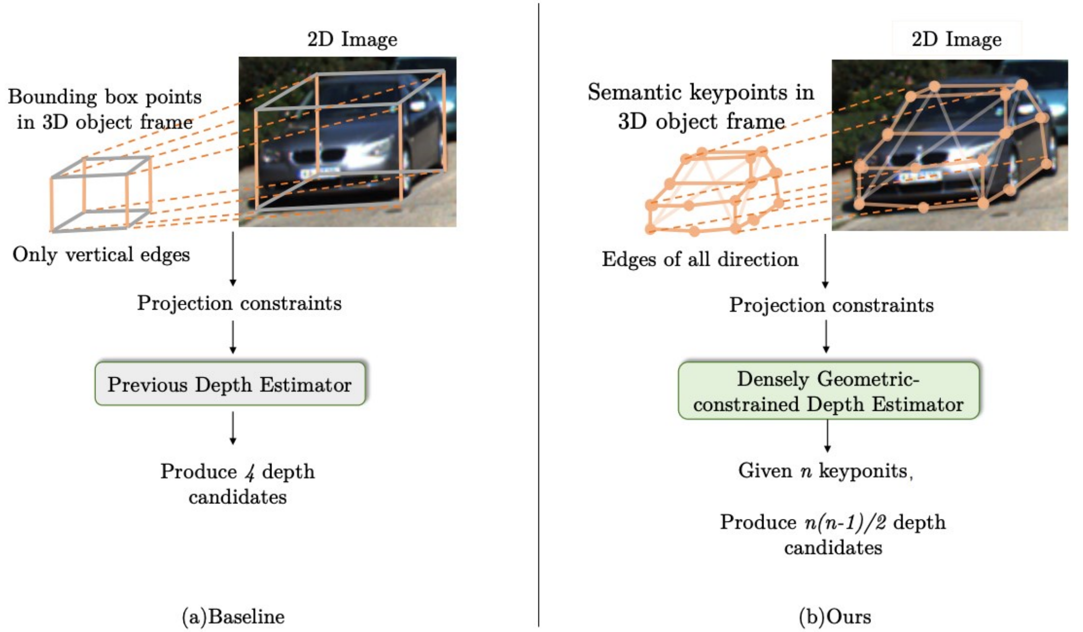
❑ Monocular 3D Object Detection



- **Regressing directly**
  - CenterNet (arXiv 19)
  - FCOS3D (CVPRW 21)
  - ……
- **Based on depth map**
  - PseudoLiDAR (CVPR 19)
  - PL++ (ICLR 20)
  - D4LCN (CVPR 20)
  - PatchNet (ECCV 20)
  - CaDDN (CVPR 21)
  - ……
- **Geometric constraints**
  - DeepMANTA (CVPR 17)
  - MonoFlex (CVPR 21)
  - AutoShape (CVPR 21)
  - Epro-PnP (CVPR 22)
  - DCD (ECCV 22)
  - ……

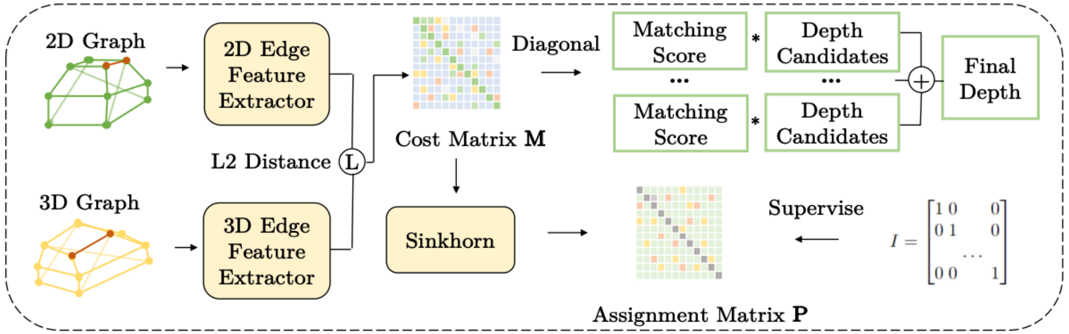
❖ Previous Work

❑ Densely Geometric-Constrained Depth Estimator (DCD)



Graph Matching Weighting module

Weighting depth candidates to obtain final depth



(b) GMW module

Dense geometric constraints → More depth candidates

Yingyan Li, Yuntao Chen, **Jiawei He**, Zhaoxiang Zhang. *Densely Constrained Depth Estimator for Monocular 3D Object Detection.* In European Conference on Computer Vision (ECCV) 2022.

❖ BA-Det: Object-centric Temporal Optimization

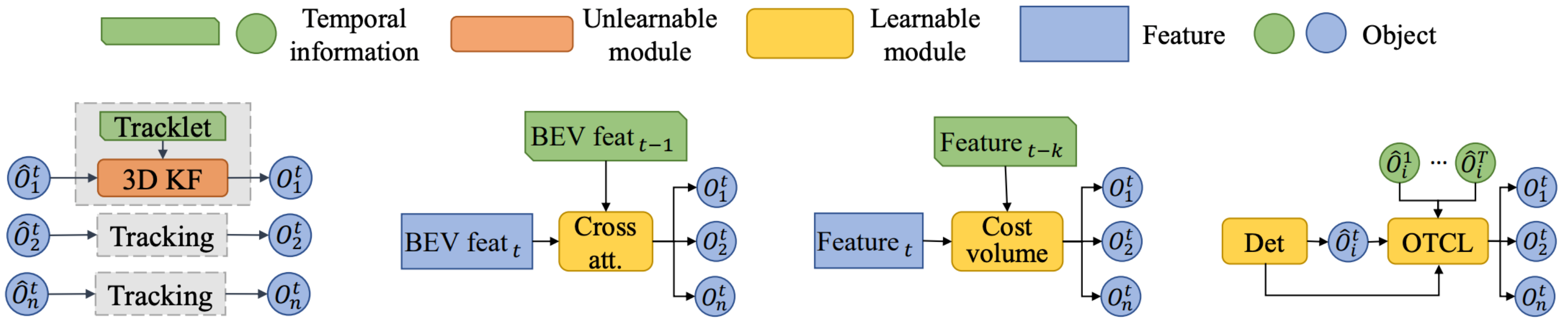
3D Video Object Detection with Learnable Object-Centric Global Optimization (CVPR 2023)

❑ Existing Video-based Methods

a) Temporal filtering: Without learning

b) Temporal BEV: Short-term, feature drifting (dynamic objects)

c) Stereo from video: Short-term (Two-frame), ignoring dynamic objects



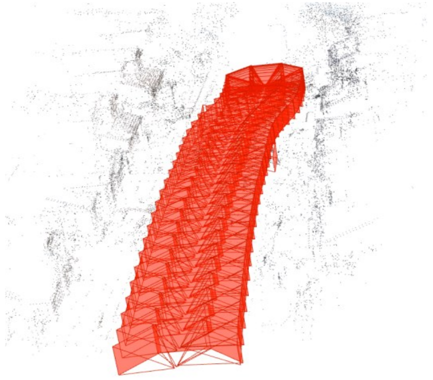
(a) Temporal Filtering (b) Temporal BEV (c) Stereo from Video (d) BA-Det (Ours)

❖ BA-Det: Object-centric Temporal Optimization

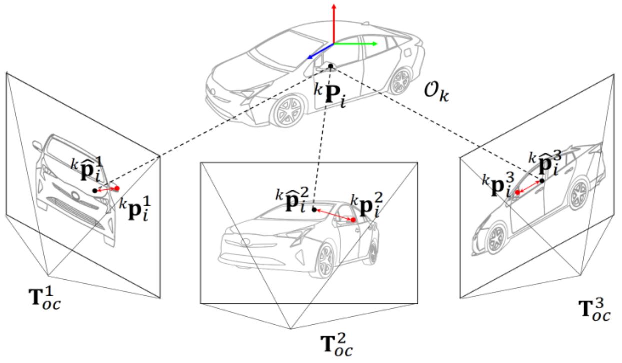
3D Video Object Detection with Learnable Object-Centric Global Optimization (CVPR 2023)

❑ Object-centric Geometric Constraints in Video

- From Scene-level BA to Object-centric BA
  - Optimizing camera pose → optimizing object pose
  - Correspondence: hand-craft sparse feature → learnable dense feature



Scene-level BA



Object-centric BA

❖ BA-Det: Object-centric Temporal Optimization

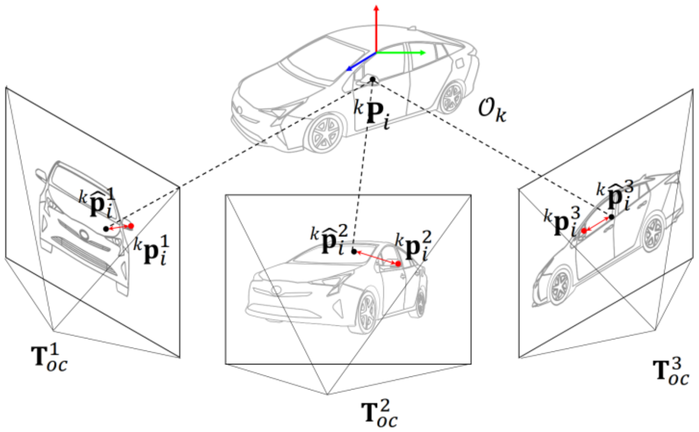
3D Video Object Detection with Learnable Object-Centric Global Optimization (CVPR 2023)

❑ Object-centric Geometric Constraints in Video

- Object-centric BA

$$\bar{\mathcal{T}}_k, \bar{\mathcal{P}}_k = \arg\min_{\mathcal{T}_k, \mathcal{P}_k} \frac{1}{2} \sum_{i=1}^{m} \sum_{t=1}^{T} ||{}^k\mathbf{p}_i^t - \Pi({}^k\mathbf{T}_{co}^t, {}^k\mathbf{P}_i, \mathbf{K})||_2^2$$

- Advantages compared with existing work
  - Object-centric manner
    - ✓ Handling both static and moving objects
  - Dense temporal correspondence learning
    - ✓ Utilizing longer temporal information

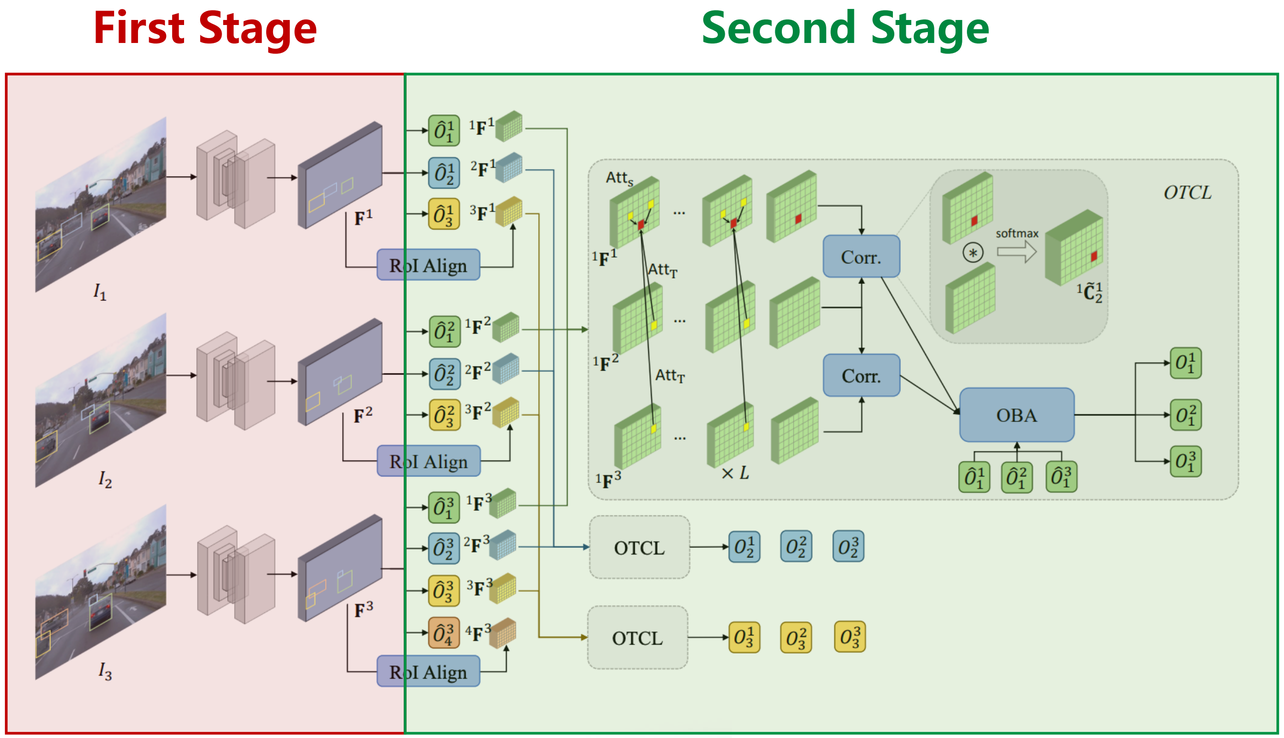


❖ BA-Det: Object-centric Temporal Optimization

3D Video Object Detection with Learnable Object-Centric Global Optimization (CVPR 2023)

❑ BA-Det: Object-centric Global Optimizable Detector

- **First Stage**
  - ✓ Single-frame object detection
  - ✓ Temporal object association
- **Second Stage**
  - ✓ Temporal/spatial aggregation
  - ✓ Correspondence learning with featuremetric OBA loss



❖ BA-Det: Object-centric Temporal Optimization

3D Video Object Detection with Learnable Object-Centric Global Optimization (CVPR 2023)

❑ Object-centric Temporal Correspondence Learning Module

- Featuremetric Object Bundle Adjustment Loss
  - Featuremetric OBA

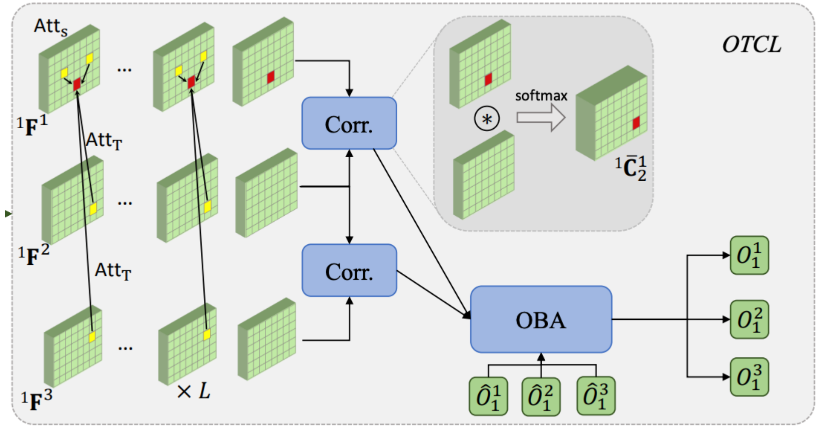
$$\bar{\mathcal{T}}_k, \bar{\mathcal{P}}_k = \arg\min_{\mathcal{T}_k, \mathcal{P}_k} \frac{1}{2} \sum_{i=1}^{m} \sum_{t=1}^{T} ||\mathbf{f}[{}^k\mathbf{p}_i^t] - \mathbf{f}[\Pi({}^k\mathbf{T}_{oc}^t, {}^k\mathbf{P}_i, \mathbf{K})]||_2^2$$

  - featuremetric reprojection loss

$$\mathcal{L}_{\text{rep}}^k = \sum_{i=1}^{m} \sum_{t=1}^{T} ||{}^k e_i^t||_2^2 = \sum_{i=1}^{m} \sum_{t=1}^{T} \sum_{t'=1}^{T} ||{}^k\mathbf{f}_i^t - {}^k\mathbf{f}_i^{t'}||_2^2$$

  - L2 norm to cosine distance

$$\mathcal{L}_{\text{OBA}}^k = -\sum_{i=1}^{m} \sum_{t=1}^{T} \sum_{t'=1}^{T} \log({}^k\widetilde{\mathbf{C}}_t^{t'}[{}^k\bar{\mathbf{p}}_i^t, {}^k\bar{\mathbf{p}}_i^{t'}]).$$



Supervised on correlation between temporal RoI features

❖ BA-Det: Object-centric Temporal Optimization

3D Video Object Detection with Learnable Object-Centric Global Optimization (CVPR 2023)

❑ Results on Waymo Open Dataset (WOD)

	LEVEL_1				LEVEL_2			
	3D AP_{70}	3D APH_{70}	3D AP_{50}	3D APH_{50}	3D AP_{70}	3D APH_{70}	3D AP_{50}	3D APH_{50}
M3D-RPN [2]	0.35	0.34	3.79	3.63	0.33	0.33	3.61	3.46
PatchNet [28]	0.39	0.37	2.92	2.74	0.38	0.36	2.42	2.28
PCT [42]	0.89	0.88	4.20	4.15	0.66	0.66	4.03	3.99
MonoJSG [23]	0.97	0.95	5.65	5.47	0.91	0.89	5.34	5.17
GUPNet [27]	2.28	2.27	10.02	9.94	2.14	2.12	9.39	9.31
DEVIANT [18]	2.69	2.67	10.98	10.89	2.52	2.50	10.29	10.20
CaDDN [33]	5.03	4.99	17.54	17.31	4.49	4.45	16.51	16.28
DID-M3D [31]	-	-	20.66	20.47	-	-	19.37	19.19
BEVFormer [22]†	-	7.70	-	30.80	-	6.90	-	27.70
DCD [21]	12.57	12.50	33.44	33.24	11.78	11.72	31.43	31.25
MonoFlex [51] (Baseline)	11.70	11.64	32.26	32.06	10.96	10.90	30.31	30.12
BA-Det(Ours)†	**16.60**	**16.45**	**40.93**	**40.51**	**15.57**	**15.44**	**38.53**	**38.12**

+ 43%

Condition on depth range

	Method	3D AP_{70}			3D APH_{70}			3D AP_{50}			3D APH_{50}		
		0-30	30-50	50-∞	0-30	30-50	50-∞	0-30	30-50	50-∞	0-30	30-50	50-∞
L1	DCD [21]	32.47	5.94	1.24	32.30	5.91	1.23	62.70	26.35	10.16	62.35	26.21	10.09
	MonoFlex [51]	30.64	5.29	1.05	30.48	5.27	1.04	61.13	25.85	9.03	60.75	25.71	8.95
	BA-Det(Ours)†	**37.74**	**11.04**	**3.86**	**37.46**	**10.95**	**3.79**	**71.07**	**37.15**	**14.89**	**70.46**	**36.79**	**14.61**
L2	DCD [21]	32.30	5.76	1.08	32.19	5.73	1.08	62.48	25.60	8.92	62.13	25.46	8.86
	MonoFlex [51]	30.54	5.14	0.91	30.37	5.11	0.91	60.91	25.11	7.92	60.54	24.97	7.85
	BA-Det(Ours)†	**37.61**	**10.72**	**3.37**	**37.33**	**10.63**	**3.31**	**70.83**	**36.14**	**13.62**	**70.23**	**35.79**	**13.37**

❖ BA-Det: Object-centric Temporal Optimization

3D Video Object Detection with Learnable Object-Centric Global Optimization (CVPR 2023)

❑ Ablation Study and Discussions

- Ablation study

	LEVEL_1			
	3D AP_{70}	3D APH_{70}	3D AP_{50}	3D APH_{50}
MonoFlex (baseline)	11.70	11.64	32.26	32.06
Our first-stage prediction	13.57	13.48	34.70	34.43
+3D Tracking	14.01	13.93	35.19	34.92
+ Learnable global optimization	15.85	15.75	38.06	37.76
+ Tracklet rescoring	16.43	16.30	40.07	39.70
+ Bbox interpolation	16.60	16.45	40.93	40.51

- ORB feature vs. our learnable feature

	$\bar{L}_t$	LEVEL_1			
		3D AP_{70}	3D APH_{70}	3D AP_{50}	3D APH_{50}
MonoFlex (baseline)	-	11.70	11.64	32.26	32.06
BA-Det+ ORB feature [34]	2.6	14.05	13.96	35.21	34.95
BA-Det+ Our feature	10	16.60	16.45	40.93	40.51

- Static (Scene-level) vs. Object-centric

	LEVEL_1			
	3D AP_{70}	3D APH_{70}	3D AP_{50}	3D APH_{50}
MonoFlex (baseline)	11.70	11.64	32.26	32.06
Initial prediction	13.57	13.48	34.70	34.43
Static BA	14.73	14.62	37.89	37.56
Ours	16.60	16.45	40.93	40.51

- Latency of each module

Total latency	181.5ms
First-stage detector	132.6ms
Object tracking	6.6ms
Feature correspondence	23.0ms
Object bundle adjustment	19.3ms

❖ BA-Det: Object-centric Temporal Optimization

3D Video Object Detection with Learnable Object-Centric Global Optimization (CVPR 2023)

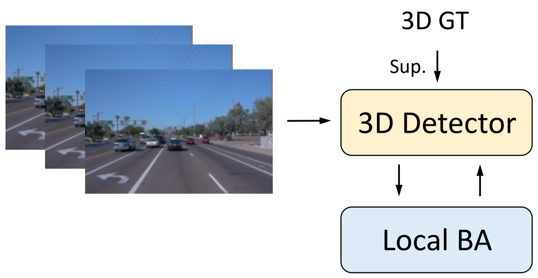
❑ Qualitative Results

❖ BA²-Det: From 3D Labels to 2D Labels

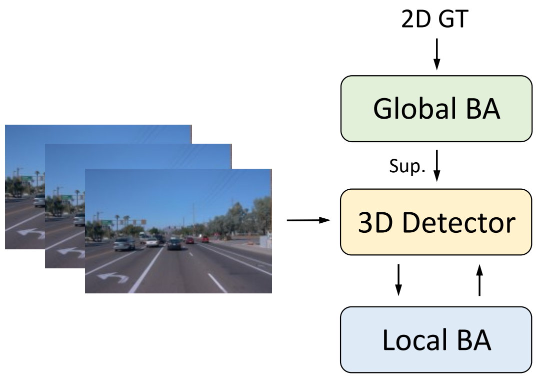
2D Supervised Monocular 3D Object Detection by Global-to-Local 3D Reconstruction (arXiv:2306.05418)

❑ Motivation

- Camera-based 3D object detector (e.g., BA-Det) depends on 3D labels
  - 3D labels → 2D labels: recover 3D structure from video
  - Scene-level global reconstruction + Object-level local reconstruction



3D-fully-supervised **BA-Det**



2D-supervised **BA²-Det**

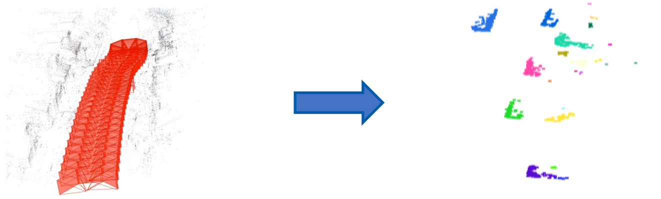
❖ BA²-Det: From 3D Labels to 2D Labels

2D Supervised Monocular 3D Object Detection by Global-to-Local 3D Reconstruction (arXiv:2306.05418)

❑ Notable Problems

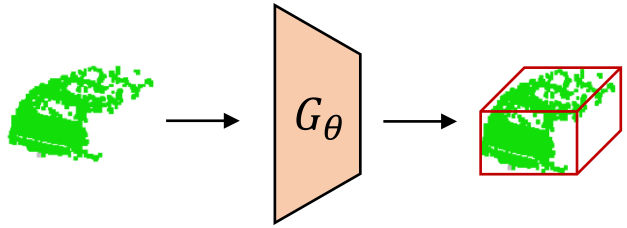
1. How to recover 3D location of each object with images?

   ✓ Global BA + object clustering on point cloud



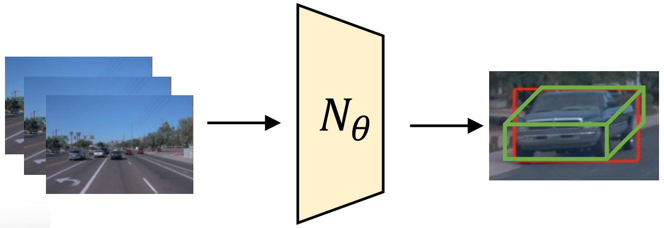
2. How to estimate 3D bounding boxes from object clusters?

   ✓ Fitting 3D pseudo boxes (complete objects) + learning



3. How to generalize static pseudo labels to dynamic objects?

   ✓ Video-based detector (Local BA) + iterative self-retraining

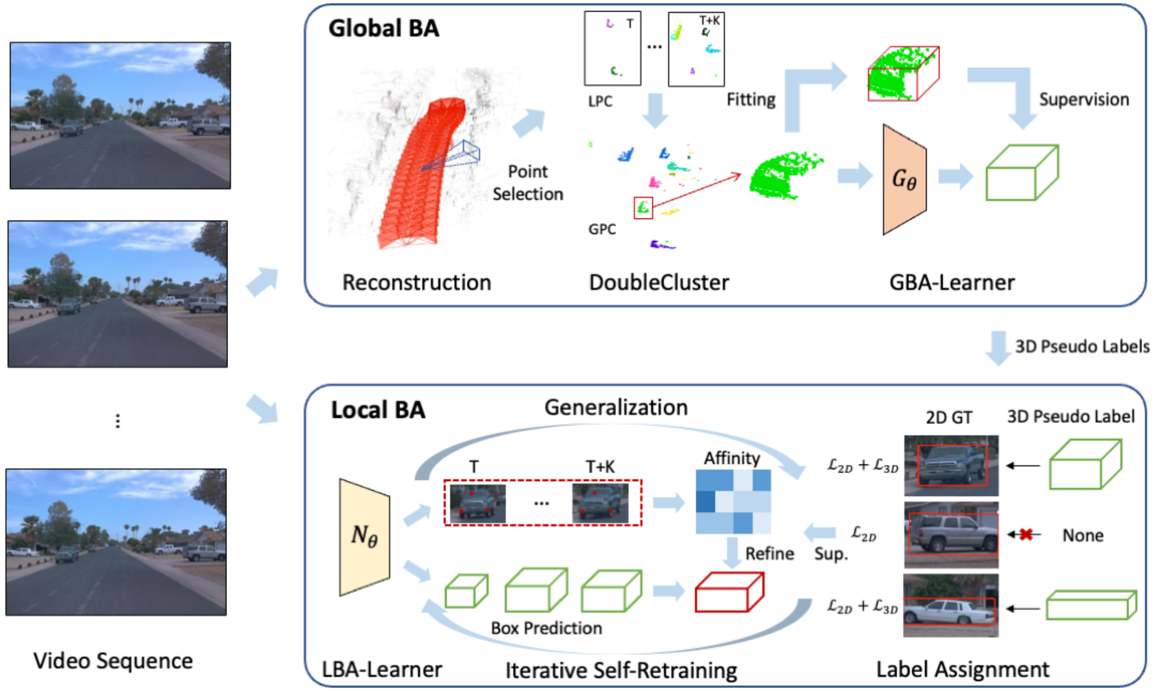


❖ BA²-Det: From 3D Labels to 2D Labels

2D Supervised Monocular 3D Object Detection by Global-to-Local 3D Reconstruction (arXiv:2306.05418)

❑ Global-to-local Pipeline

- ■ Global BA
  - ✓ Scene-level Structure-from-Motion
  - ✓ DoubleClustering
  - ✓ GBA-Learner (cluster → box)
- ■ Local BA
  - ✓ BA-Det with 2D label assignment (LBA-Learner)
  - ✓ Iterative self-retraining

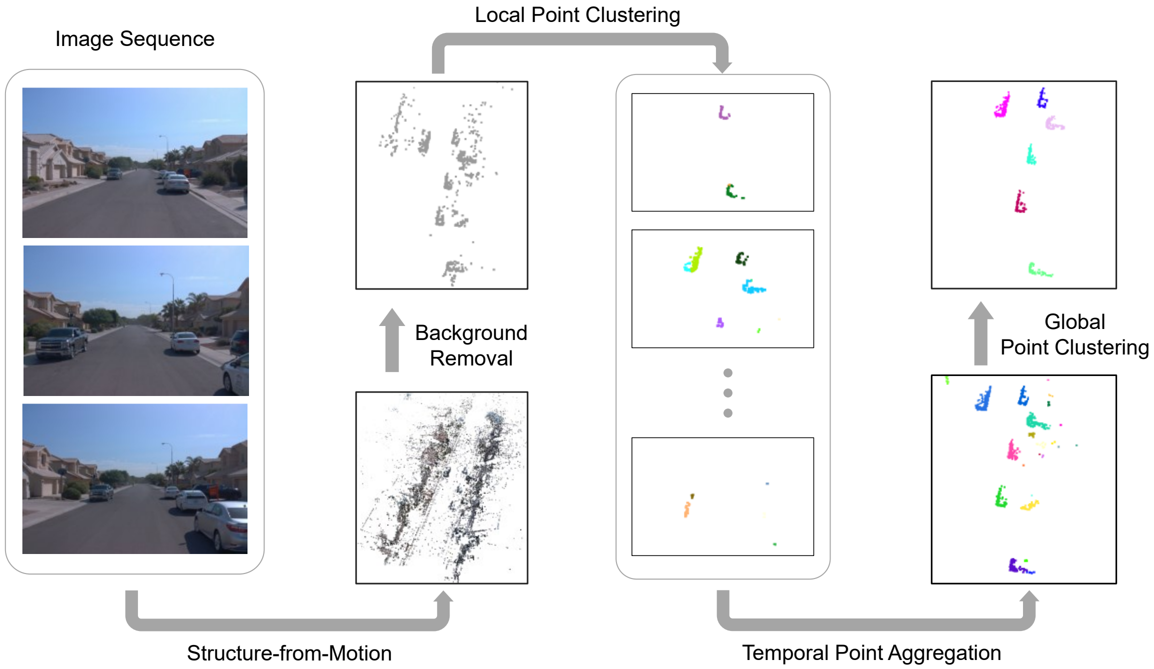


❖ BA²-Det: From 3D Labels to 2D Labels

2D Supervised Monocular 3D Object Detection by Global-to-Local 3D Reconstruction (arXiv:2306.05418)

❑ DoubleCluster

- Local Point Clustering
  - Main cluster for each object in each frame
- Global Point Clustering
  - Temporal cluster merging
  - Main cluster for each object in all frames



❖ BA²-Det: From 3D Labels to 2D Labels

2D Supervised Monocular 3D Object Detection by Global-to-Local 3D Reconstruction (arXiv:2306.05418)

❑ Experiments

➢ Main results on WOD

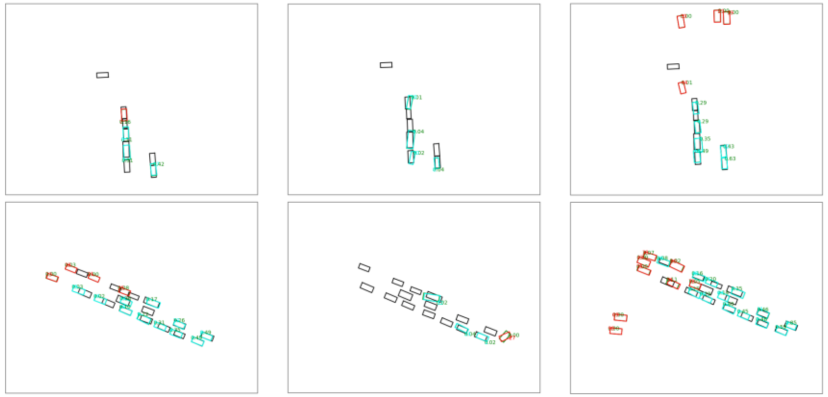
Method	3D Sup.	3D AP_5	3D APH_5	3D AP_{50}	3D APH_{50}	LET APL_{50}	LET AP_{50}	LET APH_{50}
PatchNet [28]	100%†	-	-	2.92	2.74	-	-	-
M3D-RPN [2]	100%†	-	-	3.79	3.63	-	-	-
PCT [47]	100%†	-	-	4.20	4.15	-	-	-
MonoJSG [22]	100%†	-	-	5.65	5.47	-	-	-
GUPNet [27]	100%†	-	-	10.02	9.94	-	-	-
BA-Det [10] Stage 1	100%	70.33	69.41	34.70	34.43	50.63	67.30	66.50
BA-Det [10] Stage 1	10%	53.68	52.30	15.44	15.22	28.21	44.21	43.23
BA-Det [10]	100%	72.96	71.78	40.93	40.51	54.45	68.32	67.36
BA-Det [10]	10%	57.29	55.27	19.70	19.27	32.53	46.91	45.52
SfM [42]+BA-Det [10]	0%	27.84	8.80	2.89	0.75	7.34	10.75	3.31
BA²-Det(Ours)	0%	60.01	44.81	10.39	8.98	22.24	32.60	23.86

+ 115%   Outperform some 3D fully-supervised methods

Method	3D Sup.	3D AP_5 0-30	3D AP_5 30-50	3D AP_5 50-∞	3D APH_5 0-30	3D APH_5 30-50	3D APH_5 50-∞	LET APL_{50} 0-30	LET APL_{50} 30-50	LET APL_{50} 50-∞	LET AP_{50} 0-30	LET AP_{50} 30-50	LET AP_{50} 50-∞
BA-Det [10]	100%	87.80	72.52	48.45	86.91	71.52	46.98	66.15	57.97	36.44	82.74	69.58	45.77
BA-Det [10]	10%	73.25	54.00	34.50	71.38	52.22	32.53	38.31	35.57	22.40	56.98	47.28	31.11
SfM [42]+BA-Det [10]	0%	46.87	25.88	9.09	14.26	8.86	2.84	11.35	7.74	2.60	17.59	10.12	3.48
BA²-Det (Ours)	0%	77.38	54.95	33.74	64.54	37.57	21.64	25.00	23.97	14.63	39.24	31.73	20.30

➢ Ablation study

N_θ w/ 3D	N_θ w/ 2D	G_θ	r_y w/ d	Iter.	OBA	3D AP_5	3D APH_5	LET APL_{50}	LET AP_{50}
✓						20.97	6.70	4.27	7.28
	✓					28.40	11.34	5.02	8.62
	✓	✓				33.75	11.94	9.63	16.80
	✓	✓	✓			41.17	28.73	12.23	21.41
	✓	✓	✓	✓		56.33	42.05	17.87	29.62
	✓	✓	✓	✓	✓	60.01	44.81	22.24	32.60



(a) BA-Det (10% labeled videos).   (b) SfM + BA-Det (Baseline).   (c) BA²-Det (Ours).

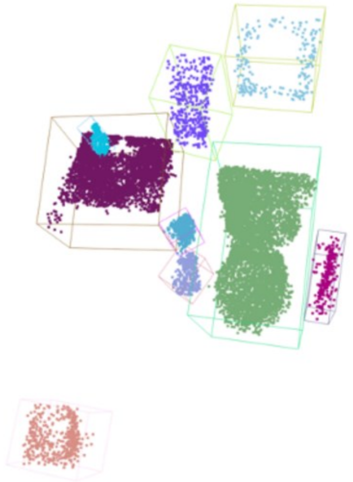
❖ BA²-Det: From 3D Labels to 2D Labels

2D Supervised Monocular 3D Object Detection by Global-to-Local 3D Reconstruction (arXiv:2306.05418)

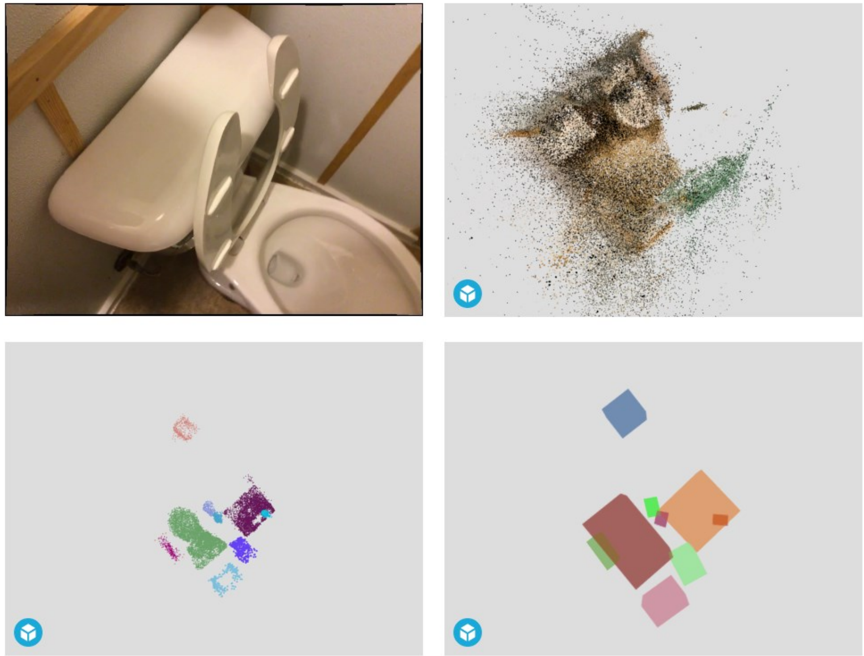
❑ Experiments

- Using SAM instead of 2D gt boxes

(a) Input image examples.



(b) Detected 3D boxes from the video.



❖ BA²-Track: Association with Pseudo 3D Location

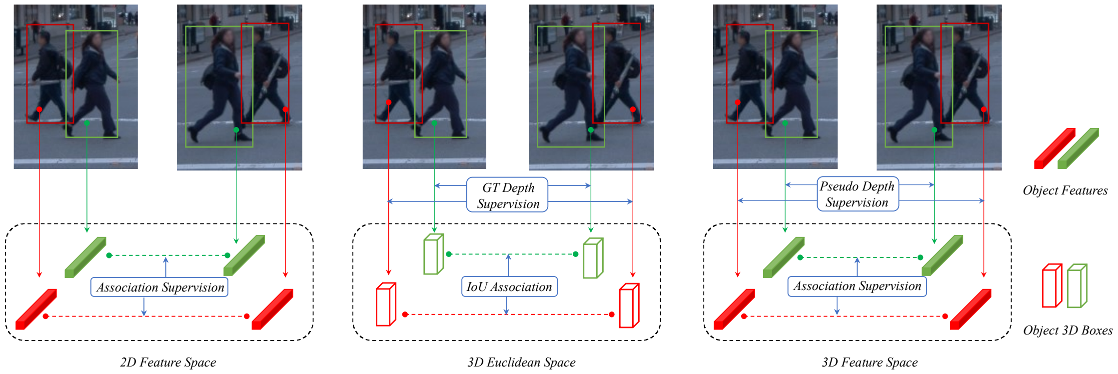
Tracking Objects with 3D Representation from Videos (arXiv:2306.05416)

❑ Motivation

- Association is ***hard*** for ***2D*** multiple object tracking
  - Object occlusion/inaccurate motion model
- Lift 2D object in 3D space
- BA²-Det is an example to obtain 3D representation in any 2D video with ego-motion (even w/o off-the-shelf depth)
- BA²-Det + association learning

*3D association is **much easier***

Method	IDS	Early Termination	Wrong Assosication
CenterPoint	2891	2890	1
Immortal Tracker(Ours)	114	113	1



❖ BA²-Track: Association with Pseudo 3D Location

Tracking Objects with 3D Representation from Videos (arXiv:2306.05416)

❑ Object Association with Pseudo 3D Representation

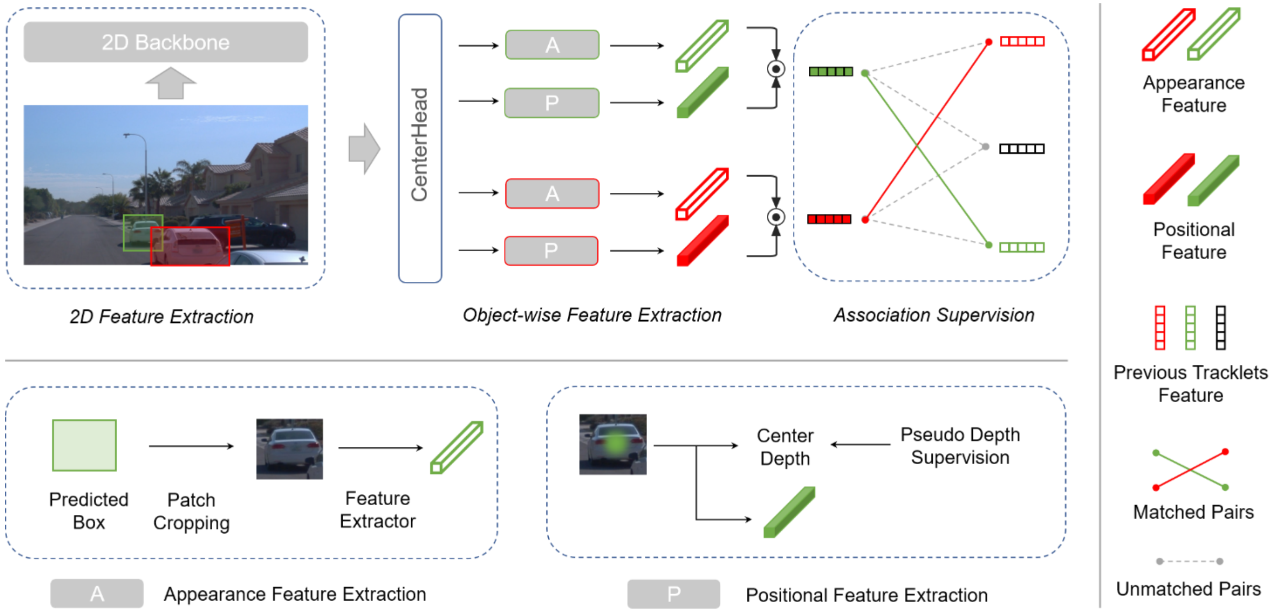
- 3D object feature from center
- 2D/3D feature concatenation

$$^{(0)}\mathbf{f}_j^t = [^{(2D)}\mathbf{f}_j^t, ^{(3D)}\mathbf{f}_j^t]$$

- Cross-graph GCN

$$^{(l+1)}\mathbf{f}_j^t = \mathrm{MLP}(^{(l)}\mathbf{f}_j^t + \frac{||^{(l)}\mathbf{f}_j^t||_2 {}^{(l)}\mathbf{m}_j^{t-1}}{||^{(l)}\mathbf{m}_j^{t-1}||_2}), l \in [0, L-1]$$

- Graph matching from GMTracker



❖ BA^2-Track: Association with Pseudo 3D Location

Tracking Objects with 3D Representation from Videos (arXiv:2306.05416)

❑ Experiments

■ 2D MOT

SOTA performance on WOD and KITTI

Method	Backbone	Split	Category	MOTA ↑	IDF1 ↑
IoU baseline [29]	ResNet-50	val	Vehicle	38.25	-
Tracktor++ [1, 29]	ResNet-50	val	Vehicle	42.62	-
RetinaTrack [29]	ResNet-50	val	Vehicle	44.92	-
QDTrack [12]	ResNet-50	val	Vehicle	55.6	66.2
P3DTrack (Ours)	DLA-34	val	Vehicle	55.9	65.6

Method	+Label	+Data	HOTA	AssA	ID Sw.	MOTA
QD-3DT [19]	3D GT		72.77	72.19	206	85.94
Mono3DT [18]	3D GT		73.16	74.18	379	84.28
OC-SORT [4]		PD	76.54	76.39	250	90.28
PermaTrack [49]		PD	78.03	78.41	258	91.33
RAM [48]		PD	79.53	80.94	210	91.61
QDTrack [12]			68.45	65.49	313	84.93
TrackMPNN [39]			72.30	70.63	481	87.33
CenterTrack [66]			73.02	71.20	254	88.83
LGM [50]			73.14	72.31	448	87.60
DEFT [5]			74.23	73.79	344	88.38
P3DTrack (Ours)			**74.59**	**76.86**	**219**	85.60

■ 3D MOT

Better than monocular 3D MOT method QD-3DT

	Fully Sup.	MOTA$_{50}$ ↑	Mismatch$_{50}$ ↓	MOTA$_{30}$ ↑	Mismatch$_{30}$ ↓
QD-3DT [11]	✓	0.0308	0.00550	0.1867	0.01340
CC-3DT [9]	✓	0.0480	0.00180	0.2032	0.00690
SfM [42]+BA-Det [10]+Immortal [48]		0.0011	<0.00001	0.0652	0.00038
BA^2-Det (Ours)		0.0352	0.00002	0.1522	0.00008

■ Ablation study

+ 1.1 MOTA and 1.6 IDF1 with pseudo 3D rep.

	MOTA ↑	IDF1 ↑	FP ↓	FN ↓	ID Sw. ↓
Baseline	51.0	62.3	8709	331056	9100
+low-quality dets [62]	53.4	64.3	13058	309653	9005
+GNN	56.5	66.5	20381	278752	10129
+3D representation	57.6	68.1	33587	258066	9920

Better than pretrained/geometric-based depth

3D rep from	MOTA ↑	IDF1 ↑	FP ↓	FN ↓	ID Sw. ↓
P3DTrack (Ours)	57.6	68.1	33587	258066	9920
SfM [44]	55.5	66.8	55823	249935	11145
MiDaS v3 [38]	55.7	63.4	34514	259471	21058

❖ References

- Yingyan Li, Yuntao Chen, **Jiawei He**, Zhaoxiang Zhang. Densely Constrained Depth Estimator for Monocular 3D Object Detection. In European Conference on Computer Vision (**ECCV**) 2022.
- **Jiawei He**, Yuntao Chen, Naiyan Wang, Zhaoxiang Zhang. *3D Video Object Detection with Learnable Object-Centric Global Optimization.* In IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**) 2023.
- **Jiawei He**, Lue Fan, Yuqi Wang, Yuntao Chen, Zehao Huang, Naiyan Wang, Zhaoxiang Zhang. *Tracking Objects with 3D Representation from Videos.* arXiv:2306.05416.
- **Jiawei He**, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang. *2D Supervised Monocular 3D Object Detection by Global-to-Local 3D Reconstruction.* arXiv:2306.05418.

❖ Summary

- BA-Det → BA^2-Det → BA^2-Track
- 3D detection → 2D/3D MOT
- Fully supervise → Weakly-supervise → Self-supervise
- Object reconstruction → Scene+object reconstruction

Demos: https://ba2det.site/