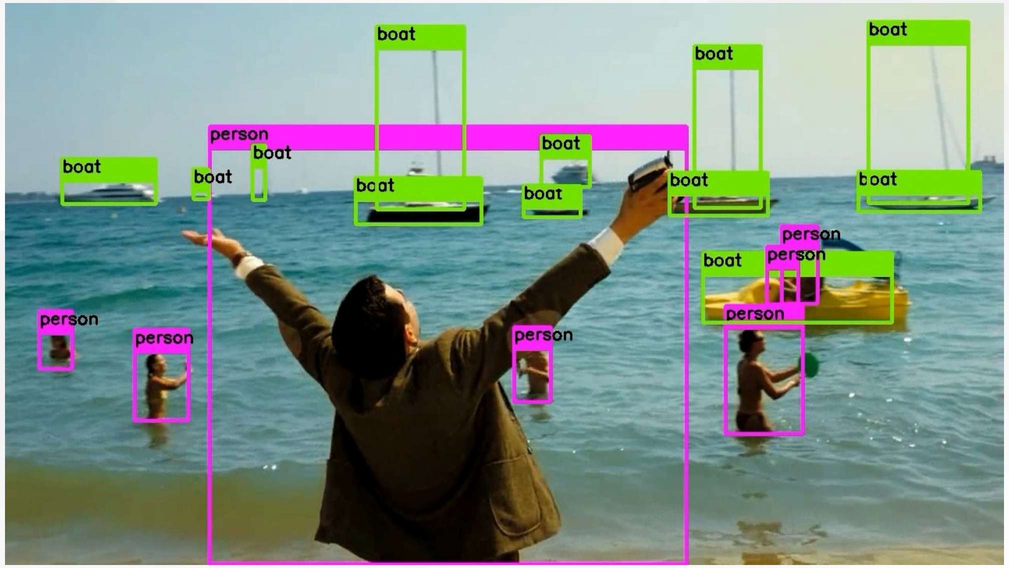
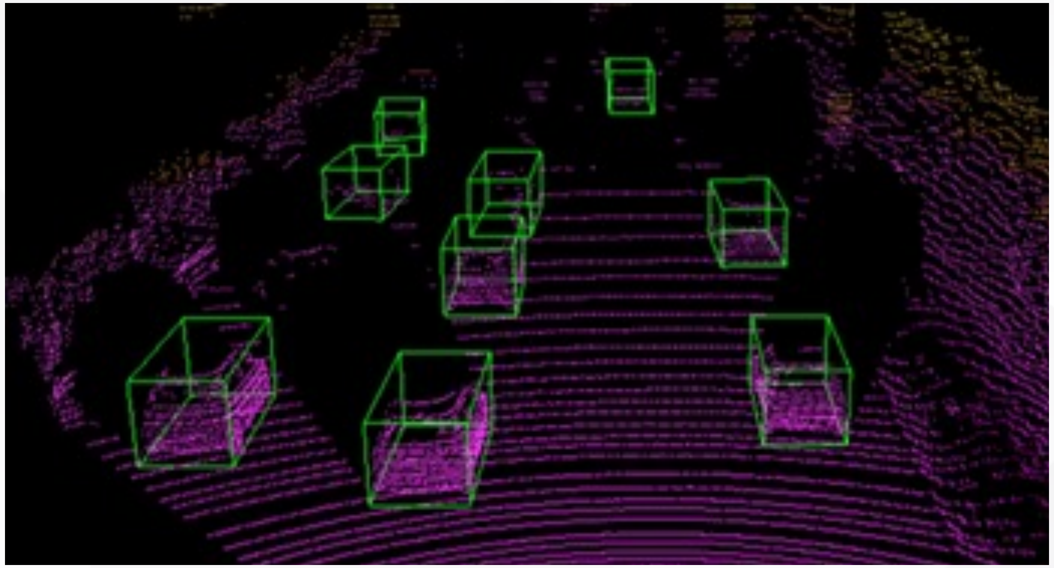
Cross-Modal 3D Object Detection and Tracking

Chao Ma

Shanghai Jiao Tong University

3D Object Detection

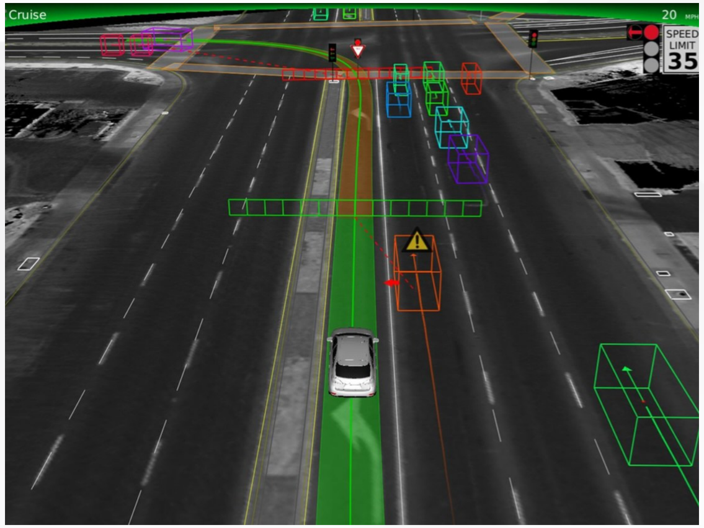




- **Input：** 2D Images
- **Information：** (R, G, B)
- **Dense/Sparse：** Dense
- **Output：** 2D BBX、location
- **DOF：** 4

- **Input：** 2D Images/Cloud Points/…
- **Information：** (R, G, B | X, Y, Z, I, …)
- **Dense/Sparse：** Dense Image & Sparse Points
- **Output：** 3D BBX、Location、Orientation、Speed
- **DOF：** 9 （Decrease to 7 when ground is fixed）

3D Object Detection Applications



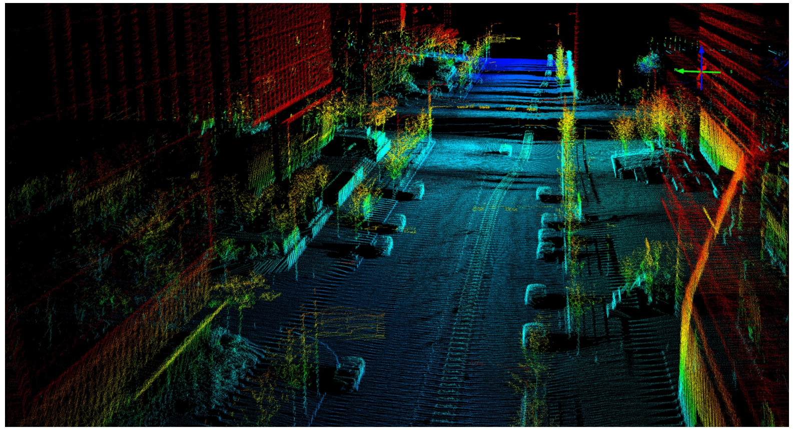
Auto-Driving

AR / VR

Robotics

Background: 3D Object Detection

LiDAR



- Modality： Point cloud
- Input： (X, Y, Z, I, …)
- Advantages： accurate location
- Disadvantages： sparse, unordered

?

Fusion

Camera

- Modality： 2D Image
- Input： (R, G, B, …)
- Advantages： dense, rich semantics
- Disadvantages： lack of depth

Fusion-based 3D Object Detection

人工智能研究院
Artificial Intelligence Institute

1 Result Level

Methods: adopt off-the-shelf 2D object detectors.
Disadvantages: The performance of 2D detectors set an upper bound on 3D detection.

- ✓ F-PointNets 2018 CVPR
- ✓ F-ConvNet 2019 IROS

2 Proposal Level

Methods: perform fusion at the region proposal level
Disadvantages: slow and cumbersome

- ✓ MV3D 2017 CVPR
- ✓ AVOD 2018 IROS

3 Point Level

Methods: fetch point-wise image features by projecting point clouds onto image plane.

a

Methods: construct BEV camera features before fusing with LiDAR BEV features.
Disadvantages: Feature blurring

- ✓ ContFuse 2018 ECCV
- ✓ MMF 2019 CVPR
- ✓ 3D-CVF 2020 ECCV

b

Methods: augment each LiDAR point with image features or segmentation scores.

- ✓ MVX-Net 2019 ICRA
- ✓ **PointPainting 2020 CVPR**

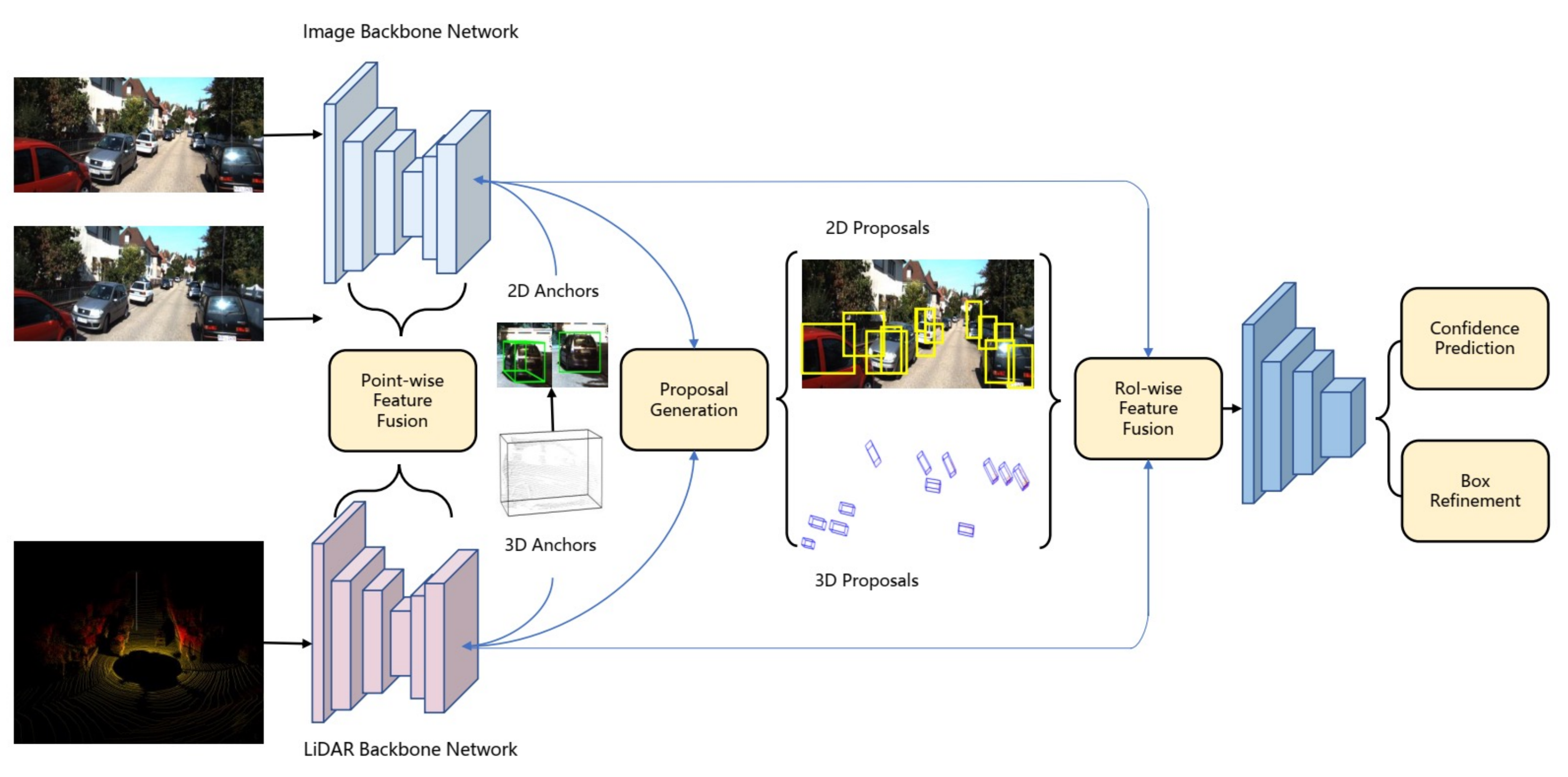
Observations on KITTI

3D Detection results on KITTI

Method	Modality	3D AP(%)			2D AP(%)		
		Easy	Moderate	Hard	Easy	Moderate	Hard
MV3D	RGB+LiDAR	71.09	62.35	55.12	-	-	-
AVOD	RGB+LiDAR	73.59	65.78	58.38	95.17	89.88	82.83
AVOD-FPN	RGB+LiDAR	81.94	71.88	66.38	94.70	88.92	84.13
F-PointNet	RGB+LiDAR	81.20	70.39	62.19	95.85	**95.17**	85.42
ContFuse	RGB+LiDAR	82.54	66.22	64.04	-	-	-
VoxelNet	LiDAR	77.49	65.11	57.73	-	-	-
Second	LiDAR	83.13	73.66	66.20	93.72	90.68	85.63
PointPillars	LiDAR	82.58	74.31	68.99	94.00	91.19	88.17
PointRCNN	LiDAR	86.96	75.64	70.70	95.92	91.90	87.11

Cloud point 3D detectors perform better than cross-modal approaches

Two Stage Cross-Modal Fusion



- Stage 1: Point-pixel fusion for proposal generation
- Stage 2: ROI-wise feature fusion for 3d bounding box refinement

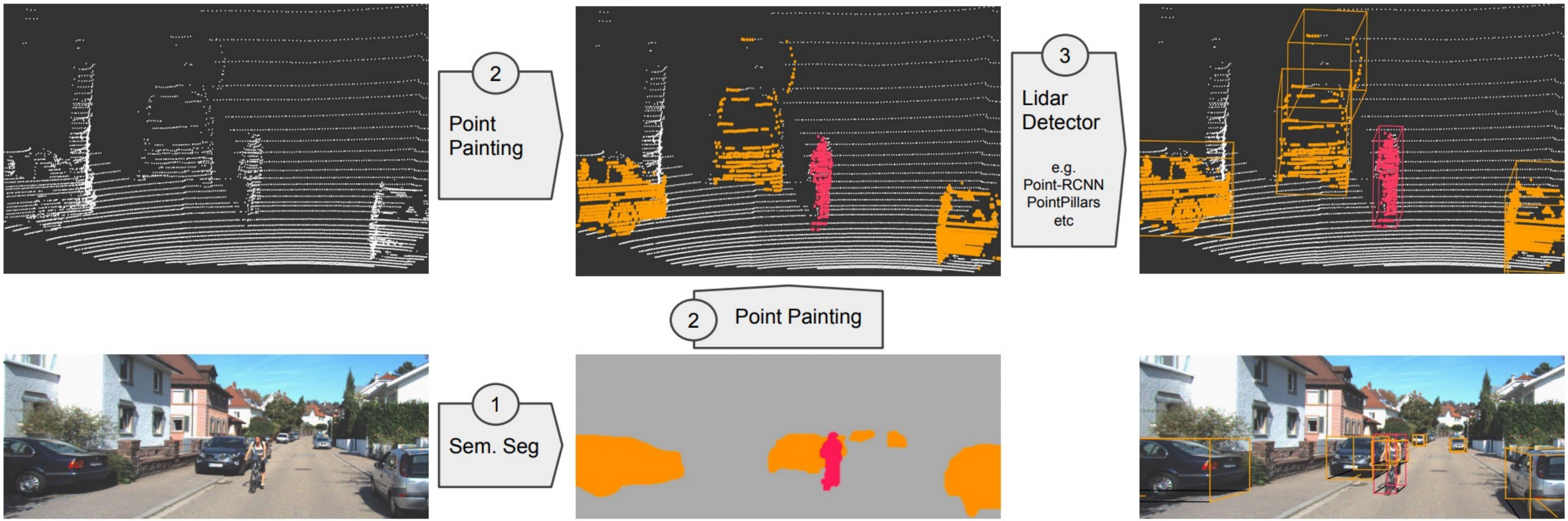
Ming Zhu, Chao Ma*, Pan Ji, Xiaokang Yang, Cross-Modality 3D Object Detection, in WACV 2021

Results on KITTI

Method	Modality	3D AP(%)			2D AP(%)		
		Easy	Moderate	Hard	Easy	Moderate	Hard
MV3D	RGB+LiDAR	71.09	62.35	55.12	-	-	-
AVOD	RGB+LiDAR	73.59	65.78	58.38	95.17	89.88	82.83
AVOD-FPN	RGB+LiDAR	81.94	71.88	66.38	94.70	88.92	84.13
F-PointNet	RGB+LiDAR	81.20	70.39	62.19	95.85	**95.17**	85.42
ContFuse	RGB+LiDAR	82.54	66.22	64.04	-	-	-
VoxelNet	LiDAR	77.49	65.11	57.73	-	-	-
Second	LiDAR	83.13	73.66	66.20	93.72	90.68	85.63
PointPillars	LiDAR	82.58	74.31	68.99	94.00	91.19	88.17
PointRCNN	LiDAR	86.96	75.64	70.70	95.92	91.90	87.11
Ours	RGB+LiDAR	**87.22**	**77.28**	**72.04**	**96.21**	93.45	**88.68**

Ming Zhu, Chao Ma*, Pan Ji, Xiaokang Yang, Cross-Modality 3D Object Detection, in WACV 2021

Image Representation for Lidar Points



Sourabh Vora, Alex H. Lang, Bassam Helou, and Oscar Beijbom, PointPainting: Sequential Fusion for 3D Object Detection, in CVPR 2020

Image Representation for Lidar Points

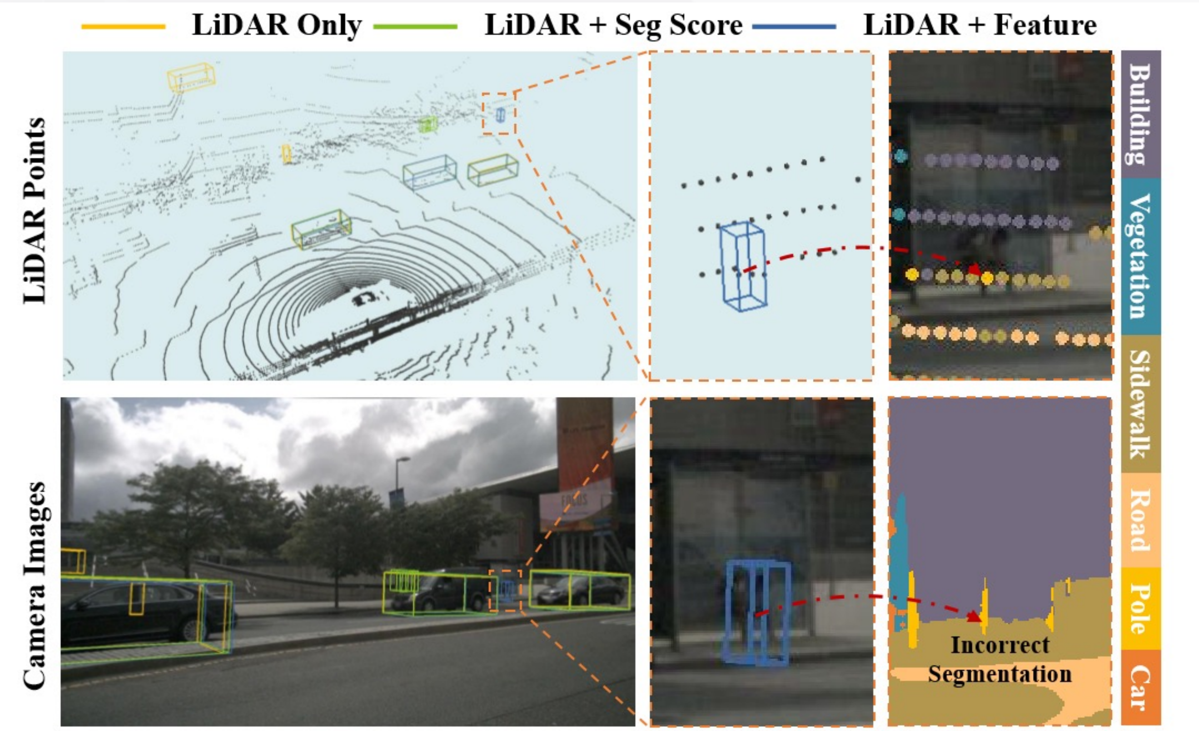
Segmentation Scores

- Provide semantic labels
- Straightforward and compact semantic cues

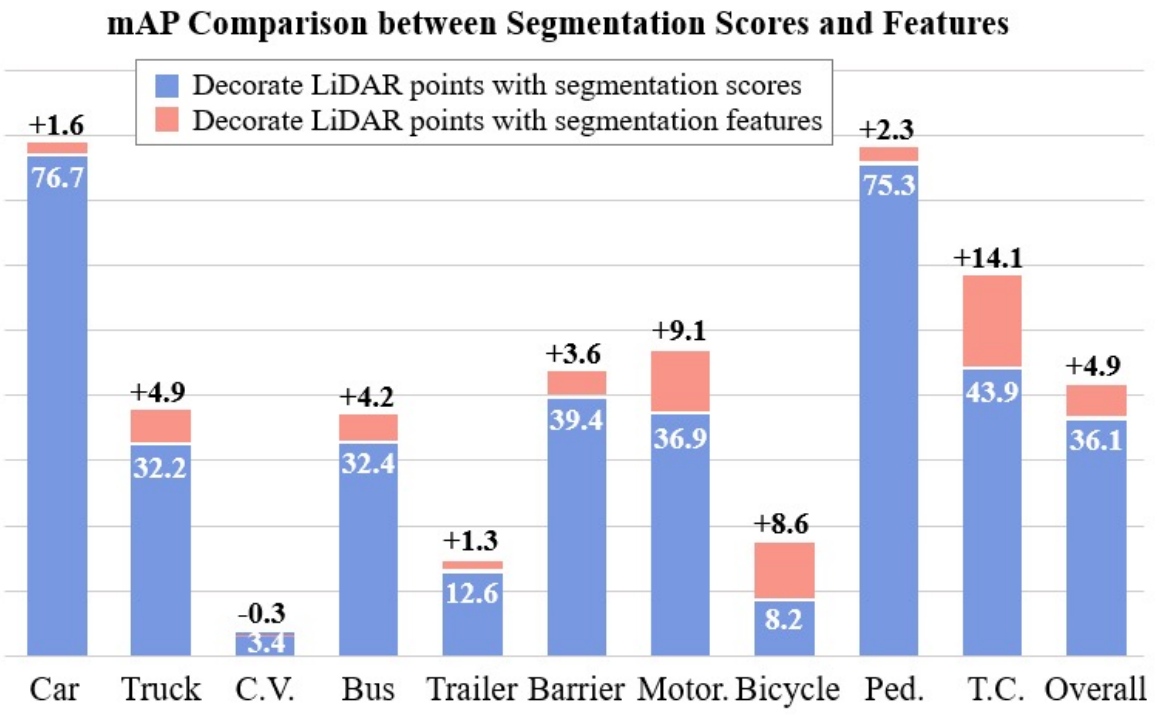
VS

CNN Features

- Provide richer semantic cues rather than the object class only
- Larger receptive field



- ***PointPainting fails due to segmentation failures on small objects***



- ***CNN Feature is better than Segmentation scores***

Motivation from CNN Training

Data Augmentation


Input


Zoom in/out


H-Flipping


V-Flipping


Random Crop


Rotation

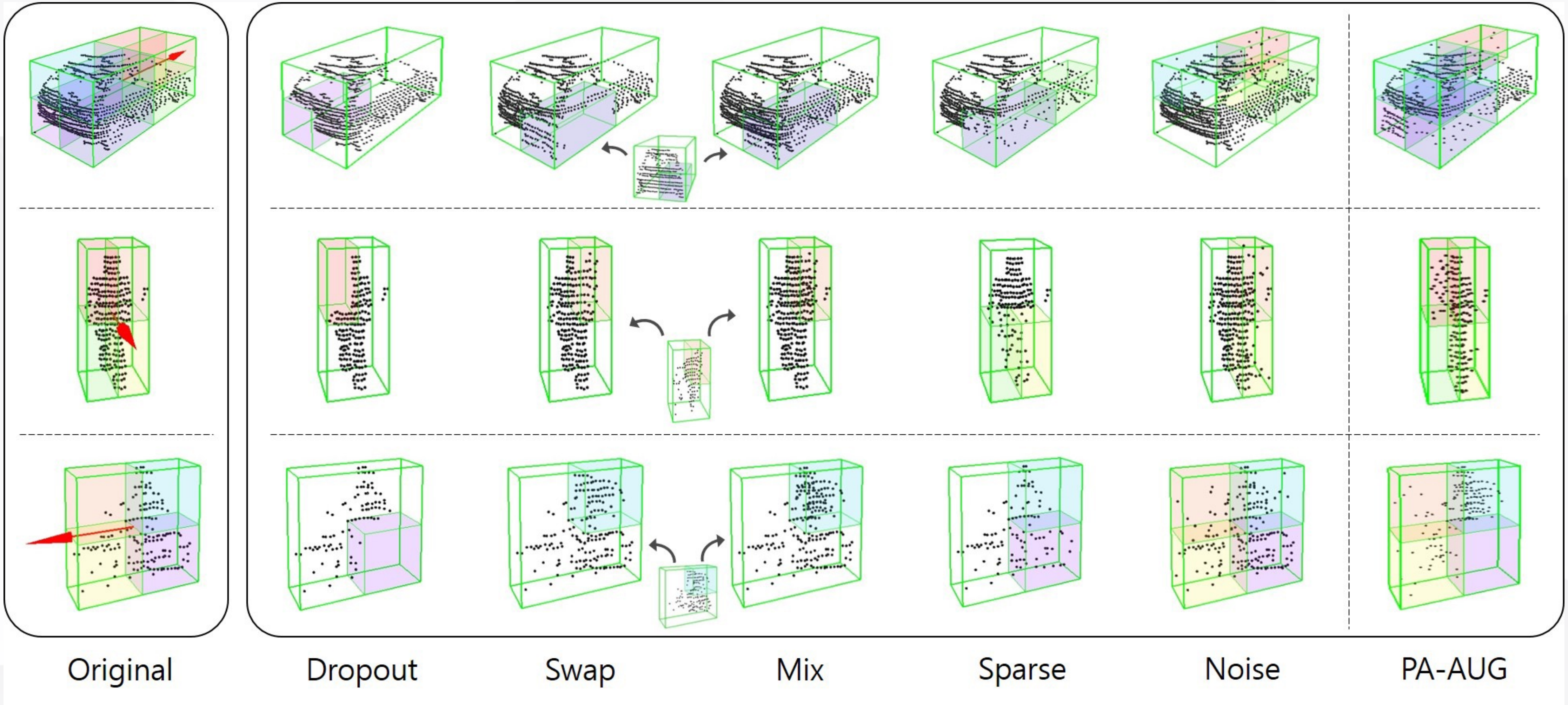

Coloring


Padding


CutMix

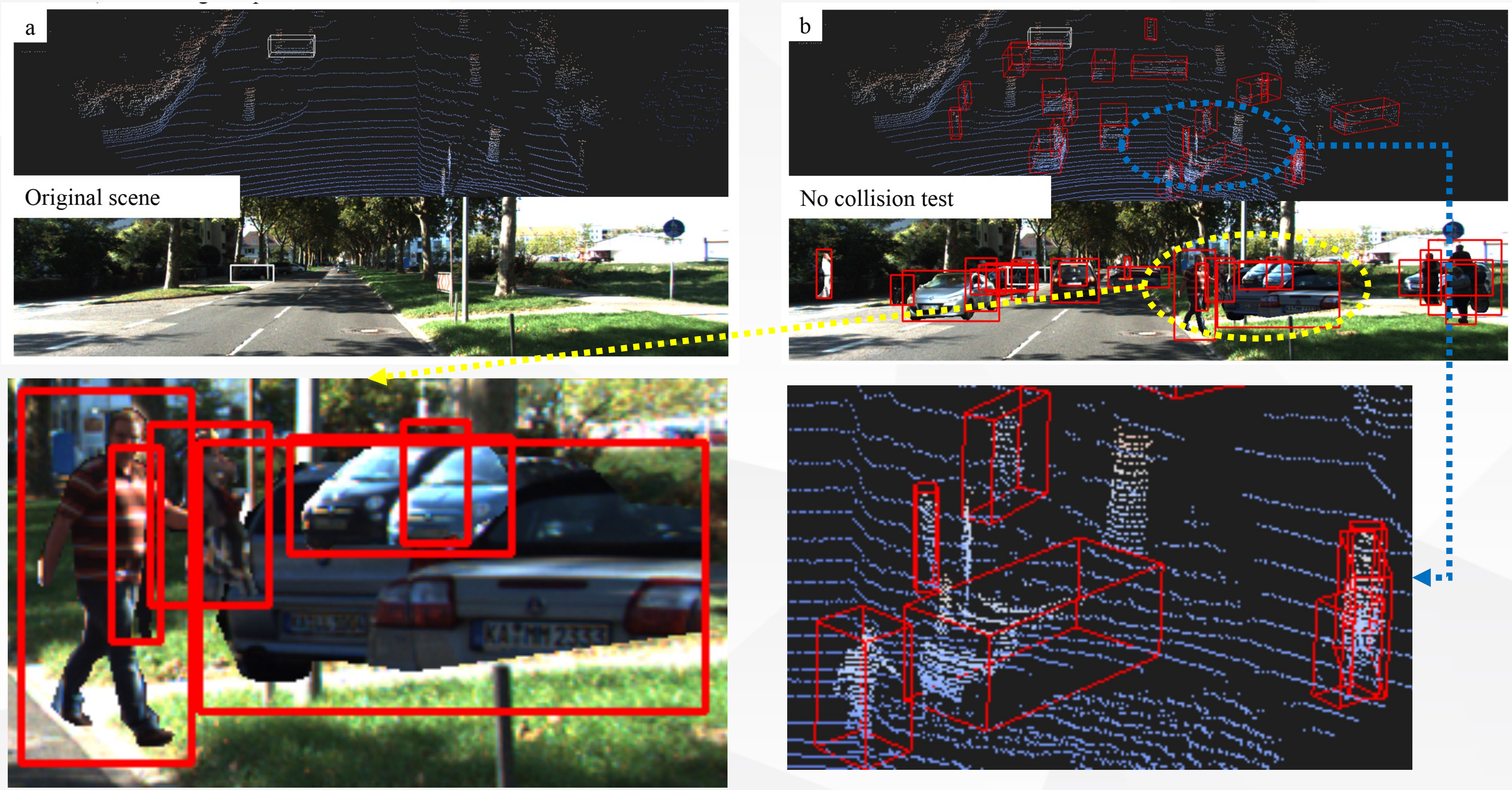

CutOut

Data Augmentation on Cloud Points



Challenge on Cross-Modal Data Augmentation

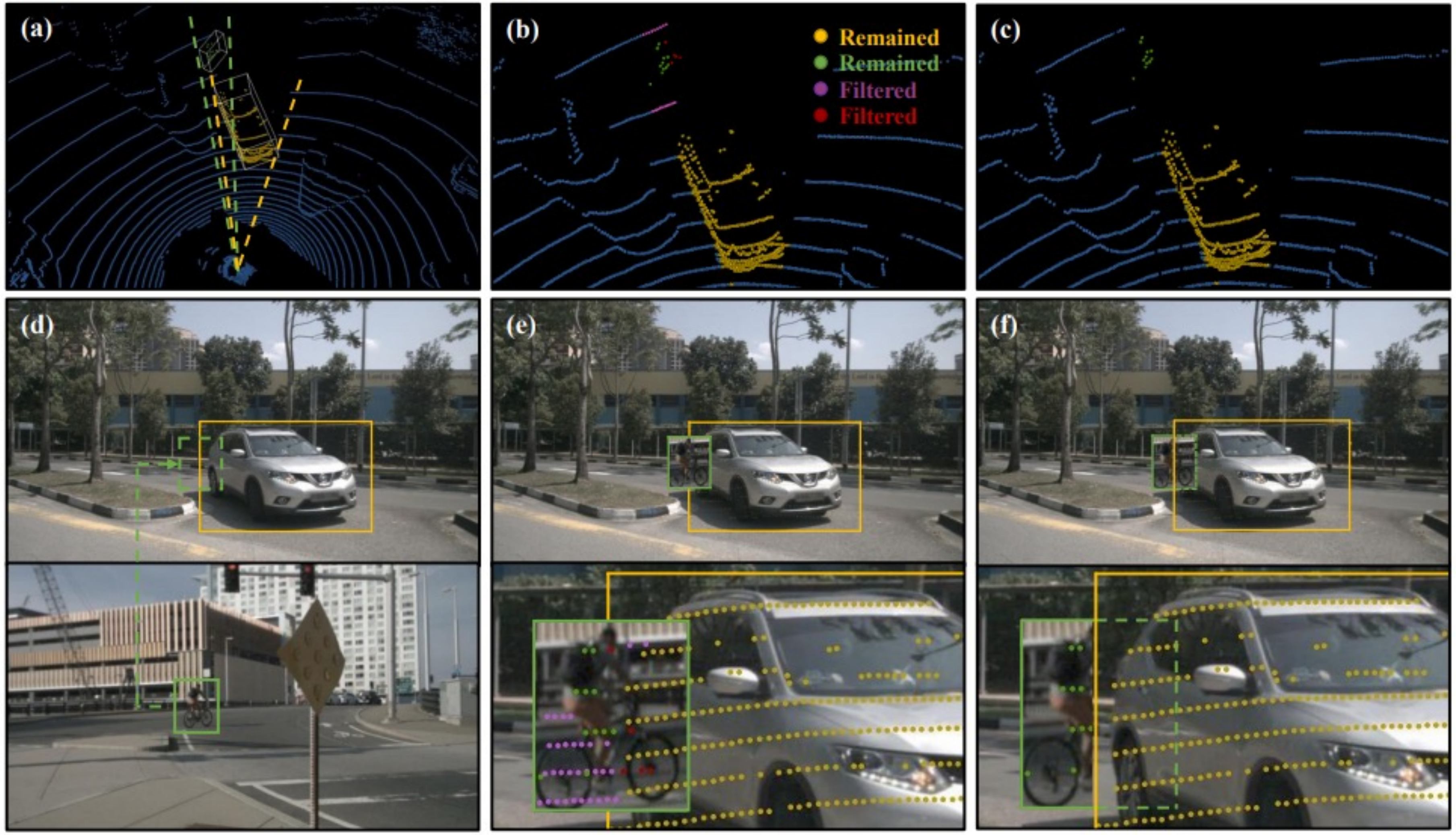
人工智能研究院
Artificial Intelligence Institute



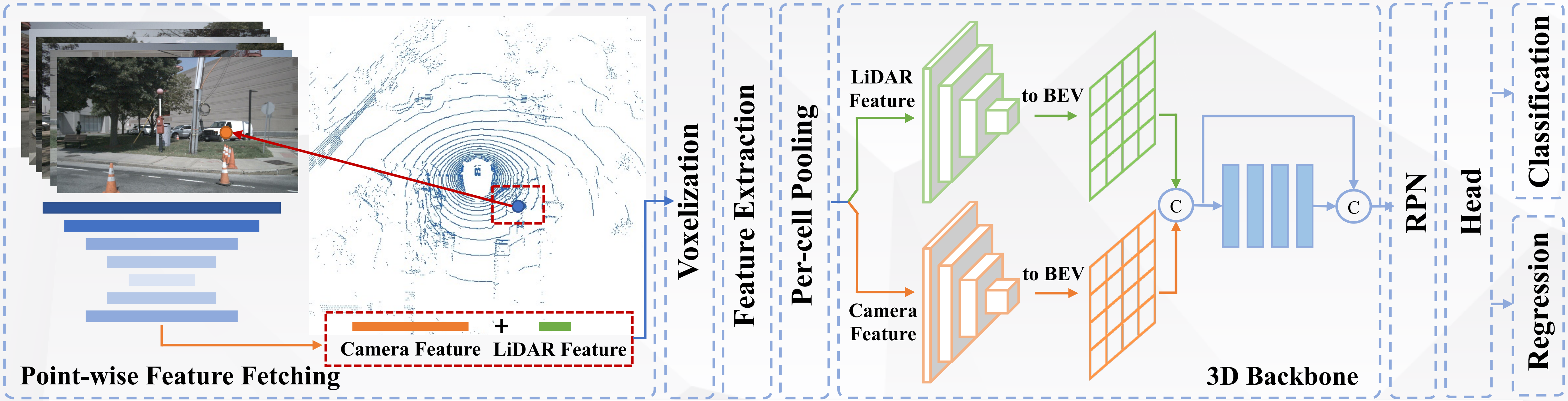
Data Augmentation for Cross-modality

人工智能研究院
Artificial Intelligence Institute

- **Methods**: simultaneously attach a virtual object onto Lidar scene and images.
- **Challenge**: consistency preservation between camera and LiDAR data.



PointAugmenting Network Architecture



- **Lidar only Baseline:** CenterPoint
- **Point-wise Feature Fetching**: . LiDAR points are projected onto image plane and then appended by the fetched point-wise CNN features
- **3D Detection**: a late fusion mechanism across modalities

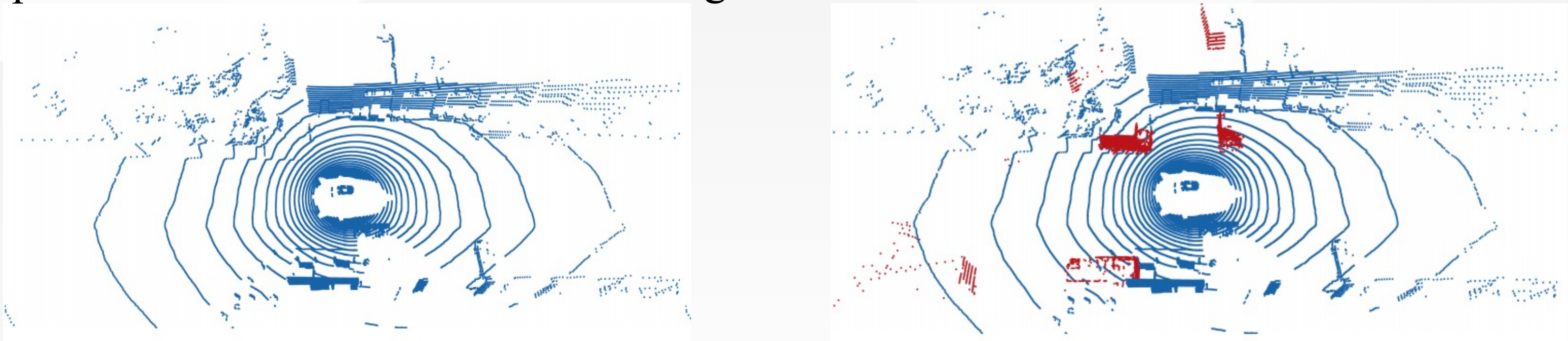
Chunwei Wang, Chao Ma*, Ming Zhu, Xiaokang Yang, PointAugmenting: Cross-Modal Augmentation for 3D Object Detection, in CVPR 2021

Data Augmentation for Cross-modality

人工智能研究院
Artificial Intelligence Institute

- **Data Augmentation for Lidar Points**

GT-Paste: pastes virtual objects in the forms of ground-truth boxes and LiDAR points from other scenes to the training scenes.



Method	Car	Truck	C.V.	Bus	Trailer	Barrier	Motor.	Bicycle	Ped.	T.C.	mAP	NDS
CenterPoint w/o GT-Paste	74.2	30.9	3.7	27.0	12.5	37.2	30.3	1.7	68.2	42.4	32.8	42.3
CenterPoint w/ GT-Paste	78.6	39.2	2.0	33.5	13.5	46.8	32.2	8.6	74.2	47.5	37.6	49.5
Gains of GT-Paste	+4.4	+8.3	-1.7	+6.5	+1.0	+9.6	+1.9	+6.9	+6.0	+5.1	+4.8	+7.2

Table 1. Effectiveness of the GT-Paste data augmentation scheme. Applying GT-Paste data augmentation for LiDAR points achieves an improvement of +4.8% 3D mAP. We use CenterPoint as baseline with 1/8 training data on the nuScenes dataset.

Extend to Cross-modality – Consistency Destruction

propose a simple yet effective cross-modal augmentation method to make GT-Paste applicable to both point clouds and images.

SJTU

nuScenes datatset

- *Rank 2 on nuScenes Leaderboard (rank 1 with single model)*

Method	mAP	NDS	Car	Truck	C.V.	Bus	Trailer	Barrier	Motor.	Bicycle	Ped.	T.C.
PointPillars [9]	30.5	45.3	68.4	23.0	4.1	28.2	23.4	38.9	27.4	1.1	59.7	30.8
3DSSD [25]	42.6	56.4	81.2	47.2	12.6	61.4	30.5	47.9	36.0	8.6	70.2	31.1
PointPainting [19]	46.4	58.1	77.9	35.8	15.8	36.2	37.3	60.2	41.5	24.1	73.3	62.4
CBGS [35]	52.8	63.3	81.1	48.5	10.5	54.9	42.9	65.7	51.5	22.3	80.1	70.9
CenterPoint [27]	60.3	67.3	85.2	53.5	20.0	63.6	56.0	71.1	59.5	30.7	84.6	78.4
Ours	**66.8**	**71.0**	**87.5**	**57.3**	**28.0**	**65.2**	**60.7**	**72.6**	**74.3**	**50.9**	**87.9**	**83.6**

Table 2. Performance comparisons of 3D object detection on the nuScenes test set. We report the NDS, mAP, and mAP for each class.

Waymo datatset

Method	Vehicle		Pedestrian		Cyclist		All	
	L1 mAP	L2 mAP	L1 mAP	L2 mAP	L1 mAP	L2 mAP	L1 mAP/mAPH	L2 mAP/mAPH
CenterPoint [27]	66.70	62.00	73.55	68.64	72.51	70.00	70.92 / 68.26	66.88 / 64.36
Ours	67.41	62.70	75.42	70.55	76.29	74.41	73.04 / 70.39	69.22 / 66.70
Gains of fusion	+0.71	+0.70	+1.87	+1.91	+3.78	+4.41	+2.12 / +2.13	+2.34 / +2.34

Table 3. Performance comparisons of 3D object detection on the Waymo validation set. We show the mAP and mAPH in the L1 and L2 difficulty levels. The results of CenterPoint are reproduced by ourselves.

Ablation Studies

人工智能研究院
Artificial Intelligence Institute

1 Cross-Modal Network Design

	Seg Score	DetFeat.	CC	LF	mAP	NDS
(a)					37.4	49.9
(b)	✓		✓		42.3	51.4
(c)		✓	✓		46.0	53.9
(d)		✓		✓	47.5	55.6

Table 4. Comparison of fusion policies. Seg Score: decorating LiDAR points with segmentation scores as suggested by PointPainting [19]. DetFeat: decorating LiDAR points with image features from the detection task. CC: fusing LiDAR and image features by point-wise concatenation. LF: our late fusion mechanism.

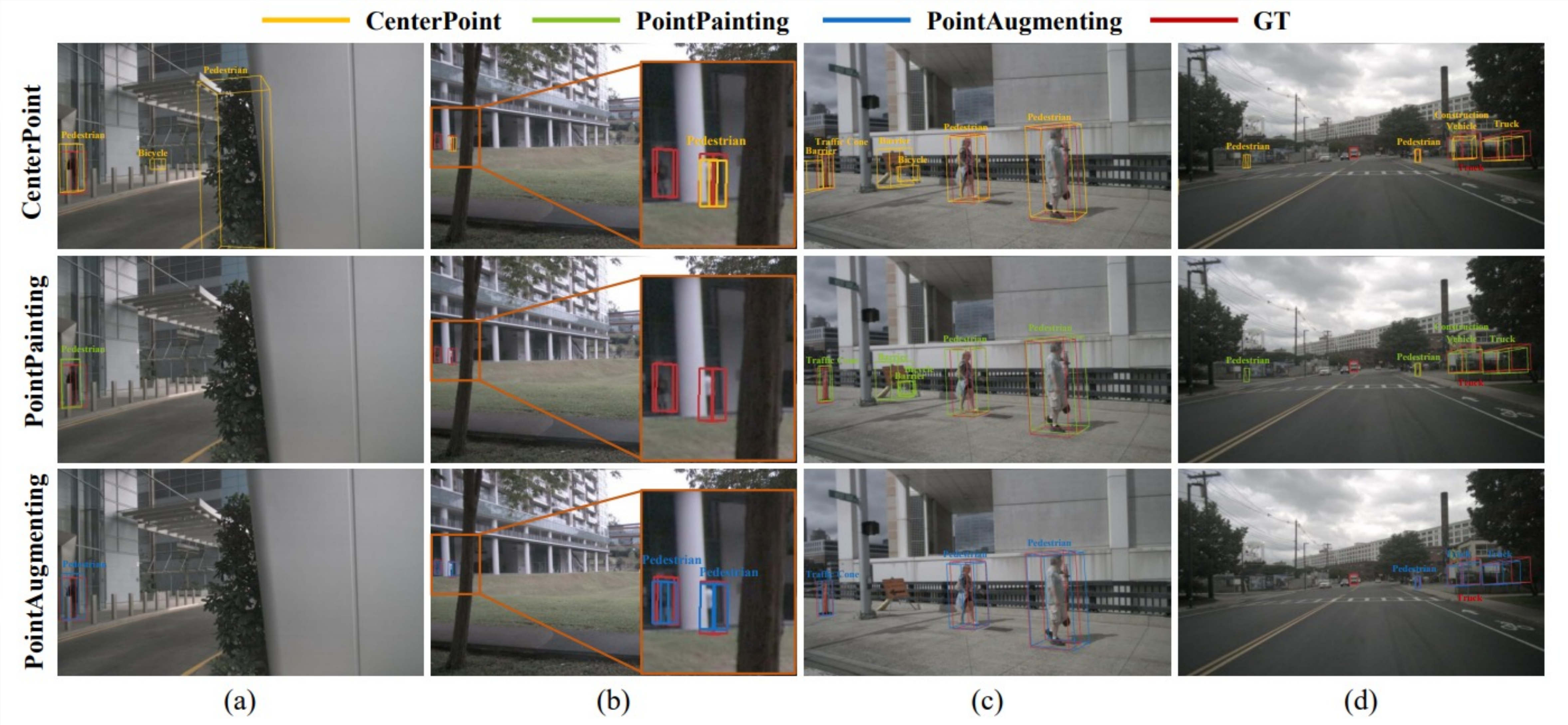
2 Cross-Modal Data Augmentation

	Naive	CM	Fade	Fusion	mAP	NDS
(e)					32.8	42.3
(f)	✓				37.6	49.5
(g)		✓			37.4	49.9
(h)				✓	42.6	50.0
(i)		✓		✓	47.5	55.6
(j)		✓	✓	✓	48.8	56.8

Table 5. Effectiveness of cross-modal data augmentation. Naive: the original GT-Paste applied to CenterPoint. CM: Our cross-modal GT-Paste data augmentation. Fade: the training strategy that discontinues our data augmentation in the last 5 epochs. Fusion: adding camera stream by our late fusion mechanism.

SJTU

Result Comparison

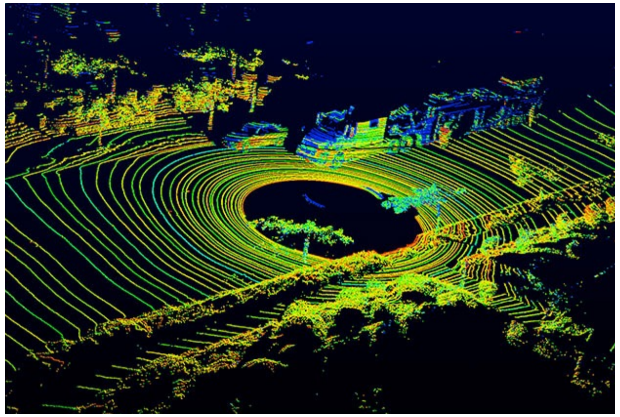


PointAugmenting: Cross-Modal Augmentation for 3D Object Detection

Chunwei Wang, Chao Ma, Ming Zhu, Xiaokang Yang
Shanghai Jiao Tong University

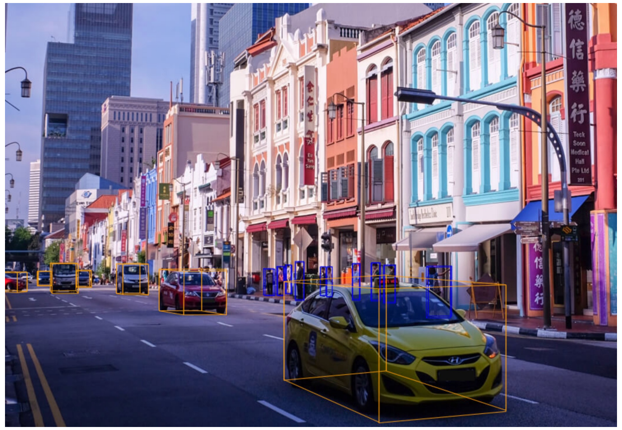
—— CVPR 2021 ——

Background: Cross-Modal Fusion



Modality 1 - LiDAR

- Input： (x, y, z, i).
- Advantages： position accuracy.
- Disadvantages： lack of appearance.



Modality 2 - Camera

- Input： (R, G, B).
- Advantages： rich appearance.
- Disadvantages： lack of depth.

Fusion

Position

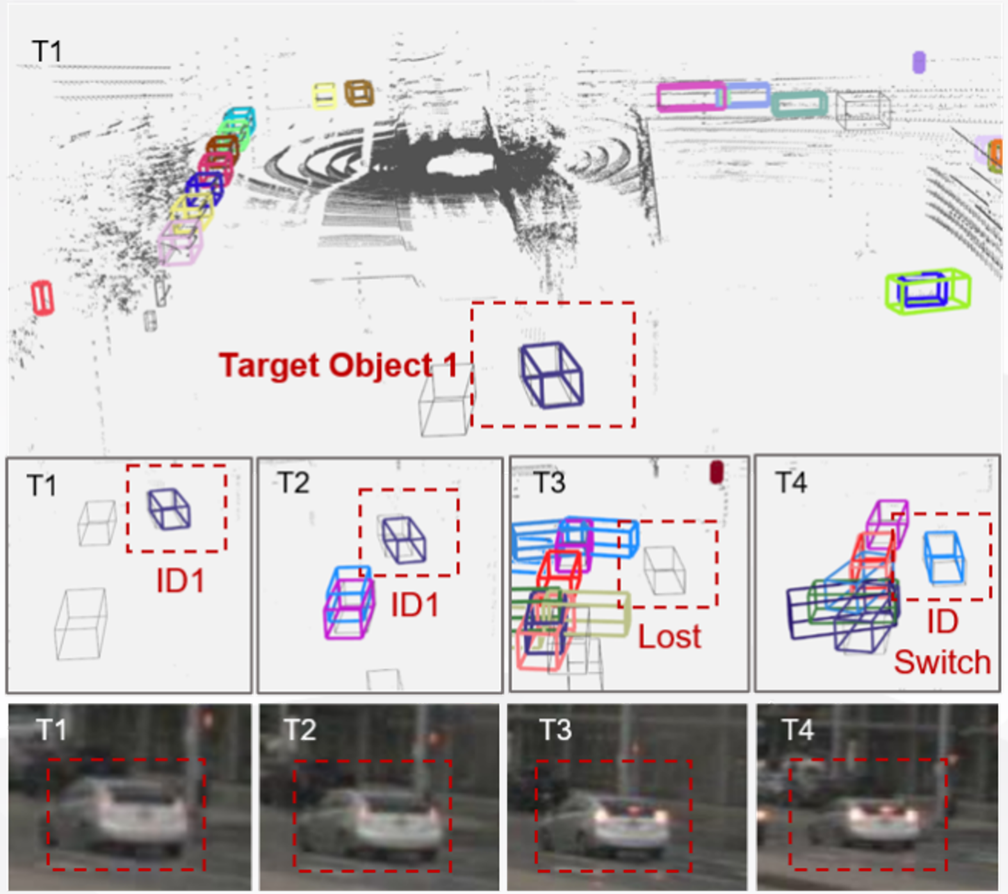
Appearance

Background: Multi-Object Tracking

人工智能研究院
Artificial Intelligence Institute

Cross-Modal Clues for Data Association

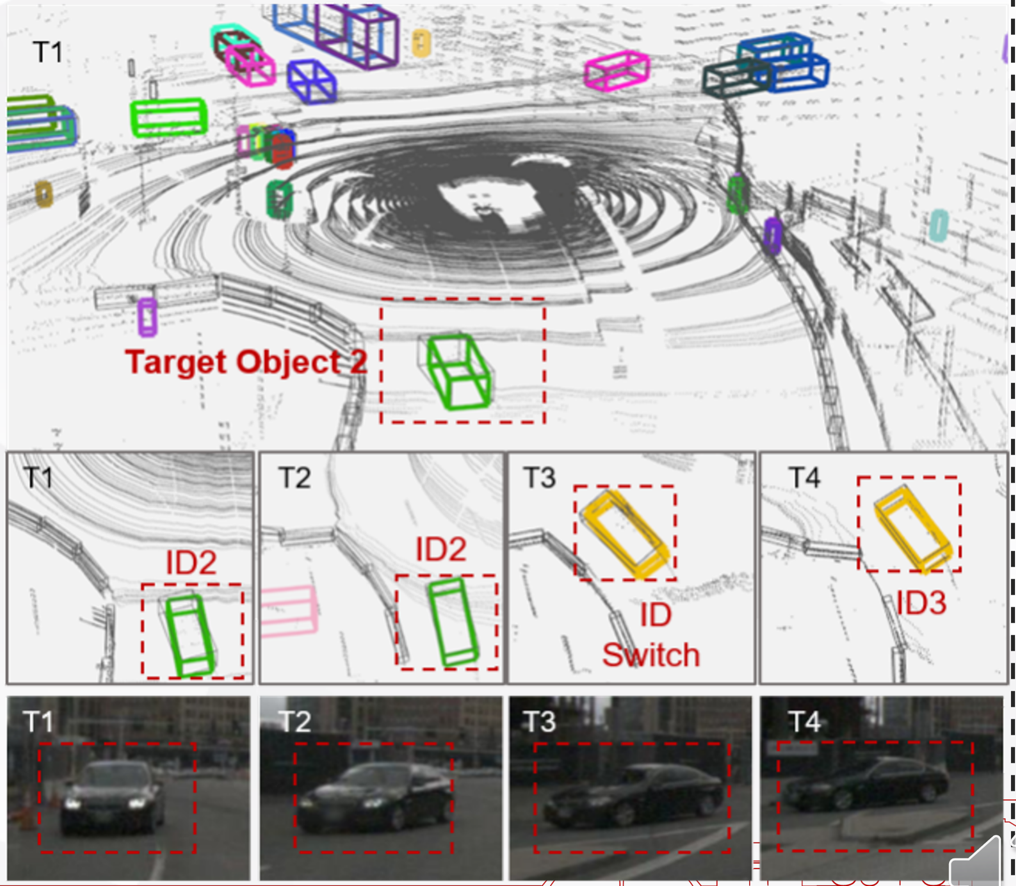
Scene 1 : Noisy Detections



Position clues

Appearance clues

Scene 2 : Large Motions



Related Work

1 Point Cloud-Based

Clue: 3D position information.
Disadvantages: The position association will be agnostic under noisy location perceptions.

- ✓ AB3DMOT 2020 IROS
- ✓ PnPNet 2020 CVPR

2 Image-Based

Clue: Appearance information and 2D position information.
Disadvantages: 2D location is visually distorted and easily occluded.

- ✓ RetinaTrack 2020 CVPR
- ✓ JDE 2020 ECCV
- ✓ CenterTrack 2020 ECCV

3 Fusion-Based

Previous:

Methods: fetch instance-level 3D features or 2D features for each detected instances.
Disadvantages: Time-consuming post-processing.

- ✓ GNN3DMOT 2020 CVPR
- ✓ mmMOT 2019 ICCV
- ✓ JRMOT 2020 IROS

Challenges:

- Robustness – cross-modal clues
- Effectiveness – unified model

Network Architecture

人工智能研究院
Artificial Intelligence Institute

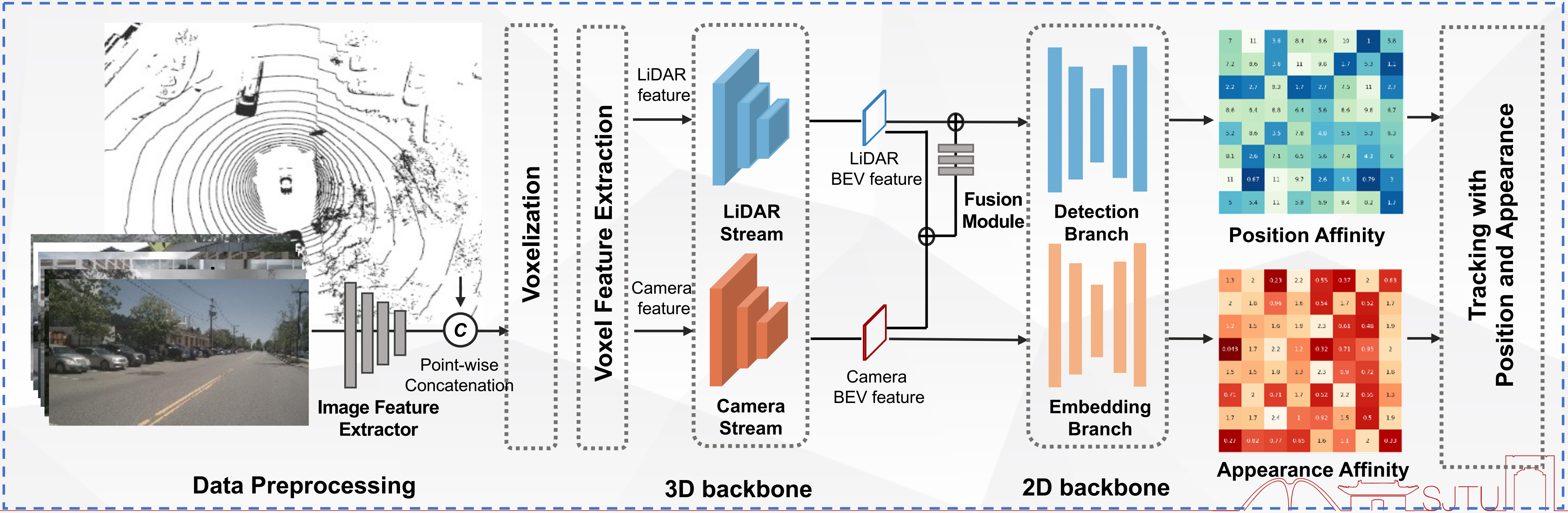
AlphaTrack

Methods: An ***end-to-end model*** that jointly output position and appearance clues, which facilitate the ***cross-modal association mechanism***.

→ ***Effectiveness***

→ ***Robustness***

Overall Architecture



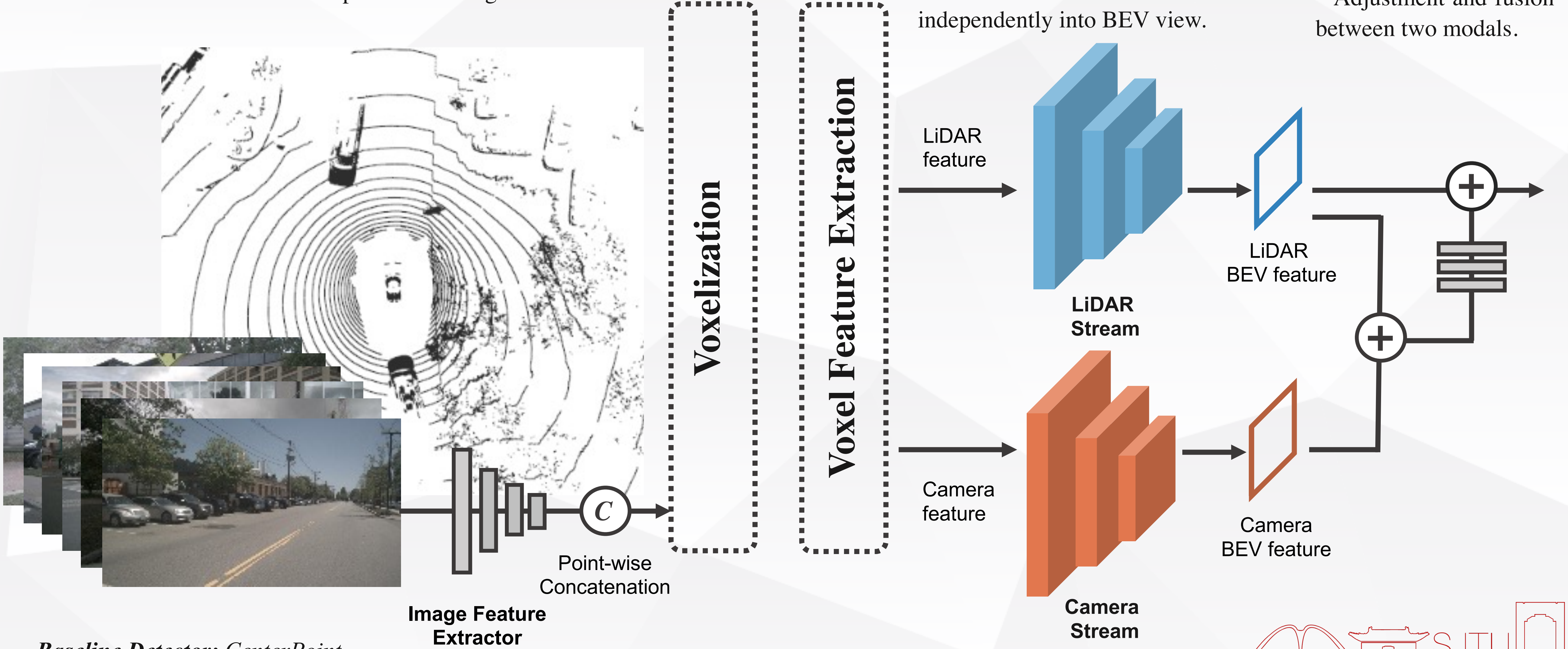
Cross-Modal 3D Detection

人工智能研究院
Artificial Intelligence Institute

1. Data Preprocessing:
- Fetch point-wise image features

2. Parallel 3D backbone
- Process two kinds of features independently into BEV view.

3. Fusion Module
- Adjustment and fusion between two modals.



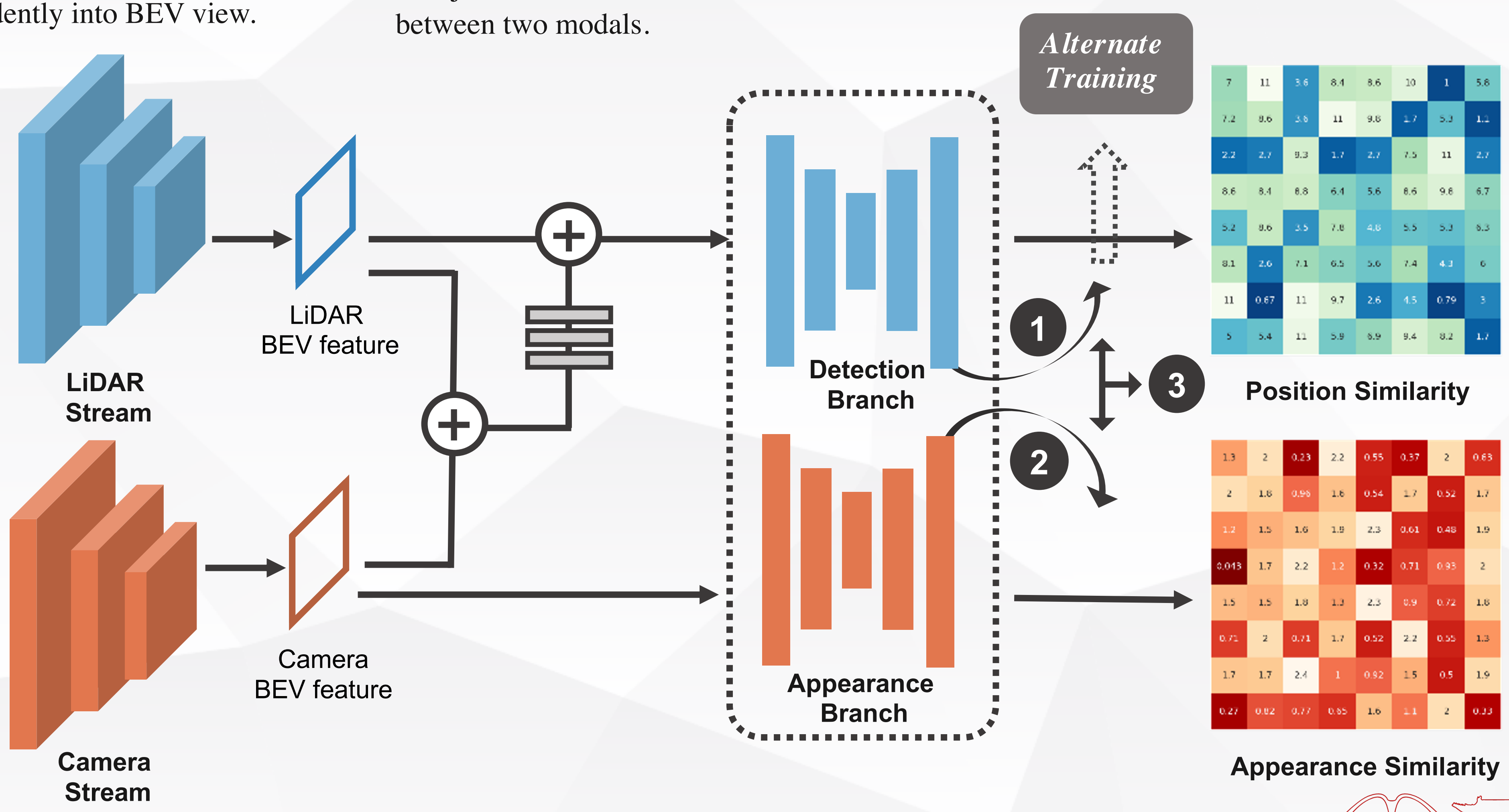
Baseline Detector: *CenterPoint*

Cross-Modal 3D Tracking

2. Parallel 3D backbone
- Process two kinds of features independently into BEV view.

3. Fusion Module
- Adjustment and fusion between two modals.

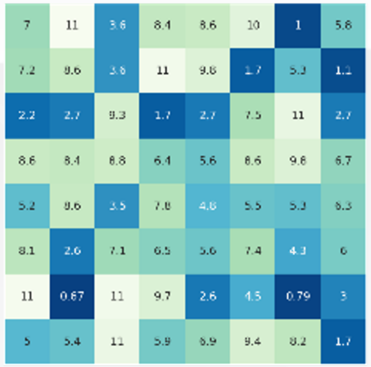
4. Joint output of location and appearance
- Alternative training and jointly output.



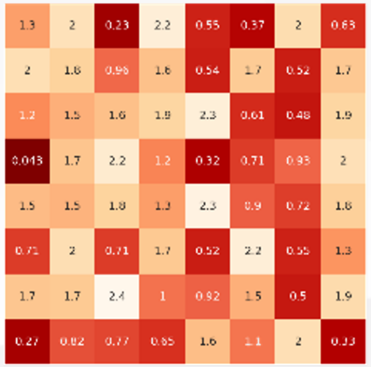
Multi-Modal 3D Tracking

5. Three-stage Tracking algorithm

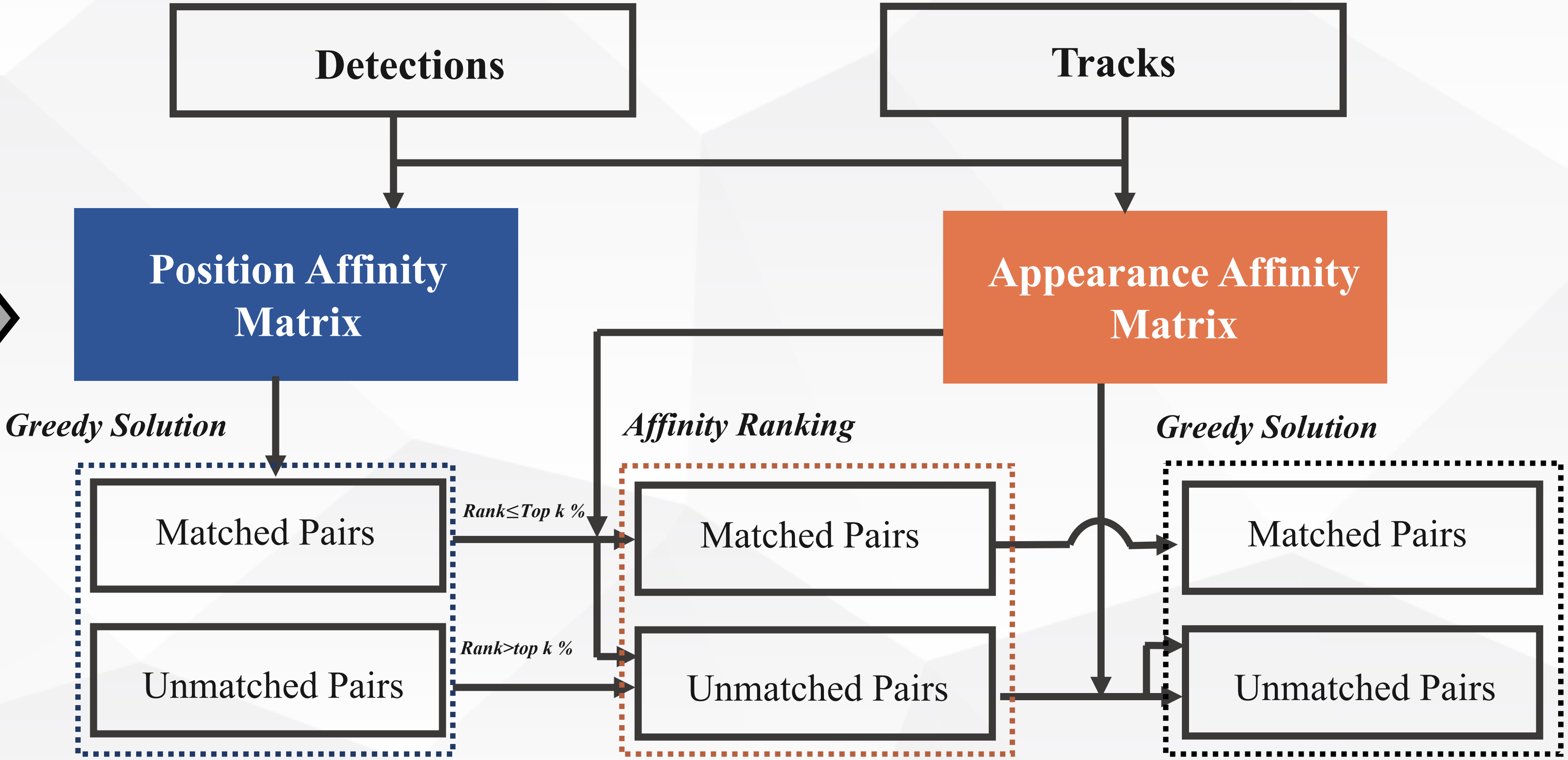
- Implement position and appearance clues explicitly.



Position Similarity



Appearance Similarity



Stage1: Baseline

Stage2: Filtering

Stage3: Re-Matching

Ablation Studies

人工智能研究院
Artificial Intelligence Institute

The effectiveness of network designs

	FI	FM	AE	Bicycle	Bus	Car	Motorcycle	Pedestrian	Trailer	Truck	mAP↑/AMOTA↑
(a)	-	-	-	35.92 / 40.89	67.23 / 79.86	84.73 / 82.93	57.41 / 54.59	82.85 / 73.61	35.30 / 48.85	54.83 / 65.20	59.75 / 63.72
(b)	Seg	EF	-	52.76 / 51.74	69.21 / 79.60	85.50 / 82.81	61.42 / 62.12	85.49 / 74.57	39.90 / 48.56	56.98 / 64.59	64.47 / 66.28
(c)	Feat	EF	-	57.02 / 58.27	71.75 / 82.17	86.84 / 83.85	72.25 / 77.21	86.77 / 74.26	41.93 / 51.57	59.36 / 67.84	67.99 / 70.74
(d)	Feat	LF	-	62.09 / 62.61	**74.56** / 83.03	87.50 / 84.41	**75.78** / **78.18**	86.96 / 73.71	**43.86** / 53.97	61.57 / 70.41	70.33 / 72.33
(e)	Feat	LF	Uniform	56.94 / 59.13	70.02 / 80.07	85.96 / 82.77	70.26 / 74.67	85.86 / 72.66	38.17 / 46.57	59.13 / 68.44	67.99 / 69.19
(f)	Feat	LF	Alter	**64.26** / **65.86**	74.05 / **83.67**	**87.60** / **85.27**	74.94 / **78.18**	**87.15** / **74.83**	43.32 / **54.64**	**61.78** / **70.49**	**70.44** / **73.27**
Gains from a to f				+28.14 / +24.91	+6.82 / +3.81	+2.87 / +2.34	+17.53 / +23.59	+4.30 / +1.22	+8.02 / +5.79	+6.95 / +5.29	+10.69 / +9.55

Cross-Modal Fusion Scheme:

(b) Vs (c): *Image feature representations* provide richer information than *segmentation scores*.

(b) Vs (d): *Late fusion* at BEV level fuse cross-modal features better than *early fusion* at point level.

Joint appearance branch:

(e) Vs (f): *Alternative training* facilitate joint output of both position and appearance embedding.

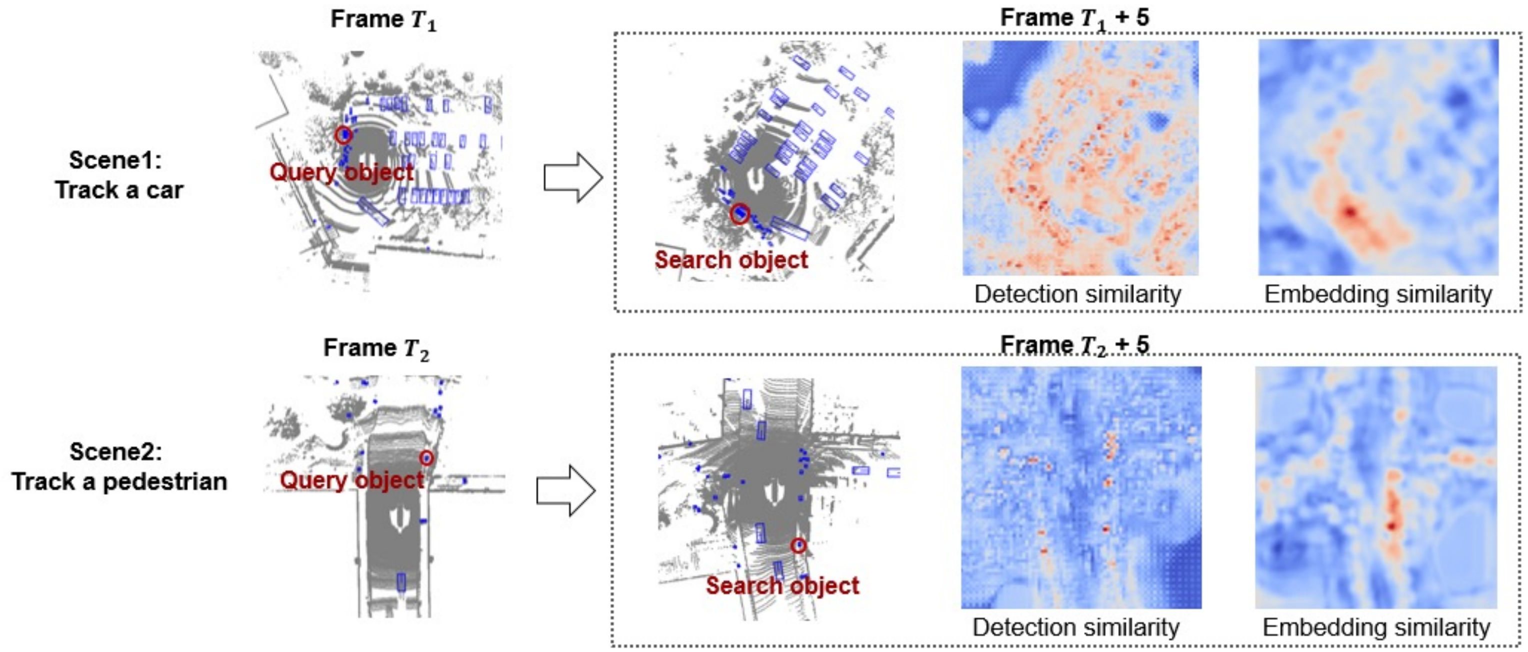
(d) Vs (f): *Appearance embedding* improve tracking association largely in additional to position.

Ablation Studies

Feature retrieval performance

- Appearance embedding feature map is ***instance-aware*** while detection feature map is ***object-agnostic***.
- Our joint appearance embedding show better discriminative power than others.

APP	Det	ATPR↑(%)	AMOTA↑(%)
-	CenterPoint	-	63.72
AlignedReID [28]	CenterPoint	66.92	54.56
PointNet [6]	CenterPoint	41.94	51.82
AlphaTrack (ours)	CenterPoint	**92.68**	**64.93**



	Motion	Sum	Conv	Filter	Re-Match	AMOTA↑(%)	IDS↓
(a1)	Kalman					68.73	1021
(b1)	Kalman	√				69.12	967
(c1)	Kalman		√			67.68	1152
(d1)	Kalman			√		68.53	3432
(e1)	Kalman			√	√	**70.00**	**929**
(a2)	Velocity					72.39	642
(b2)	Velocity	√				72.77	639
(c2)	Velocity		√			70.76	994
(d2)	Velocity			√		73.21	715
(e2)	Velocity			√	√	**73.27**	**575**

The effectiveness of association mechanisms

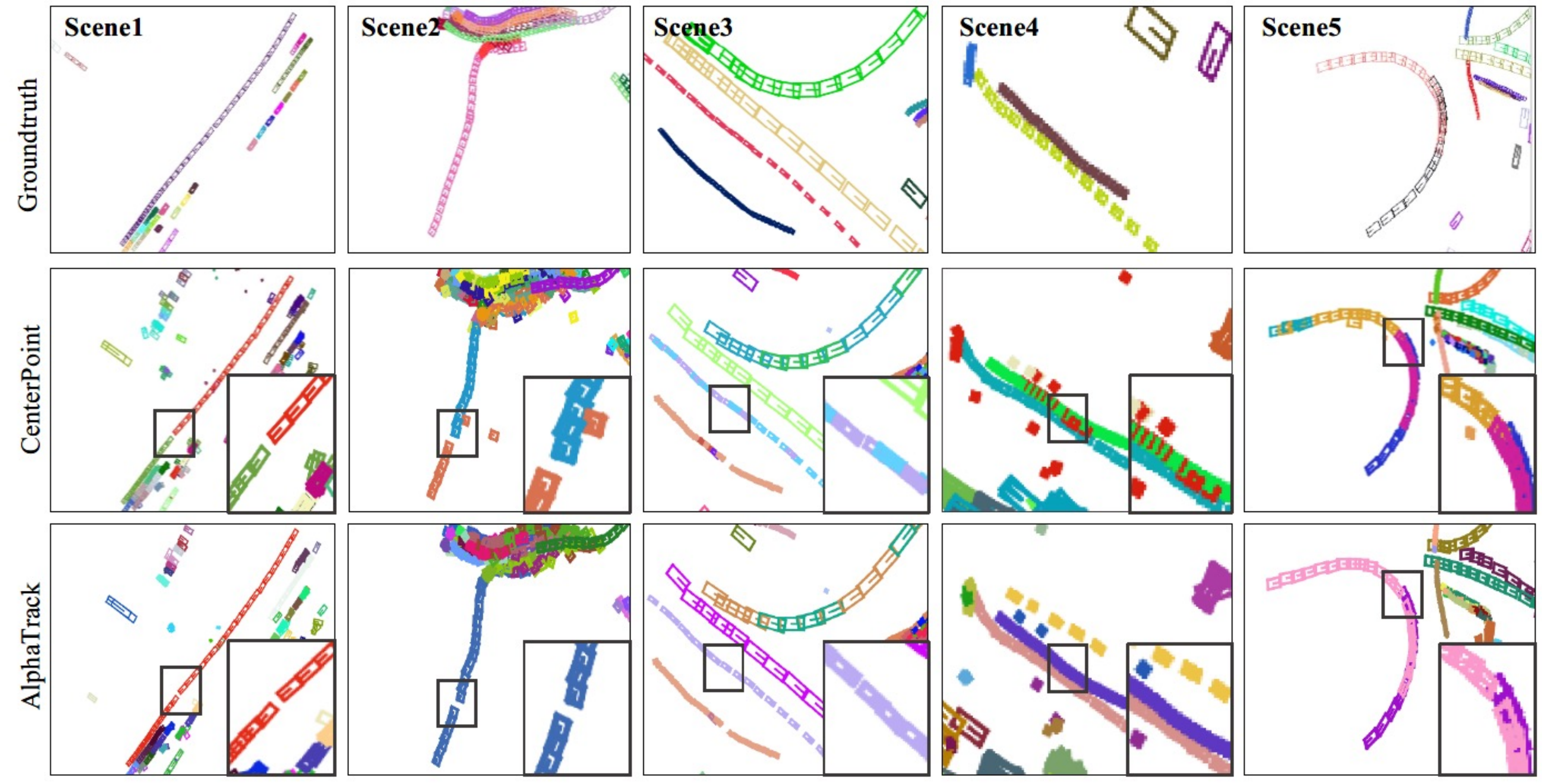
- The *explicit application* of two association clues is superior to simple fusion methods.
- The complementary association clues are effective for *two common motion models*.

Experiments Results

nuScenes test set

Method	Bicycle	Bus	Car	Motor	Ped	Trailer	Truck	AMOTA↑(%)	AMOTP↓(%)	FP↓	FN↓	IDS↓
StanfordIPRL-TRI [21]	25.5	64.1	71.9	48.1	74.5	49.5	51.3	55.0	79.8	17353	33216	950
CenterPoint-single [1]	32.1	71.1	82.9	59.1	76.7	65.1	59.9	63.8	55.5	18612	22928	760
EagerMOT	**58.3**	74.1	81.0	62.5	74.4	63.6	59.7	67.7	**55.0**	17705	24925	1156
Octopus-Traker	41.2	74.5	83.2	69.4	**79.0**	64.5	63.5	67.9	56.2	**16971**	22272	781
AlphaTrack (ours)	47.1	**74.9**	**84.2**	**74.2**	78.3	**70.1**	**64.2**	**70.4**	57.5	18247	**21126**	**718**

Qualitative Result



Detection Comparison

nuScenes：scene - 1066

Tracking Comparison

nuScenes： scene - 1066

Test Evaluation

nuScenes： scene - 0084

Taking-Home Message

- Decorating point cloud with CNN features in the BEV map is helpful for 3D detection
- Cross-modal data augmentation is critical for 3D detection
- Appearance information from images is effective for 3D tracking

Brainstorm

- Is there better correspondence in the BEV map?
- Can mask augmentation work better?
- Can tracking benefit detection?