Omnidirectional Multi-Object Tracking

Kai Luo[1,*] Hao Shi[2,*] Sheng Wu[1] Fei Teng[1] Mengfei Duan[1] Chang Huang[1]
Yuhang Wang[1] Kaiwei Wang[2] Kailun Yang[1,†]
[1]Hunan University [2]Zhejiang University

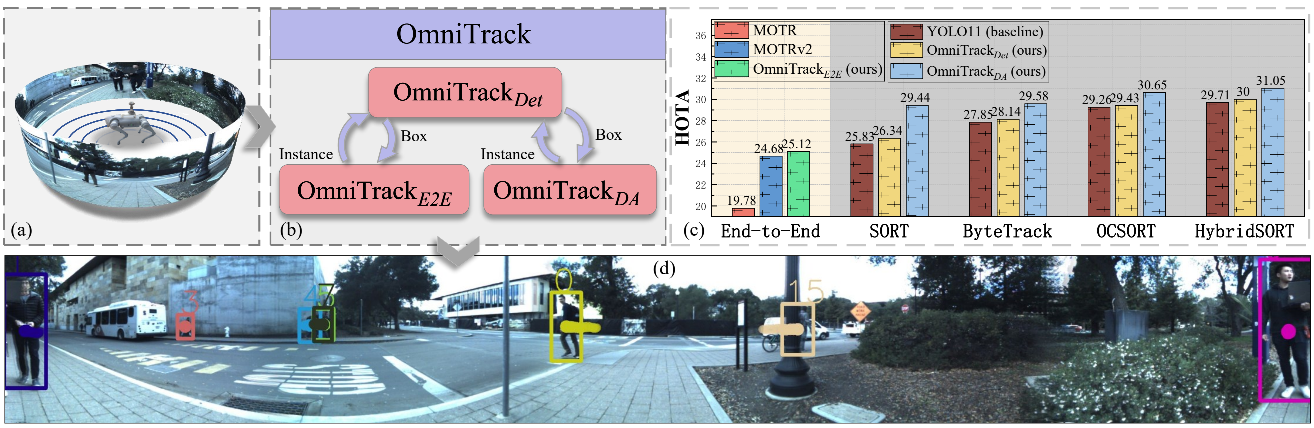

Figure 1. Comparison of OmniTrack's overall structure and performance. (a) shows the input panoramic image. (b) illustrates the proposed OmniTrack method. (c) presents a performance comparison with other multi-object tracking algorithms. (d) visualizes tracking results.

Abstract

Panoramic imagery, with its 360° field of view, offers comprehensive information to support Multi-Object Tracking (MOT) in capturing spatial and temporal relationships of surrounding objects. However, most MOT algorithms are tailored for pinhole images with limited views, impairing their effectiveness in panoramic settings. Additionally, panoramic image distortions, such as resolution loss, geometric deformation, and uneven lighting, hinder direct adaptation of existing MOT methods, leading to significant performance degradation. To address these challenges, we propose OmniTrack, an omnidirectional MOT framework that incorporates Tracklet Management to introduce temporal cues, FlexiTrack Instances for object localization and association, and the CircularStatE Module to alleviate image and geometric distortions. This integration enables tracking in panoramic field-of-view scenarios, even under rapid sensor motion. To mitigate the lack of panoramic MOT datasets, we introduce the QuadTrack dataset—a comprehensive panoramic dataset collected by a quadruped robot, featuring diverse challenges such as panoramic fields of view, intense motion, and complex environments. Extensive experiments on the public JRDB dataset and the newly introduced QuadTrack benchmark demonstrate the state-of-the-art performance of the proposed framework. OmniTrack achieves a HOTA score of 26.92% on JRDB, representing an improvement of 3.43%, and further achieves 23.45% on QuadTrack, surpassing the baseline by 6.81%. The established dataset and source code are available at https://github.com/xifen523/OmniTrack.

1. Introduction

Panoramic cameras, with a 360° Field of View (FoV), capture comprehensive surrounding information, making them essential for applications like autonomous driving [10, 67], robotic navigation [60, 64], and human-computer interaction [28, 69]. For small-scale mobile robots, such as quadrupedal robots, panoramic cameras are especially advantageous, allowing complete environmental awareness within a single compact setup, as illustrated in Fig. 1(a).

Despite progress in Multi-Object Tracking (MOT), panoramic MOT remains underexplored. Existing MOT algorithms [14, 48], developed for pinhole cameras, struggle in panoramic settings due to inherent challenges like resolution loss, geometric distortion, and uneven color and brightness distribution when unfolded (Fig. 1 (d)). These challenges often lead to performance degradation when applying pinhole-based algorithms to panoramic images, limiting their effectiveness for panoramic scene perception.

To address these challenges, developing an MOT algo-

*Equal contribution
†Correspondence: kailun.yang@hnu.edu.cn

rithm capable of comprehensive perception in panoramic images with panoramic FoV is a pressing problem. To this end, this paper, for the first time, proposes an omnidirectional multi-object tracking framework, **OmniTrack**, specifically designed for such tasks in 360° panoramic imagery. OmniTrack unifies two mainstream MOT paradigms—Tracking-By-Detection (TBD) and End-To-End (E2E) tracking—and introduces a feedback mechanism that effectively reduces uncertainty in panoramic FoV with rapid sensor motion, enabling fast and accurate target localization and association.

This framework consists of three core components: a *CircularStatE Module*, *FlexiTrack Instance*, and *Tracklets Management*. The CircularStatE Module is designed to mitigate wide-angle distortion and enhance consistency in lighting and color. The FlexiTrack Instance exploits the temporal continuity of objects, guiding the perception module to focus on key areas within the panoramic FoV and aiding in localization and association. This approach helps mitigate the difficulty of object localization in panoramic FoV. The Tracklets Management module collects and manages trajectory data, providing prior knowledge to the FlexiTrack Instance. Through these components, OmniTrack unifies the two MOT paradigms: disabling data association within Tracklets Management results in an End-to-End tracker, OmniTrack$_{E2E}$, while enabling association yields a TBD-style tracker, OmniTrack$_{DA}$. By employing the same data association strategy, as shown in Fig. 1 (c), the framework of OmniTrack$_{DA}$ achieves significantly stronger performance. Disabling both the FlexiTrack Instance and Tracklets Management reduces the system to a panoramic object detector, OmniTrack$_{Det}$, as shown in Fig. 1 (b).

Moreover, to support panoramic MOT research, we developed **QuadTrack**, a dataset collected with a $360^{\circ} \times 70^{\circ}$ panoramic camera mounted on a quadrupedal robot. This mobile platform's biomimetic gait introduces realistic, complex motion characteristics, challenging existing MOT algorithms. Collected across five campuses in two cities, QuadTrack includes $19,200$ images, encompassing a wide variety of dynamic, real-world scenarios. In contrast to typical MOT datasets [5, 8, 15, 17, 53, 74] that use static or linearly moving platforms, QuadTrack provides a new benchmark for evaluating MOT performance in panoramic-FoV scenarios with rapid and non-linear sensor motion.

At a glance, our work makes the following contributions:

- To address the gap in omnidirectional multi-object tracking, we propose **OmniTrack**, a novel framework that unifies both E2E and TBD tracking paradigms. This approach reduces uncertainty and enhances perceptual and association performance in panoramic-FoV scenarios.
- We present **QuadTrack**, a new panoramic MOT dataset with complex motion dynamics, providing a challenging benchmark for panoramic-FoV multi-object tracking.
- Extensive experiments on JRDB and QuadTrack datasets show OmniTrack's superior performance, achieving a 26.92% HOTA on JRDB and 23.45% on QuadTrack test splits, advancing the state-of-the-art in panoramic MOT.

2. Related Work

Panoramic scene understanding. Panoramic perception enables a holistic understanding of a 360° scene in a single shot [3, 13, 19, 22, 25, 37, 38]. Main areas include panoramic scene segmentation [10, 33, 34, 62, 71, 81, 82], panoramic estimation [2, 4, 12, 58, 65], panoramic layout estimation [46, 59, 75], panoramic generation [42, 66, 83], and panoramic flow estimation [44, 60], *etc*. [23, 28, 39, 54]. Researchers typically unfold panoramas into equirectangular projections or polyhedral projections to adapt algorithms designed for limited-FoV data [35, 44, 65]. They also apply techniques such as deformable convolutions to handle severe distortions in high-latitude regions [60, 77].

Recently, researchers have recognized the advantages of omnidirectional images for tracking, particularly their ability to maintain continuous observation of targets without the out-of-view issues present in limited field-of-view setups. Jiang *et al*. [36] propose a 500FPS omnidirectional tracking system using a three-axis active vision mechanism to capture fast-moving objects in complex environments. The 360VOT benchmark [31] is introduced for omnidirectional object tracking, focusing on spherical distortions and object localization challenges. Huang *et al*. [32] present 360Loc for omnidirectional localization that tackles cross-device challenges by generating lower-FoV query frames from 360° data. Another work by Xu *et al*. [70] introduces an extended bounding FoV (eBFoV) representation to alleviate spherical distortions in panoramic videos. Unlike previous methods, this work first explores extremely challenging panoramic-FoV and intense-motion panoramic tracking for mobile robots, *e.g*., aiming to enhance the robot's spatiotemporal understanding of objects in its surroundings.

Multi-object tracking. Object tracking primarily follows two paradigms: Tracking-By-Detection (TBD) [14, 21, 30, 43, 48, 55, 56, 80] and End-To-End (E2E) [18, 24, 45, 76]. Among these, TBD is currently one of the most prevalent, with frameworks following the design principles of SORT [68]. First, the detection network [11, 26] is used to locate bounding boxes for objects, then the target's current position is predicted based on its historical trajectory, and the predicted results are associated with detection results [41]. Many subsequent works have refined this approach: DeepSORT [44] introduced a ReID model to incorporate appearance information for association, and ByteTrack [78] designed a confidence-based, stage-wise association strategy. Other methods [1, 20, 73] introduced motion compensation modules to mitigate camera motion, and OC-SORT [9] optimized the motion estimation mod-

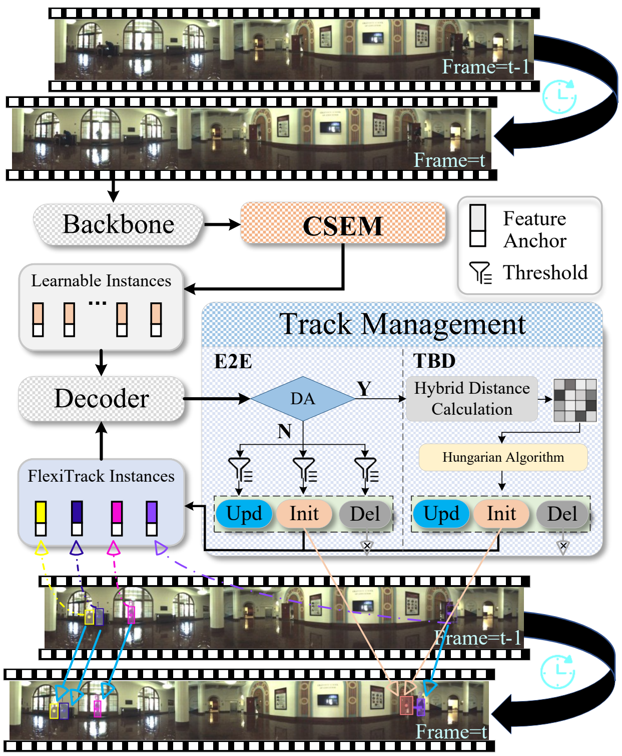

Figure 2. The proposed OmniTrack pipeline. **CSEM** refers to the CircularStatE Module 3.4 , **DA** stands for data association, **E2E** denotes the End-to-End tracking paradigm, **TBD** refers to the Track-By-Detection tracking paradigm, **Upd** refers to updating tracks, **Init** to initializing tracks, and **Del** to deleting tracks.

ule. Additionally, E2E methods have continued to evolve. TrackFormer [51] and MOTR [76] proposed transformer-based, End-to-End tracking approaches. Recent improvements [48, 79] have enhanced detector performance and improved data association accuracy in occlusion scenarios. Unlike existing methods that focus on narrow-FoV pinhole camera data with linear sensor motion, we address the challenges of MOT in panoramic-FoV scenarios, tackling issues such as geometric distortion and complex motion.

3. OmniTrack: Proposed Framework

In this section, we introduce OmniTrack, a panoramic multi-object tracking framework that addresses the unique challenges in panoramic-FoV images, including extensive search spaces, geometric distortion, resolution loss, and lighting inconsistencies. OmniTrack is designed with a feedback mechanism to iteratively refine object detection, integrating trajectory information back into the detector to enhance tracking stability across panoramic-FoV scenes (Sec. 3.1). Specifically, we propose the OmniTrack framework, which consists of three key components:

- **Tracklets Management** (Sec. 3.2): Manages object trajectory lifecycles and provides temporal priors to the perception module.
- **FlexiTrack Instance** (Sec. 3.3): Rapidly locates and associates objects across the panoramic view by leveraging temporal context.
- **CircularStatE Module** (Sec. 3.4): Mitigates geometric distortion and improves consistency across the panoramic FoV, enhancing feature reliability.

Algorithm 1: OmniTrack Inference Process

Input: A Panoramic video/image sequence $\mathbb{V}$
Output: Tracks $\mathcal{T}$ of the video/image sequence

Initialization: $\mathcal{T} \leftarrow \emptyset$;
Define the Initialize threshold $\tau_{\mathcal{I}}$;
Define the Update threshold $\tau_{\mathcal{U}}$;
for *frame* f_k *in* $\mathbb{V}$ **do**
  /* As shown in Fig. 2 */
  $\{\mathcal{S}_3, \mathcal{S}_4, \mathcal{S}_5\} \leftarrow \texttt{Backbone}(f_k)$;
  $\mathcal{I}_L \leftarrow \texttt{CSEM}(\{\mathcal{S}_3, \mathcal{S}_4, \mathcal{S}_5\})$;
  $\mathcal{I}_F \leftarrow \mathcal{T}_{f_{k-1}}$;
  $\mathcal{D}_k^F, \mathcal{D}_k^L \leftarrow \texttt{Decoder}(\mathcal{I}_F, \mathcal{I}_L)$;
  if *DA* **then**
    /* Data Association */
    $\mathcal{C} \leftarrow \texttt{Distance Calculation}(\mathcal{D}_k^F + \mathcal{D}_k^L, \mathcal{T}_{f_{k-1}})$;
    $\{\texttt{Update, Initialize, Delate}\} \leftarrow \texttt{Hungarian Algorithm}(\mathcal{C})$;
    $\mathcal{T}_{f_k} \leftarrow \{\texttt{Update, Initialize, Delete}\}$
  else
    /* End-to-End */
    for d *in* $\{\mathcal{D}_k^F \cup \mathcal{D}_k^L\}$ **do**
      if $d \in \mathcal{D}_k^F$ & $d.score > \tau_{\mathcal{U}}$ **then**
        $\texttt{Update} \leftarrow d$;
      if $d \in \mathcal{D}_k^L$ & $d.score > \tau_{\mathcal{I}}$ **then**
        $\texttt{Initialize} \leftarrow d$;
      else
        $\texttt{Delete} \leftarrow d$;
    $\mathcal{T}_{f_k} \leftarrow \{\texttt{Update, Initialize, Delete}\}$
Return: $\mathcal{T}$

In green is the key of our method.

3.1. Feedback Mechanism

The OmniTrack framework, illustrated in Fig. 2, incorporates a feedback mechanism that iteratively refines detections by integrating trajectory information back into the detector. This mechanism operates on the principle of reducing information entropy, thereby enhancing stability in Panoramic-FoV and improving MOT performance.

In traditional MOT [1, 9, 20, 78], detection and association are decoupled, leading to higher entropy as each frame's detection $H(x_t)$ is calculated independently:

$$H(x_t) = -\sum_{i=1}^{n} P(x_t^i) \log P(x_t^i), \tag{1}$$

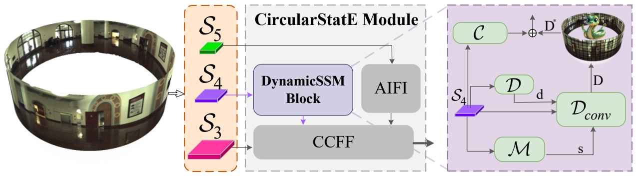

Figure 3. The proposed **CircularStatE Module** fuses multi-scale features to generate learnable instances. The **DynamicSSM Block** mitigates distortions in panoramic-FoV images, enhancing feature stability across uneven lighting and color distributions.

where x_t^i denotes the position of the i-th target in frame t, with probability distribution $P(x_t^i)$. The global association entropy $H(\{y_t\})$ depends on the joint probability distribution of target positions across all frames:

$$H(\{y_t\}) = -\sum_{i=1}^{n} P(\{x_1^i, x_2^i, \ldots, x_T^i\}) \times \log P(\{x_1^i, x_2^i, \ldots, x_T^i\}). \quad (2)$$

The cumulative entropy across all frames, accounting for independent matching, is formulated as:

$$H_{\text{independent}} = \sum_{t=1}^{T} H(x_t) + H(\{y_t\}). \quad (3)$$

In contrast, OmniTrack's feedback mechanism allows detections from frame $t-1$ to inform those in frame t, reducing per-frame uncertainty. Specifically, the conditional entropy of frame t, given prior feedback y_{t-1}, is:

$$H(x_t|y_{t-1}) = -\sum_{i=1}^{n} P(x_t^i|y_{t-1}^i) \log P(x_t^i|y_{t-1}^i). \quad (4)$$

The total entropy with feedback becomes:

$$H_{\text{feedback}} = \sum_{t=1}^{T} H(x_t|y_{t-1}), \quad (5)$$

where $H_{\text{feedback}} < H_{\text{independent}}$, indicating a reduction in uncertainty over time. This feedback-driven approach thus enhances tracking stability in panoramic-FoV scenarios.

3.2. Tracklets Management

To reduce uncertainty in target localization and association while incorporating temporal information, OmniTrack incorporates a Tracklets Management module. During training, this module caches temporal data for instances with confidence scores exceeding a threshold τ, providing historical context to improve detection consistency in subsequent frames. During inference, Tracklets Management oversees trajectory lifecycle management by updating, deleting, or initializing instances based on their confidence scores. In scenarios without data association, trajectories are managed directly, forming OmniTrack$_{E2E}$ (Alg. 1, Lines 14-21). When data association is enabled, Tracklets Management utilizes TBD-based methods [9, 72] to enhance tracking, referred to as OmniTrack$_{DA}$ (Alg. 1, Lines 10-12)

3.3. FlexiTrack Instance

As described in Eq. (2), the global association entropy is significantly high under panoramic-FoV conditions, making the association task challenging. Benefiting from the Feedback Mechanism (Sec. 3.1), which integrates trajectory information into the detector to reduce information entropy. This approach eliminates the need for global search across the entire field of view, making it especially effective for panoramic-scale perception tasks. Based on this insight, we introduce *FlexiTrack Instance*.

Each FlexiTrack Instance (see Fig. 2) shares the Decoder network structure with Learnable Instances, consisting of a feature vector $\mathcal{X} \in \mathbb{R}^{128}$ and an anchor $\mathcal{Y} \in \mathbb{R}^{128}$, as shown in Fig. 2. By sharing the decoder, FlexiTrack Instances can seamlessly adapt to various MOT paradigms, enhancing flexibility and allowing integration across different approaches without additional modifications. To enhance robustness, noise is added to both feature vectors and anchors during training, minimizing dependency on historical data and improving generalization:

$$\mathcal{X}' = \mathcal{X} + \mathcal{N}_X, \quad \mathcal{Y}' = \mathcal{Y} + \mathcal{N}_Y, \quad (6)$$

where $\mathcal{N}_X$ and $\mathcal{N}_Y$ represent the noise components added to the feature vector and anchor, respectively. To initialize all FlexiTrack Instances, let $\mathcal{I_F}$ denote the set of initial *instances*, and N the total number of trajectories. Each instance ${\mathcal{I_F}}^i$ is composed of a feature vector $\mathcal{X}_i$ and an anchor $\mathcal{Y}_i$, as:

$$\mathcal{I_F} = \left\{ {\mathcal{I_F}}^i \mid {\mathcal{I_F}}^i = (\mathcal{X}_i', \mathcal{Y}_i'), i \in \{1, 2, \ldots, N\} \right\}. \quad (7)$$

$\mathcal{X}_i' \in \mathbb{R}^{d_\mathcal{X}}$ and $\mathcal{Y}_i' \in \mathbb{R}^{d_\mathcal{Y}}$ are the feature vector and anchor of the i-th trajectory, with $d_\mathcal{X} = d_\mathcal{Y} = 128$ representing their respective dimensions. This enables $\mathcal{I_F}$ to inherit trajectory information, guiding the perception module to quickly locate the object and establish temporal associations.

3.4. CircularStatE Module

The panoramic image provides an exceptionally panoramic FoV, capable of capturing 360° scenes. However, this inevitably introduces issues such as geometric distortions and inconsistencies in color and brightness in real-world high-dynamic-range scenes. To address these challenges, this paper proposes the *CircularStatE Module*, which alleviates distortions and improves the consistency of image features, thereby enhancing the performance of perception models.

The *DynamicSSM Block*, which is central to the *CircularStatE Module*, is responsible for mitigating distortions

Datasets	Data		Domain		Trk Len	No. Seq	No. Smp	No. T
	Cov.	Pano.	Platform	Movement				
KITTI MOT [27]	*n.a.*	✗			*n.a.*	21	8k	749
Waymo [50]	220°	✗			20s	103k	20m	*n.a.*
nuScenes [8]	360°	✗			20s	1000	40k	*n.a.*
BDD100K MOT [74]	*n.a.*	✗			40s	2000	398k	*n.a.*
SportsMOT [15]	*n.a.*	✗			*n.a.*	240	150k	3401
DanceTrack [61]	*n.a.*	✗			*n.a.*	100	105k	990
JRDB [49]	360°	✗			≤117s	54	20k	*n.a.*
MOT17 [53]	*n.a.*	✗			≤85s	14	11k	1331
MOT20 [17]	*n.a.*	✗			≤133s	8	13k	3833
QuadTrack (ours)	360°	✓			60s	32	19k	332

Table 1. Typical datasets for 2D tracking. Abbreviations: (Autonomous Car), (Mobile Robot), (Quadruped Robot), (Internet images/videos), (Wheels), (Gait), (Stationary), Cov. (Coverage), Pano. (Panoramic camera), Trk Len (Track Length), No. Seq (The number of sequences), No. Smp (The number of samples), and No. T (the number of tracks).

and refining the feature map. The operation is broken down into the following steps:

Distortion and Scale Calculation. The first step is to compute both the distortion and scale information from the input feature map S_4:

$$\mathbf{d}, \mathbf{s} = \mathcal{D}(S_4),\ \sigma(\mathcal{M}(S_4)), \tag{8}$$

where, $\mathbf{d}$ and $\mathbf{s}$ represent the distortion and scale, respectively, both of which have dimensions $\mathbb{R}^{B\times C\times W\times H}$.

Mitigate Distortion. To correct distortions, we apply a dynamic convolution $\mathcal{D}_{conv}$ to refine the feature map. The operation can be expressed as:

$$\mathbf{D} = \mathcal{D}_{conv}(\mathbf{d} \odot \mathbf{s}, S_4), \tag{9}$$

where the symbol $\odot$ represents the Hadamard product, ensuring effective integration of scale adjustments.

Improve Consistency. Following distortion correction, a State Space Model (SSM) [16] is applied to enhance light and color consistency in the panoramic image. The input to this step is the output from the previous stage, denoted as $\mathbf{D}\in\mathbb{R}^{B\times C\times W\times H}$, and can be represented as follows:

$$\mathbf{D}^*[b,c,x,y] = \frac{1}{N}\sum_{d\in\{scan\}} F_{S6}(S_d(\mathbf{D}[b,c,x,y])), \tag{10}$$

where N represents the number of scans, S_d represents the scanning function, and F_{S6} is the transformation function for the S6 block [16].

Feature Fusion. Finally, the outputs from the dynamic convolution branch and the residual branch are fused. The fusion module $\mathcal{F}$ combines the refined feature map $\mathbf{D}^*$ with a processed version of S_4 (obtained via a CNN operation $\mathcal{C}(S_4)$) to yield the final output feature map $\mathbf{F}$:

$$\mathbf{F} = \mathcal{F}(\mathcal{C}(S_4) \oplus \mathbf{D}^*). \tag{11}$$

$\oplus$ denotes the feature fusion operation, combining details from both branches for optimal feature representation.

4. QuadTrack: a Dynamic 360° MOT Dataset

Most existing MOT datasets [17, 53, 61] are captured using pinhole cameras, which are characterized by a narrow-FoV and linear sensor motion. However, when panoramic-FoV capture devices experience even slight movements, the entire scene can change drastically, posing significant challenges for object tracking. QuadTrack addresses this challenge by providing a benchmark specifically designed to test MOT algorithms under dynamic, non-linear motion conditions. It enables evaluating algorithm robustness in tracking objects with panoramic, non-uniform motion.

4.1. Dataset Collection and Challenges

To acquire a dataset with a panoramic FoV and complex motion dynamics, we utilized a quadruped robot dog as the data collection platform. This platform was selected for its biomimetic gait, which emulates the natural locomotion patterns of quadrupedal animals, introducing additional challenges for motion tracking due to its inherent complexity and variability. The robot measures $70cm\times31cm\times40cm$, with a maximum payload capacity of $7kg$. It can navigate vertical obstacles up to $15cm$ and inclines up to 30°, making it highly maneuverable in everyday environments. With 12 joint motors, the robot replicates realistic walking motions at speeds up to $2.5m/s$. For sensing, we used a Panoramic Annular Lens (PAL) camera to capture wide-angle scenes with a FoV of $360°\times70°$. The camera has a pixel size of $3.45\mu m\times3.45\mu m$, a resolution of 5 million effective pixels, and supports a maximum output of 2048×2048 pixels at 40.5FPS. Mounted on the quadruped robot (see Fig. 4 (b)), the camera ensures an

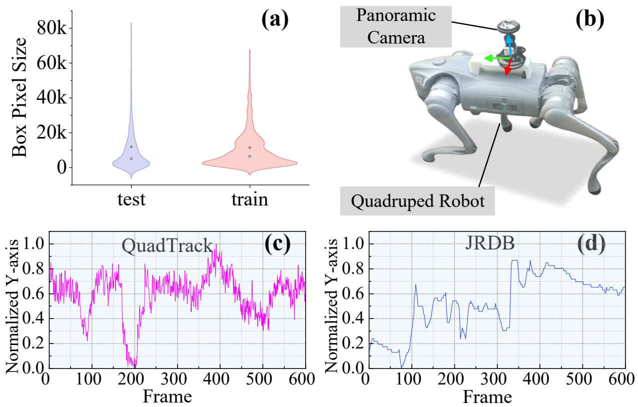

Figure 4. (a) shows the bounding box (bbox) size distribution for the training and validation sets, whereas (b) depicts the data collection platform and panoramic camera setup. (c) and (d) compare the normalized Y-axis pixel positions of trajectories between the QuadTrack ( ) and JRDB [49] ( ) datasets, illustrating the significant vertical motion of the sensor in QuadTrack.

unobstructed, optimal view. Using this platform, the outdoor data collection spans morning, noon, afternoon, and evening, in diverse unconstrained environments across five campuses in two cities.

With the biomimetic gait of the quadruped robot, the collected panoramic images naturally exhibited characteristic shaking, particularly along the Y-axis (Fig. 4 (c) and (d)). Compared to the JRDB dataset [49], our QuadTrack dataset introduces more complex motion challenges. Additionally, the data faces challenges such as uneven exposure, color inconsistencies due to the panoramic FoV, and increased motion blur, as rapid relative displacement between moving objects and the background intensifies the blurring effect. More details can be found in the supplementary.

4.2. Data Distribution and Comparative Analysis

Unlike existing panoramic MOT datasets [17, 27, 53], which rely on pinhole cameras, QuadTrack, as shown in Tab. 1, is the first to be captured using a single 360° panoramic camera. With a panoramic FoV (360°×70°), QuadTrack significantly differs from traditional MOT datasets [17, 53]. In contrast to autonomous driving datasets [8, 50, 74], which often feature more predictable motion, QuadTrack incorporates complex, biologically inspired gait movements. Moreover, unlike internet-sourced datasets [15, 61], QuadTrack is designed to better reflect real-world application scenarios. While many existing datasets [8, 50, 52, 74] consist of short video sequences, QuadTrack emphasizes long-term tracking, with each video lasting 60 seconds. To further challenge data association, we downsampled the dataset to 10FPS, resulting in 600 frames per sequence, spread across 32 sequences. In total, QuadTrack includes $19,200$ frames and $189,876$ bounding boxes.

As illustrated in Fig. 4 (a), the distribution of both the

	Method	HOTA↑	OSPA↓	IDF1↑	MOTA ↑
E2E	TrackFormer [51]	19.16	0.95	19.66	17.79
	MOTRv2 [79]	18.22	**0.93**	19.30	12.30
	OmniTrack$_{E2E}$ (ours)	**21.56**	0.94	**22.87**	**25.01**
TBD	SORT [7]	23.49	0.90	26.11	24.59
	DeepSORT [68]	22.15	0.95	23.46	24.88
	ByteTrack [78]	25.00	0.86	27.95	26.59
	Bot-SORT [1]	22.90	0.91	24.27	23.08
	OC-SORT [9]	25.04	**0.84**	27.89	25.64
	HybridSORT [72]	25.01	0.85	27.82	25.03
	DiffMOT [48]	19.96	0.95	20.26	20.05
	OmniTrack$_{DA}$ (ours)	**26.92**	**0.84**	**30.26**	**26.60**

Table 2. Comparison with state-of-the-art methods on the JRDB test set [49].

	Method	HOTA↑	OSPA↓	IDF1↑	MOTA ↑
E2E	TrackFormer [51]	19.62	0.97	17.75	**3.16**
	MOTRv2 [79]	16.42	**0.96**	17.08	-0.06
	OmniTrack$_{E2E}$ (ours)	**19.87**	0.98	**19.47**	-5.89
TBD	SORT [7]	14.57	0.98	15.60	4.81
	DeepSORT [68]	21.16	0.96	22.56	5.12
	ByteTrack [78]	20.66	**0.94**	22.56	8.68
	Bot-SORT [1]	15.77	0.99	15.65	5.92
	OC-SORT [9]	20.83	**0.94**	22.60	7.65
	HybridSORT [72]	16.64	0.96	17.38	6.79
	DiffMOT [48]	16.40	0.97	16.62	6.21
	OmniTrack$_{DA}$ (ours)	**23.45**	**0.94**	**26.41**	**9.68**

Table 3. Comparison with state-of-the-art methods on the QuadTrack test set.

training and test sets is consistent, ensuring a reliable and balanced evaluation of MOT methods. This similarity in the distribution between the sets reduces the potential for bias and allows for a more accurate comparison of model performance across varying conditions. The trajectories depicted in Fig. 4 (c) and (d) highlight the increased complexity of multi-object tracking under panoramic FoV conditions. Notably, the motion along the Y-axis is significantly more intense compared to JRDB [49], further increasing the difficulty of object detection and association.

5. Experiments

5.1. Experiment Setup

Datasets. We conduct experiments on two datasets: JRDB [49] and QuadTrack. JRDB is a panoramic dataset designed for crowded human environments, comprising 10 training sequences, 7 validation sequences, and 27 test sequences. The panoramic images in this dataset are stitched using a wheeled mobile robot equipped with five pinhole cameras. It includes both outdoor and indoor scenes, characterized by significant occlusion and the presence of small objects. Additionally, some objects exhibit rapid relative motion to the robot, which presents substantial challenges for MOT algorithms. Detailed information regarding the QuadTrack dataset is elaborated in Sec. 4.

	Association Method	Detection Method	HOTA ↑	IDF1↑	OSPA ↓	MOTA↑	DetA ↑	AssA ↑	FPS ↑
Track-By- Detection (TBD)	SORT [7]	YOLO11 [63] (baseline)	25.83	29.56	0.915	31.02	27.62	24.51	49.18
		OmniTrack$_{Det}$ (ours)	26.34 (+0.51)	31.11 (+1.55)	0.907 (-0.008)	34.21 (+3.19)	30.52 (+2.90)	22.96 (-1.55)	12.14
		OmniTrack$_{DA}$ (ours)	29.44 (+3.10)	33.27 (+2.16)	0.927(+0.020)	33.44 (-0.77)	35.16 (+4.64)	25.06 (+2.10)	11.78
	ByteTrack [78]	YOLO11 [63] (baseline)	27.85	32.20	0.896	34.46	31.49	25.15	50.36
		OmniTrack$_{Det}$(ours)	28.14 (+0.29)	32.97 (+0.77)	0.870 (-0.026)	37.36 (+2.90)	32.94 (+1.45)	24.29 (-0.86)	12.24
		OmniTrack$_{DA}$ (ours)	29.58 (+1.44)	34.54 (+1.57)	0.859 (-0.011)	38.14 (+0.78)	34.71 (+1.77)	25.49 (+1.20)	11.83
	OC-SORT [9]	YOLO11 [63] (baseline)	29.26	33.69	0.874	34.22	31.81	27.48	46.33
		OmniTrack$_{Det}$ (ours)	29.43 (+0.17)	34.11 (+0.42)	0.851 (-0.023)	38.72 (+4.50)	34.48 (+2.67)	25.39 (-2.09)	11.59
		OmniTrack$_{DA}$ (ours)	30.65 (+1.22)	34.83 (+0.72)	0.838 (-0.013)	36.37 (-2.35)	35.58 (+1.10)	26.76 (+1.37)	11.13
	HybridSORT [72]	YOLO11 [63] (baseline)	29.71	34.16	0.877	34.71	31.70	28.39	44.34
		OmniTrack$_{Det}$ (ours)	30.00 (+0.29)	34.09 (-0.07)	0.853 (-0.024)	32.32 (-2.39)	35.02 (+3.32)	26.09 (-2.30)	11.65
		OmniTrack$_{DA}$ (ours)	31.05 (+1.05)	36.06 (+1.97)	0.850 (-0.003)	38.13 (+5.81)	35.08 (+0.06)	27.78 (+1.69)	10.96
E2E	TrackFormer [51]	*n.a.*	22.22	23.38	0.959	23.83	30.30	16.93	7.38
	MOTR [76]	*n.a.*	19.78	23.25	0.928	25.44	25.51	15.61	12.73
	MOTRv2 [79]	*n.a.*	24.68	25.49	**0.911**	17.05	26.83	22.97	**13.01**
	OmniTrack$_{E2E}$ (ours)	*n.a.*	**25.12**	**27.42**	0.925	**34.99**	**33.35**	**19.17**	11.64

Table 4. Results on the JRDB validation set [49]. The first four groups compare methods under the TBD paradigm, whereas the last group presents a comparison under the E2E paradigm. In the TBD paradigm, each method is evaluated under three detection methods: the baseline with YOLO11 [63] as the detector, the OmniTrack$_{Det}$ detector, and OmniTrack$_{DA}$. The numbers represent the improvement relative to the previous line's method. The FPS metric is measured on a single RTX 3090 GPU with an image resolution of 4160×480.

Exp.	I_{dn}	I_{ft}	HOTA↑	IDF1↑	OSPA↓	MOTA↑
①	-	-	0.01	0.00	1.00	0.00
②		✓	3.80	1.91	0.99	-0.01
③	✓		24.32	26.20	0.93	29.25
④	✓	✓	**25.12**	**27.42**	**0.93**	**34.99**

Table 5. Analysis of FlexiTrack Instance: I_{dn} represents an instance generated using Ground Truth (GT), whereas I_{ft} refers to a FlexiTrack Instance.

Metrics. We employ the CLEAR metrics [6], including MOTA, DetA, and AssA, alongside IDF1 [57], OPSA [49], and HOTA [47] for a comprehensive tracking performance evaluation. MOTA is primarily influenced by detector performance, IDF1 measures identity preservation, and HOTA integrates association and localization accuracy, making it increasingly pivotal for tracking assessment.

Implementation details. To enable a fair comparison of various MOT algorithms, we retrained models on the JRDB dataset. For End-To-End (E2E) algorithms [51, 76, 79], we trained using the default parameters from the source code on JRDB. For the MOT algorithms [7, 9, 72, 78] based on the TBD paradigm, we selected the advanced YOLO11-X [63] as the baseline detector for training on JRDB. Additionally, OmniTrack$_{Det}$ was obtained by masking the Track Management module after training the OmniTrack model and saving the detection results. The AdamW optimizer [40] was used, with the learning rate set to 10^{-5}. For additional experimental details, please refer to the supplementary.

5.2. Comparison with State of the Art

Tracking on JRDB test set. In Tab. 2, we compare our OmniTrack with state-of-the-art methods on the JRDB test set. Firstly, our approach significantly outperforms existing algorithms across all tracking metrics, whether in comparison with End-to-End or TBD paradigms. Specifically, OmniTrack achieves an impressive HOTA of 21.56% and an IDF1 of 22.87% within the End-to-End framework, surpassing the current state-of-the-art method, MOTRv2 [79], by 3.34% and 3.57%, respectively. Furthermore, in the TBD paradigm, even under the same detector conditions, OmniTrack outperforms the state-of-the-art HybridSORT [72] by 1.91% in HOTA and 2.44% in IDF1, demonstrating its superior performance.

Tracking on QuadTrack test set. In Tab. 3, we present a comparison between OmniTrack and state-of-the-art methods on the QuadTrack test set. This dataset is particularly challenging, characterized by a panoramic FoV and rapid, non-linear sensor motion, which introduces significant complexities for traditional MOT algorithms. Despite these challenges, our method outperforms existing approaches, achieving the highest HOTA scores: 19.87% for the E2E group and 23.45% for the TBD group.

5.3. Paradigm Comparison

Baseline. To further validate the advantages of OmniTrack, we conducted comparisons based on the TBD and E2E paradigms, as shown in Tab. 4. In the TBD paradigm, we evaluated several baseline tracking algorithms [7, 9, 72, 78]. Each tracking method was compared under three different detection setups: using YOLO11-X [63] as the baseline detector, OmniTrack$_{Det}$ as the detector (representing traditional TBD tracking where detection and tracking are independent), and OmniTrack$_{DA}$ with a feedback mechanism for TBD tracking. In the E2E paradigm, we used MOTR [76] as the baseline for comparison.

Result. In the TBD method, OmniTrack$_{Det}$ consistently outperforms YOLO11-X [63], showing an average improvement of 0.2% in HOTA and 0.6% in IDF1. Despite OmniTrack$_{Det}$ not having a speed advantage, it

Exp.	$\mathcal{S}_5$	$\mathcal{S}_4$	$\mathcal{S}_3$	HOTA↑	IDF1↑	OSPA↓
①	-	-	-	23.296	25.496	0.93415
②	MLP	MLP	MLP	21.951	23.535	0.92151
③	Conv	Conv	Conv	23.565	25.814	**0.90931**
④	✓	✓	✓	24.724	26.886	0.91934
⑤	✓			24.426	26.016	0.92819
⑥			✓	24.539	26.506	0.92776
⑦		✓		**25.120**	**27.423**	0.92512

Table 6. Ablation study on the CircularStatE module. S_3, S_4, and S_5 represent multi-scale features extracted from the backbone [29]. *MLP* refers to fully connected layers, *Conv* to convolutional layers. The symbol ✓ indicates the use of *DynamicSSM* 3.4

achieves notable improvements in accuracy. Furthermore, when comparing OmniTrack$_{Det}$ to OmniTrack$_{DA}$, the latter shows an average increase of 1.7% in HOTA and 1.4% in IDF1, demonstrating the effectiveness of the feedback mechanism. In the E2E paradigm, OmniTrack$_{E2E}$ achieved the best result HOTA of 25.12% and IDF1 of 27.42%.

5.4. Ablation Study

Analysis of the FlexiTrack instance. Tab. 5 compares experiments with and without denoise instances and FlexiTrack instances during the training phase. Experiments ① and ② demonstrate that FlexiTrack Instances are crucial for achieving the tracking objective. In Experiment ③, we observe that denoise instances, generated from Ground Truth (GT), significantly improve the HOTA score by providing stronger guidance. Experiments ③ and ④ further show that incorporating FlexiTrack instances after using denoise instances leads to a further improvement in the HOTA score.

Analysis of the CircularStatE module. In Tab. 6, we evaluate the effectiveness of *DynamicSSM* in the *CircularStatE*, comparing it with other common designs such as Conv and MLP. The results from experiments ②, ③, and ④ demonstrate a clear advantage for DynamicSSM. Experiments ⑤, ⑥, and ⑦ further show that applying DynamicSSM to S_4 yields the best performance. where S_5, S_4, and S_3 impact MOT results. Since S_4 contains both high-level semantic and low-level geometric features, its effect is the most pronounced.

Analysis of the initialization and update thresholds. In OmniTrack$_{E2E}$, we analyzed the impact of the *initial threshold* and *updated threshold* on tracking performance. As shown in Fig. 5, both the *initial threshold* and *updated threshold* achieved HOTA scores exceeding 25% within the range of 0.1 to 0.7. This demonstrates that OmniTrack$_{E2E}$ is robust to threshold variations, eliminating the need for fine-tuning to achieve optimal results.

Comparison of end-to-end model training. In Tab. 7, we compare the number of parameters and training time of OmniTrack$_{E2E}$ with existing End-to-End methods. Our method trains over four times faster than other End-to-End methods using default parameters on the JRDB dataset. This is achieved by implementing identity association through FlexiTrack Instances, which significantly simplifies the model design of the association component and alleviates the challenges associated with E2E model training.

Method	#Params	FLOPs	MACs	Training Time ↓
TrackFormer [51]	44.01M	335G	167G	108 hours
MOTR [76]	43.91M	1421G	709G	80 hours
MOTRv2 [79]	41.65M	1395G	696G	130 hours
OmniTrack$_{E2E}$ (ours)	63.13M	762G	369G	20 hours

Table 7. Comparison of parameters, FLOPs, MACs, and training time for various end-to-end models on the JRDB dataset [49].

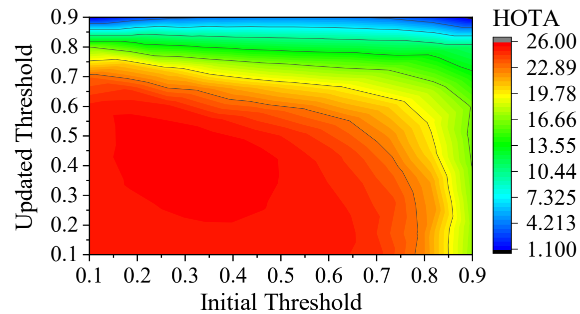

Figure 5. Effects of the trajectory initialization threshold and update threshold on the HOTA metric in OmniTrack$_{E2E}$.

6. Conclusion

This paper presents OmniTrack, a multi-object tracking framework tailored for panoramic images, effectively addressing key challenges like geometric distortion, low resolution, and lighting inconsistencies. Central to OmniTrack is a feedback mechanism that reduces uncertainty in panoramic-FoV tracking. The framework incorporates Tracklets Management for temporal stability, FlexiTrack Instance for rapid localization and association, and the CircularStatE Module to mitigate distortion and improve visual consistency. Additionally, we present QuadTrack, a cross-campus multi-object tracking dataset collected using a quadruped robot to support dynamic motion scenarios. This challenging dataset is designed to advance research in omnidirectional perception for robotics. Experiments verify that OmniTrack achieves state-of-the-art performance on public JRDB and the established QuadTrack datasets, demonstrating its effectiveness in handling panoramic tracking tasks.

Limitations. While OmniTrack demonstrates strong performance, our approach is currently limited to 2D panoramic tracking without 3D capabilities, restricting depth perception in complex scenes. Additionally, the method is centered around a mobile robotic platform. Future work could consider extending to 3D panoramic MOT or exploring human-robot collaborative perception to enhance situational awareness.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 62473139 and No. 12174341), in part by Zhejiang Provincial Natural Science Foundation of China (Grant No. LZ24F050003), and in part by Shanghai SUPREMIND Technology Co. Ltd.

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. 2, 3, 6

[2] Hao Ai and Lin Wang. Elite360D: Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion. In *CVPR*, 2024. 2

[3] Hao Ai, Zidong Cao, and Lin Wang. A survey of representation learning, optimization strategies, and applications for omnidirectional vision. *International Journal of Computer Vision*, 2025. 2

[4] Jiayang Bai, Haoyu Qin, Shuichang Lai, Jie Guo, and Yanwen Guo. GLPanoDepth: Global-to-local panoramic depth estimation. *IEEE Transactions on Image Processing*, 2024. 2

[5] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jürgen Gall. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In *ICCV*, 2019. 2

[6] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008. 7

[7] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, 2016. 6, 7

[8] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2, 5, 6

[9] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric SORT: Rethinking SORT for robust multi-object tracking. In *CVPR*, 2023. 2, 3, 4, 6, 7

[10] Yihong Cao, Jiaming Zhang, Hao Shi, Kunyu Peng, Yuhongxuan Zhang, Hui Zhang, Rainer Stiefelhagen, and Kailun Yang. Occlusion-aware seamless segmentation. In *ECCV*, 2024. 1, 2

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[12] Wenjie Chang, Yueyi Zhang, and Zhiwei Xiong. Depth estimation from indoor panoramas with neural scene representation. In *CVPR*, 2023. 2

[13] Hao Chen, Yuqi Hou, Chenyuan Qu, Irene Testini, Xiaohan Hong, and Jianbo Jiao. 360+x: A panoptic multi-modal scene understanding dataset. In *CVPR*, 2024. 2

[14] Sijia Chen, En Yu, Jinyang Li, and Wenbing Tao. Delving into the trajectory long-tail distribution for muti-object tracking. In *CVPR*, 2024. 1, 2

[15] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. SportsMOT: A large multi-object tracking dataset in multiple sports scenes. In *ICCV*, 2023. 2, 5, 6

[16] Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *ICML*, 2024. 5

[17] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian D. Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. MOT20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 2, 5, 6

[18] Shuxiao Ding, Lukas Schneider, Marius Cordts, and Juergen Gall. ADA-Track: End-to-end multi-camera 3D multi-object tracking with alternating detection and association. In *CVPR*, 2024. 2

[19] Yuan Dong, Chuan Fang, Liefeng Bo, Zilong Dong, and Ping Tan. PanoContext-Former: Panoramic total scene understanding with a transformer. In *CVPR*, 2024. 2

[20] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. StrongSORT: Make DeepSORT great again. *IEEE Transactions on Multimedia*, 2023. 2, 3

[21] Yunhao Du, Cheng Lei, Zhicheng Zhao, and Fei Su. iKUN: Speak to trackers without retraining. In *CVPR*, 2024. 2

[22] Mahsa Ehsanpour, Fatemeh Sadat Saleh, Silvio Savarese, Ian D. Reid, and Hamid Rezatofighi. JRDB-Act: A large-scale dataset for spatio-temporal action, social group and activity detection. In *CVPR*, 2022. 2

[23] Kanglong Fan, Wen Wen, Mu Li, Yifan Peng, and Kede Ma. Learned scanpaths aid blind panoramic video quality assessment. In *CVPR*, 2024. 2

[24] Ruopeng Gao and Limin Wang. MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In *ICCV*, 2023. 2

[25] Shaohua Gao, Kailun Yang, Hao Shi, Kaiwei Wang, and Jian Bai. Review on panoramic imaging and its applications in scene understanding. *IEEE Transactions on Instrumentation and Measurement*, 2022. 2

[26] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding YOLO series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2

[27] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 2013. 5, 6

[28] Ruize Han, Haomin Yan, Jiacheng Li, Songmiao Wang, Wei Feng, and Song Wang. Panoramic human activity recognition. In *ECCV*, 2022. 1, 2

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 8

[30] Cheng Huang, Shoudong Han, Mengyu He, Wenbo Zheng, and Yuhao Wei. DeconfuseTrack: Dealing with confusion for multi-object tracking. In *CVPR*, 2024. 2

[31] Huajian Huang, Yinzhe Xu, Yingshu Chen, and Sai-Kit Yeung. 360VOT: A new benchmark dataset for omnidirectional visual object tracking. In *ICCV*, 2023. 2

[32] Huajian Huang, Changkun Liu, Yipeng Zhu, Hui Cheng, Tristan Braud, and Sai-Kit Yeung. 360Loc: A dataset and benchmark for omnidirectional visual localization with cross-device queries. In *CVPR*, 2024. 2

[33] Alexander Jaus, Kailun Yang, and Rainer Stiefelhagen. Panoramic panoptic segmentation: Towards complete surrounding understanding via unsupervised contrastive learning. In *IV*, 2021. 2

[34] Alexander Jaus, Kailun Yang, and Rainer Stiefelhagen. Panoramic panoptic segmentation: Insights into surrounding parsing for mobile agents via unsupervised contrastive learning. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 2

[35] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. UniFuse: Unidirectional fusion for 360° panorama depth estimation. *IEEE Robotics and Automation Letters*, 2021. 2

[36] Mingjun Jiang, Ryo Sogabe, Kohei Shimasaki, Shaopeng Hu, Taku Senoo, and Idaku Ishii. 500-Fps omnidirectional visual tracking using three-axis active vision system. *IEEE Transactions on Instrumentation and Measurement*, 2021. 2

[37] Qi Jiang, Hao Shi, Lei Sun, Shaohua Gao, Kailun Yang, and Kaiwei Wang. Annular computational imaging: Capture clear panoramic images through simple lens. *IEEE Transactions on Computational Imaging*, 2022. 2

[38] Qi Jiang, Shaohua Gao, Yao Gao, Kailun Yang, Zhonghua Yi, Hao Shi, Lei Sun, and Kaiwei Wang. Minimalist and high-quality panoramic imaging with PSF-aware transformers. *IEEE Transactions on Image Processing*, 2024. 2

[39] Junho Kim, Jiwon Jeong, and Young Min Kim. Fully geometric panoramic localization. In *CVPR*, 2024. 2

[40] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 7

[41] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 1955. 2

[42] Jialu Li and Mohit Bansal. PanoGen: Text-conditioned panoramic environment generation for vision-and-language navigation. In *NeurIPS*, 2023. 2

[43] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. OVTrack: Open-vocabulary multiple object tracking. In *CVPR*, 2023. 2

[44] Yiheng Li, Connelly Barnes, Kun Huang, and Fang-Lue Zhang. Deep 360° optical flow estimation based on multi-projection fusion. In *ECCV*, 2022. 2

[45] Yanwei Li, Zhiding Yu, Jonah Philion, Anima Anandkumar, Sanja Fidler, Jiaya Jia, and Jose Alvarez. End-to-end 3D tracking with decoupled queries. In *ICCV*, 2023. 2

[46] Zhixin Ling, Zhen Xing, Xiangdong Zhou, Manliang Cao, and Guichun Zhou. PanoSwin: a pano-style swin transformer for panorama understanding. In *CVPR*, 2023. 2

[47] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip H. S. Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 2021. 7

[48] Weiyi Lv, Yuhang Huang, Ning Zhang, Ruei-Sung Lin, Mei Han, and Dan Zeng. DiffMOT: A real-time diffusion-based multiple object tracker with non-linear prediction. In *CVPR*, 2024. 1, 2, 3, 6

[49] Roberto Martín-Martín, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. JRDB: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 5, 6, 7, 8

[50] Jieru Mei, Alex Zihao Zhu, Xinchen Yan, Hang Yan, Siyuan Qiao, Liang-Chieh Chen, and Henrik Kretzschmar. Waymo open dataset: Panoramic video panoptic segmentation. In *ECCV*, 2022. 5, 6

[51] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. TrackFormer: Multi-object tracking with transformers. In *CVPR*, 2022. 3, 6, 7, 8

[52] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *CVPR*, 2022. 6

[53] Anton Milan, Laura Leal-Taixé, Ian D. Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 2, 5, 6

[54] Jonghyuk Park, Hyeona Kim, Eunpil Park, and Jae-Young Sim. Fully-automatic reflection removal for 360-degree images. In *WACV*, 2024. 2

[55] Zheng Qin, Sanping Zhou, Le Wang, Jinghai Duan, Gang Hua, and Wei Tang. MotionTrack: Learning robust short-term and long-term motions for multi-object tracking. In *CVPR*, 2023. 2

[56] Zheng Qin, Le Wang, Sanping Zhou, Panpan Fu, Gang Hua, and Wei Tang. Towards generalizable multi-object tracking. In *CVPR*, 2024. 2

[57] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCVW*, 2016. 7

[58] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. PanoFormer: Panorama transformer for indoor 360° depth estimation. In *ECCV*, 2022. 2

[59] Zhijie Shen, Zishuo Zheng, Chunyu Lin, Lang Nie, Kang Liao, Shuai Zheng, and Yao Zhao. Disentangling orthogonal planes for indoor panoramic room layout estimation with cross-scale distortion awareness. In *CVPR*, 2023. 2

[60] Hao Shi, Yifan Zhou, Kailun Yang, Xiaoting Yin, Ze Wang, Yaozu Ye, Zhe Yin, Shi Meng, Peng Li, and Kaiwei Wang. PanoFlow: Learning 360° optical flow for surrounding temporal understanding. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 1, 2

[61] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. DanceTrack: Multi-object tracking in uniform appearance and diverse motion. In *CVPR*, 2022. 5, 6

[62] Zhifeng Teng, Jiaming Zhang, Kailun Yang, Kunyu Peng, Hao Shi, Simon Reiß, Ke Cao, and Rainer Stiefelhagen. 360BEV: Panoramic semantic mapping for indoor bird's-eye view. In *WACV*, 2024. 2

[63] Ultralytics. YOLO vision. https://github.com/ultralytics/ultralytics, 2024. Accessed: 2024-11-10. 7

[64] Tom van Dijk, Christophe De Wagter, and Guido C. H. E. de Croon. Visual route following for tiny autonomous robots. *Science Robotics*, 2024. 1

[65] Fu-En Wang, Yu-Hsuan Yeh, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. BiFuse++: Self-supervised and efficient bi-projection fusion for 360° depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[66] Qian Wang, Weiqi Li, Chong Mou, Xinhua Cheng, and Jian Zhang. 360DVD: Controllable panorama video generation with 360-degree video diffusion model. In *CVPR*, 2024. 2

[67] Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *CVPR*, 2024. 1

[68] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017. 2, 6

[69] Zhendong Wu, Lintao Zhao, Guocui Liu, Jingchun Chai, Jierui Huang, and Xiaoqun Ai. The effect of AR-HUD takeover assistance types on driver situation awareness in highly automated driving: A 360-degree panorama experiment. *International Journal of Human-Computer Interaction*, 2024. 1

[70] Yinzhe Xu, Huajian Huang, Yingshu Chen, and Sai-Kit Yeung. 360VOTS: Visual object tracking and segmentation in omnidirectional videos. *arXiv preprint arXiv:2404.13953*, 2024. 2

[71] Shilin Yan, Xiaohao Xu, Renrui Zhang, Lingyi Hong, Wenchao Chen, Wenqiang Zhang, and Wei Zhang. PanoVOS: Bridging non-panoramic and panoramic views with transformer for video segmentation. In *ECCV*, 2024. 2

[72] Mingzhan Yang, Guangxin Han, Bin Yan, Wenhua Zhang, Jinqing Qi, Huchuan Lu, and Dong Wang. Hybrid-SORT: Weak cues matter for online multi-object tracking. In *AAAI*, 2024. 4, 6, 7

[73] Kefu Yi, Kai Luo, Xiaolei Luo, Jiangui Huang, Hao Wu, Rongdong Hu, and Wei Hao. UCMCTrack: Multi-object tracking with uniform camera motion compensation. *AAAI*, 2024. 2

[74] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 2, 5, 6

[75] Haozheng Yu, Lu He, Bing Jian, Weiwei Feng, and Shan Liu. PanelNet: Understanding 360 indoor environment via panel representation. In *CVPR*, 2023. 2

[76] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. MOTR: End-to-end multiple-object tracking with transformer. In *ECCV*, 2022. 2, 3, 7, 8

[77] Jiaming Zhang, Kailun Yang, Hao Shi, Simon Reiß, Kunyu Peng, Chaoxiang Ma, Haodong Fu, Philip H. S. Torr, Kaiwei Wang, and Rainer Stiefelhagen. Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[78] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *ECCV*, 2022. 2, 3, 6, 7

[79] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. MOTRv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *CVPR*, 2023. 3, 6, 7, 8

[80] Guangze Zheng, Shijie Lin, Haobo Zuo, Changhong Fu, and Jia Pan. NetTrack: Tracking dynamic objects with a net. In *CVPR*, 2024. 2

[81] Xu Zheng, Pengyuan Zhou, Athanasios V. Vasilakos, and Lin Wang. 360SFUDA++: Towards source-free UDA for panoramic segmentation by learning reliable category prototypes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[82] Xu Zheng, Pengyuan Zhou, Athanasios V. Vasilakos, and Lin Wang. Semantics distortion and style matter: Towards source-free UDA for panoramic segmentation. In *CVPR*, 2024. 2

[83] Shijie Zhou, Zhiwen Fan, Dejia Xu, Haoran Chang, Pradyumna Chari, Tejas Bharadwaj, Suya You, Zhangyang Wang, and Achuta Kadambi. DreamScene360: Unconstrained text-to-3D scene generation with panoramic gaussian splatting. In *ECCV*, 2024. 2