Know Your Surroundings: Panoramic Multi-Object Tracking by Multimodality Collaboration

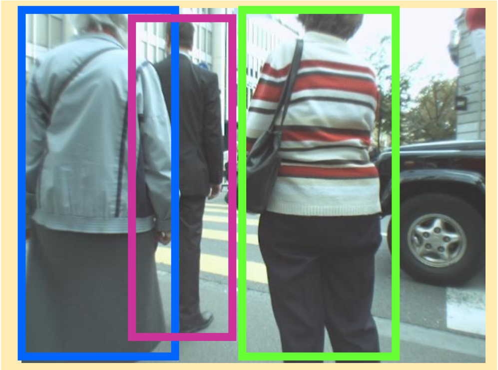
Yuhang He, Wentao Yu, Jie Han, Xing Wei, Xiaopeng Hong, Yihong Gong
Xi'an Jiaotong University (XJTU), P.R.China

Accepted as full paper, winner of 2D detection and tracking tracks

1. Multi-Object Tracking

➢ Goal

- locate the positions of interested targets, maintain their identities across frames and infer a complete trajectory for each target.



(a) Limitation of Field-of-view

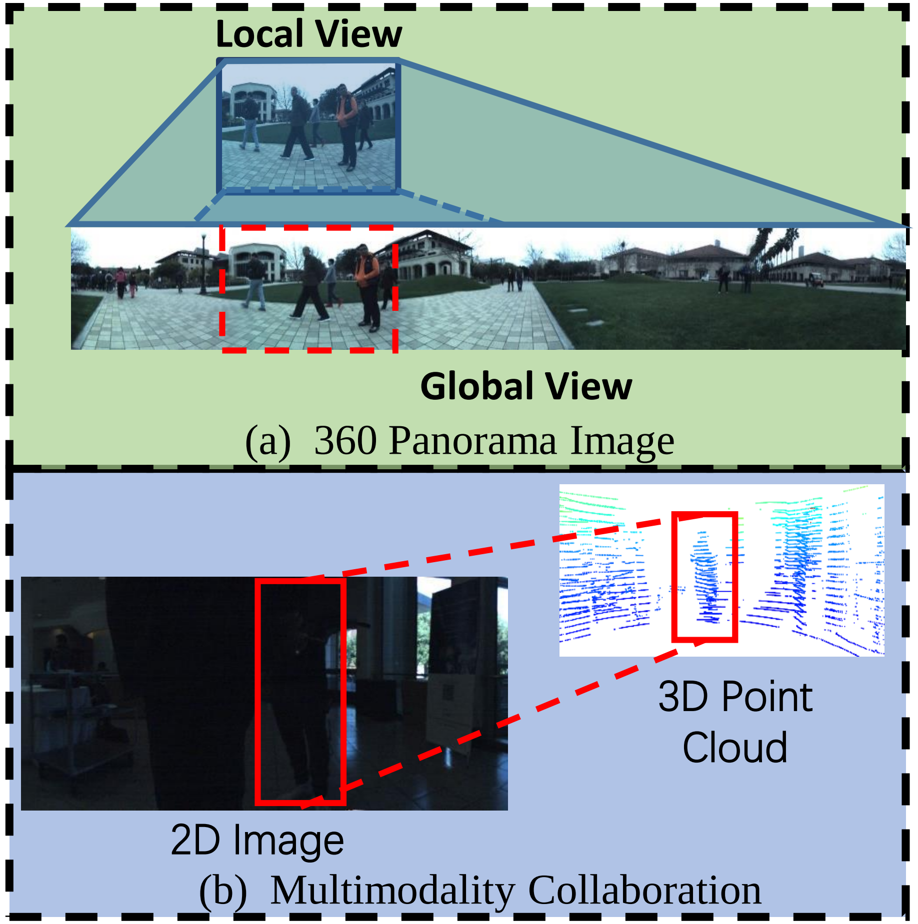
➢ Difficulties

- Limitation of camera field-of-view.
- Tracking failures in complex scenarios such as poor light conditions and background clutters.

(b) Tracking Failures in Complex Scenarios

1. Multi-Object Tracking

- Key Insights
  - A wider vision brings more information.
  - Singular modality is biased while multimodality complements each other.
- Solutions
  - Propose a MultiModality PAnoramic multi-object Tracking framework (MMPAT).
  - Take 2D 360◦ panorama images and 3D LiDAR point clouds as input and generate trajectories for targets by multimodality collaboration.

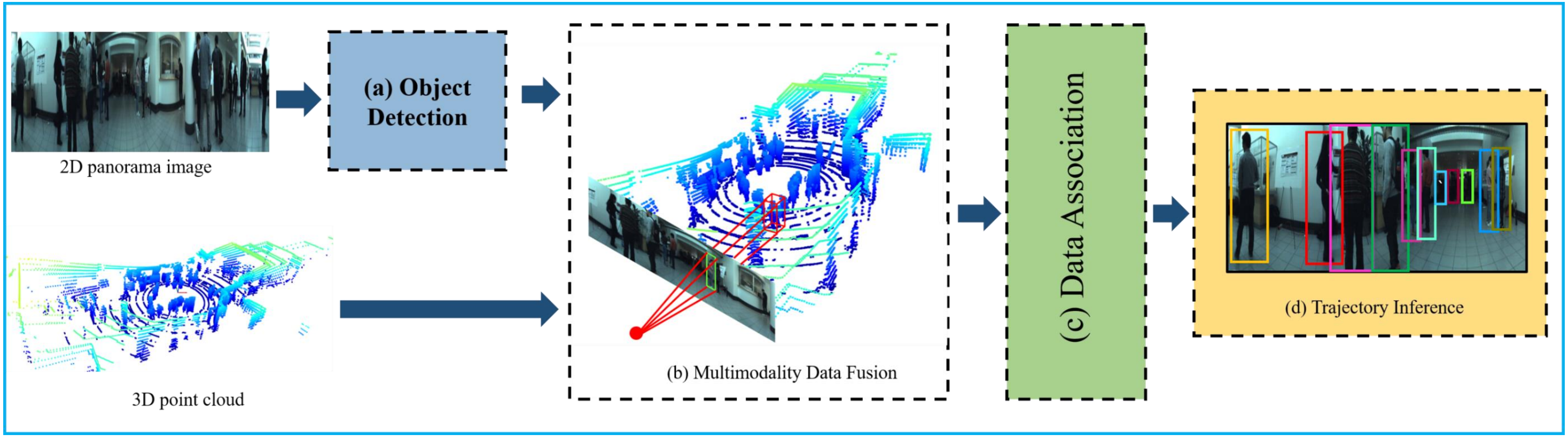


(a) 360 Panorama Image

(b) Multimodality Collaboration

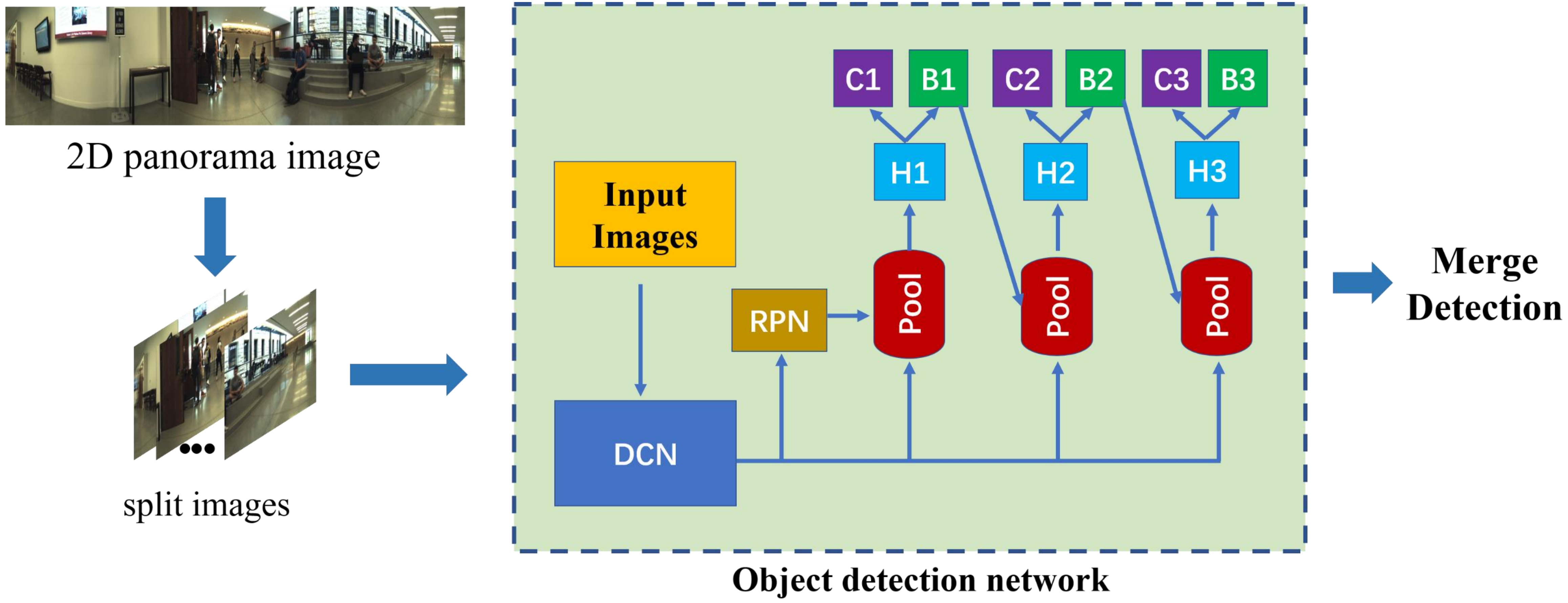
2. Panoramic Multi-Object Tracking by Multimodality Collaboration

➢ 2.1 The Proposed Framework



2. Panoramic Multi-Object Tracking by Multimodality Collaboration

➢ 2.2 Object Detection in Panorama Image



2. Panoramic Multi-Object Tracking by Multimodality Collaboration

➢ 2.2 Object Detection in Panorama Image

- **Panorama image split:**

Split the panorama image I_t into N image slices $\mathcal{J}_t = \{I_t^n\}_{n=1}^N$ along the width dimension with an overlap of 0.2.

- **Cascade object detector:**

Detect objects in each image slice by a deformable convolution network, a region proposal network and a cascade detection header.

- **Detection merge:**

Merge detection responses from all the image slices by non-maximum suppression (NMS): $\mathcal{B}_t = NMS(\mathcal{B}_t(1), \dots, \mathcal{B}_t(N))$.

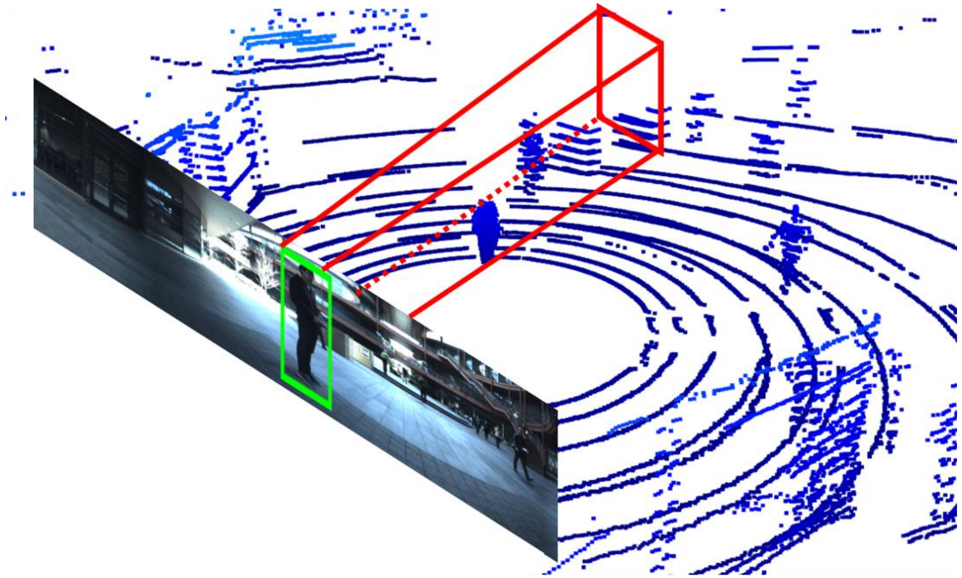
2. Panoramic Multi-Object Tracking by Multimodality Collaboration

2.3 Multimodality Data Fusion

- Perform instance segment in the 2D bounding box to filter out the background clutters.
- Collect 3D points of the target based on 3D-to-2D projection.

$$\mathcal{P} = \left\{ h \middle| \forall h \in \Omega_{ptc}, if \, \rho(h; M) \in \Omega_{box} \right\}$$

- Obtain the 3D location l_t^v of detection B_t^v by averaging the 3D points of detection B_t^v.



2. Panoramic Multi-Object Tracking by Multimodality Collaboration

➢ 2.4 Data Association

- Affinity Measurement:

$$A(u,v) = \psi_{app}(\mathcal{T}_{t-1}^{u}, \mathcal{B}_{t}^{v}) + \psi_{mot}(\mathcal{T}_{t-1}^{u}, \mathcal{B}_{t}^{v}) + \psi_{loc}(\mathcal{T}_{t-1}^{u}, \mathcal{B}_{t}^{v})$$

  - Appearance similarity: $\psi_{app}(\mathcal{T}_{t-1}^{u}, \mathcal{B}_{t}^{v}) = \frac{\sum_{\forall k \in \tau_{t-1}^{u}} [e^{k-t} \cdot \gamma(a_k^u, \phi(\mathcal{B}_t^v))]}{\sum_{\forall k \in \tau_{t-1}^{u}} e^{k-t}}$
  - Motion affinity: $\psi_{mot}(\mathcal{T}_{t-1}^{u}, \mathcal{B}_{t}^{v}) = area(\mathcal{O}_t^u \cap \mathcal{B}_t^v) / area(\mathcal{O}_t^u \cup \mathcal{B}_t^v)$
  - Location proximity: $\psi_{loc}(\mathcal{T}_{t-1}^{u}, \mathcal{B}_{t}^{v}) = \sum_{k \in \tau_{t-1}^{u}} \frac{\sigma_t(k,t) \cdot \sigma_l(\mathcal{T}_{t-1}^{u}(k)_{loc}, l_t^v)}{|\tau_k^u|}$

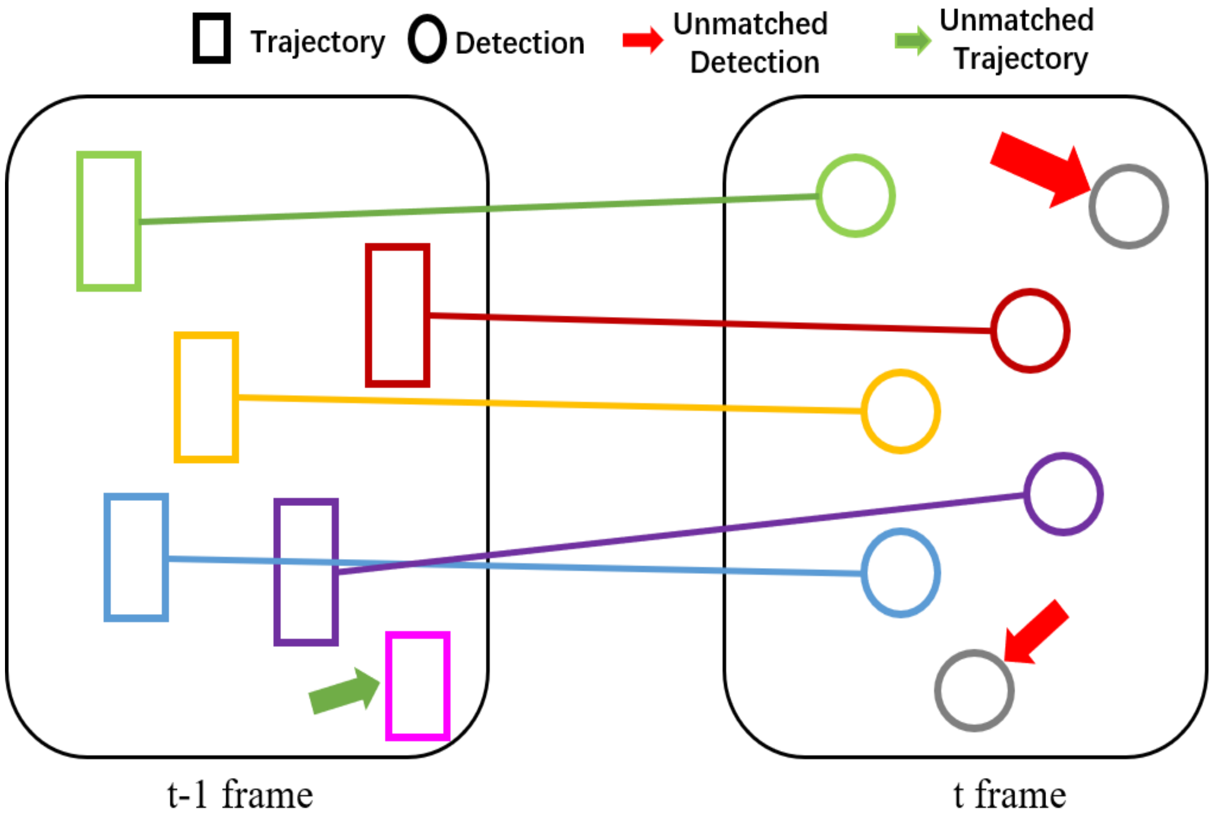
- Bipartity Graph Matching:

$$X^* = \underset{X}{\operatorname{argmax}} \|A \odot X\|_2, \qquad s.t.\ \forall u, \sum X(u,:) \leq 1, \forall v, \sum X(:,v) \leq 1,$$

2. Panoramic Multi-Object Tracking by Multimodality Collaboration

➢ 2.5 Trajectory Inference

- Detection $\mathcal{B}_t^v$ does not match with any trajectories.
- Trajectory $\mathcal{T}_{t-1}^u$ is matched with detection $\mathcal{B}_t^v$.
- Trajectory $\mathcal{T}_{t-1}^u$ does not match with any detections.



2. Panoramic Multi-Object Tracking by Multimodality Collaboration

- 2.6 Experiment:

Table 1. Detection results on the JRDB Dataset

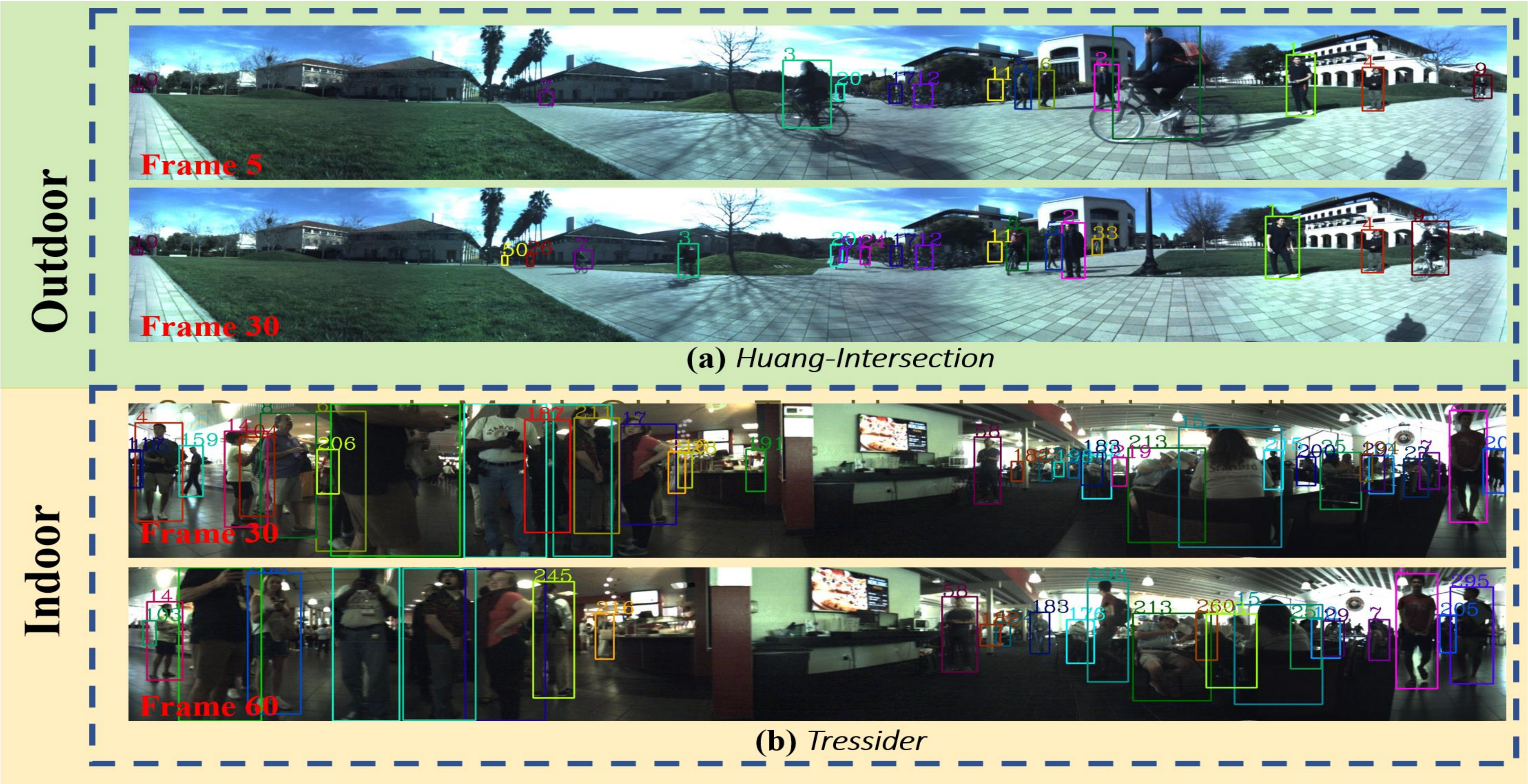
Method	AP ↑	Runtime ↓
YOLOV3 [61]	41.73	0.051
DETR [9]	48.51	0.350
RetinaNet [45]	50.38	0.056
Faster R-CNN [62]	52.17	**0.038**
Ours	**67.88**	0.070

Table 2. Tracking Results On the JRDB Dataset

Method	MOTA ↑	IDS ↓	FP ↓	FN ↓
Tracktor [1]	19.7	7026	79573	681672
DeepSORT [76]	23.2	**5296**	78947	650478
JRMOT [69]	22.5	7719	**65550**	667783
Ours	**31.7**	5742	67171	**580565**

2. Panoramic Multi-Object Tracking by Multimodality Collaboration

➢ 2.6 Experiment:



(a) *Huang-Intersection*

(b) *Tressider*

2. Error-Aware Density Isomorphism Reconstruction for Unsupervised Cross-Domain Crowd Counting

- 2.6 Experiment:

Table 3. Ablation Study on Object Detection

Method	AP ↑
Baseline	52.8
Baseline+DCN	53.1
Baseline+DCN+split	64.6
Baseline+DCN+split+mixup	68.2
Baseline+DCN+split+mixup+multiscale	69.7
Baseline+DCN+split+mixup+multiscale+softnms	70.7

Thank you!