2024 7th International Symposium on Autonomous Systems (ISAS) | 979-8-3503-6317-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISAS61044.2024.10552604

Beyond Traditional Driving Scenes: A Robotic-Centric Paradigm for 2D+3D Human Tracking Using Siamese Transformer Network

1st Shuo Xin
Institute of Cyber-Systems and Control
Zhejiang University
Hangzhou, China
1135205402@qq.com

2nd Liang Liu*
Institute of Cyber-Systems and Control
Zhejiang University
Hangzhou, China
leonliuz@zju.edu.cn

3th Xiao Kang
China North Vehicle Research Institute
Beijing, China
kangxiaotop1@126.com

4rd Zhen Zhang
Institute of Cyber-Systems and Control
Zhejiang University
Hangzhou, China
zhenz@zju.edu.cn

5th Mengmeng Wang
Institute of Cyber-Systems and Control
Zhejiang University
Hangzhou, China
mengmengwang@zju.edu.cn

6th Yong Liu*
Institute of Cyber-Systems and Control
Zhejiang University
Hangzhou, China
yongliu@iipc.zju.edu.cn

Abstract*—3D human tracking plays a crucial role in the automation intelligence system. Current approaches focus on achieving higher performance on traditional driving datasets like KITTI, which overlook the jitteriness of the platform and the complexity of the environments. Once the scenarios are migrated to jolting robot platforms, they all degenerate severely with only a 20-60% success rate, which greatly restricts the high-level application of autonomous systems. In this work, beyond traditional flat scenes, we introduce Multi-modal Human Tracking Paradigm (MHTrack), a unified multimodal transformer-based model that can effectively track the target person frame-by-frame in point and video sequences. Specifically, we design a speed-inertia module-assisted stabilization mechanism along with an alternate training strategy to better migrate the tracking algorithm to the robot platform. To capture more target-aware information, we combine the geometric and appearance features of point clouds and video frames together based on a hierarchical Siamese Transformer Network. Additionally, considering the prior characteristics of the human category, we design a lateral cross-attention pyramid head for deeper feature aggregation and final 3D BBox generation. Extensive experiments confirm that MHTrack significantly outperforms the previous state-of-the-arts on both open-source datasets and large-scale robotic datasets. Further analysis verifies each component's effectiveness and shows the robotic-centric paradigm's promising potential when deployed into dynamic robotic systems.

***Index Terms*—human tracking, robotic platform, multi-modal, Transformer, strong disturbance**

I. Introduction

With the continuous development of automation intelligence, specific human tracking as the basic building block of the above tasks, has attracted extensive attention in the field of vision involving autonomous driving, scene understanding, and robotic manipulation. Currently, most human trackers [27, 29] focus on driving scenarios where people walk on flat paved roads and the recording platforms have neglected jitter and uncomplicated surroundings, like well-known open-source datasets KITTI [5], Waymo [24], and Nuscenes [2]. However, not all tracking scenarios are so perfect, which results in trackers' severe degeneration in complex environments, especially on jolting robot platforms with only a 20-60% tracking success rate. For example, when accomplishing freight services, the robot needs to follow a specific worker along the way and it may violently sway, jump, adjust its gait, or even up and down stairs, which inevitably causes the collected data to be subjected to severe shaking. Moreover, if the on-robot tracking algorithm cannot fit the bumpy platform and continuously keep up with the target person end-to-end, then the follow-up actions on the robot will result in errors that link to the accuracy of all subsequent modules, especially in challenging indoor or outdoor complex surroundings. Till now, a unified tracking algorithm to handle this problem with higher task levels has rarely appeared.

In addition, the current tracking algorithms of pure 2D and pure 3D targets are relatively mature, but illumination, distortion, occlusion, and the hard defects of a single sensor are still long-standing problems in this field. Some pioneers [4, 11, 16, 31] attempted to merge visual contexts with other sensors in the 2D object detection and tracking fields. However, most of them rely too on visual tracking and lack depth information, which limits them from estimating 3DD direction and spatial distance between humans and robots. The multi-sensor application in 2D+3D tracking is still at the primary stage.

Based on this, this paper proposes a robotic-centric paradigm for human tracking in various and complex scenarios. To relieve the strong shaking issues on the robot platform itself, a speed-inertia module-assisted stabilization mechanism

* Corresponding Author.

is proposed along with an alternate training strategy. To handle the influence of strong disturbance in complex environments, a new 2D+3D tracking model is proposed by fusing RGB videos and point cloud data that contain different information at the same time, thus fully realizing the sufficient complementarity of information. To further improve the tracking accuracy for the human category, the cross-attention-based decoder is proposed to enrich the potential representation of the template frame with human-ware priors. Experiments show that our method successfully migrates from the well-known open-source dataset KITTI to the physical quadruped robot platform while running in real time.

II. Related Work

A. 2D tracker and the Siamese framework

The development of 2D single object tracking (abbreviated as SOT) benefits from the rapid development of deep learning and CNN. Given the initial target position in a two-dimensional video, the specific purpose of 2D SOT is to distinguish an arbitrary object in multiple consecutive frames of the video stream. As an iconic breakthrough, SiamFC [1] in 2016 firstly transitioned the mainstream of 2D tracking from traditional correlation filtering-based methods to Siamese-based structures. The end-to-end Siamese framework utilizes a cross-feature module to calculate similarities between a template branch and several search branches. Subsequently, more and more studies [7, 13, 14, 28] have attempted to further improve tracking performance based on Siamese two-branch architecture. The various methods [3, 15, 33] include regarding both appearance and motion, estimating boundary flows, using contextual structures, attention mechanisms, semantic information for discrimination, triplet loss, region proposal networks, and so on. Nowadays, the 2D SOT is relatively mature with an average tracking accuracy of over 90%. However, due to modality limitations, 2D methods lack the key depth and spatial information, making it difficult to apply to more large-scale and high-level tracking tasks. In this work, we adopt the Siamese framework mentioned above for simple and efficient frame processing.

B. 3D tracker and the limitations of points

3D SOT, which is designed to determine both the location and 3D size of objects, has a wide range of practical applications. Given the initial object location as sequence input, the 3D tracking model is to output a three-dimensional bounding box (abbreviated as 3D BBox) frame-by-frame in the point cloud format. PointNet [20] is crucial for 3D object tracking and has laid a solid foundation for its development. In 2019, SC3D [6] for the first time used pure point cloud input for target tracking to construct a pure 3D tracker. This type of method is characterized by directly using deep learning networks to extract point clouds' features, deriving subsequent P2B [21], BAT [34], V2B [9]. Other mainstream studies tried to project the point clouds into 2D planes (such as birds-eye views, and front views). Represented by [36], they no longer attempt to directly process point clouds but use familiar 2D CNN for feature extraction and aggregation. However, the view conversion process itself will seriously lose data details, causing difficulty in meeting tracking accuracy requirements.

Till now, there are still some unresolved issues in relying solely on point clouds. The first is the sparsity of point clouds. As the distance between objects increases, the point clouds become increasingly sparse, which greatly hinders the sufficient feature extraction process. Secondly, calculating Siamese branches' similarity is difficult because point clouds are disordered. Previous work used image priors [12], shape completion [32], or feature enhancement [30] to address the issues mentioned above. Although they have achieved better tracking performance, the different regions of the object are easy to be ignored or mismatched during the tracking process. Thirdly, since the human is non-rigid and its size in the point cloud space is relatively small, relying solely on points may lead to missed detections of human objects. As a result, an efficient 3D tracking paradigm is desirable for the precise 3D description of non-rigid humans.

C. Multi-modal tracker

Except for different data formats, the object tracking process of multi-modal trackers is almost identical to the above two. The core step worth exploring is how to better fuse and align image data and point cloud data to obtain more favorable information. In recent years, scholars have been exploring new methods of 2D+3D fusion like [11, 16, 26, 31], but they are mostly applied for detection tasks or 2D SOT, not for 3D tracking. To address this issue, this paper explores a novel alignment-guided attention module for 2D+3D fusion to sufficiently realize the interaction between multi-modalities.

III. A Robotic-centric human tracking paradigm

The tracking algorithm proposed in this paper models human tracking as a bottom-up learning problem, as Fig. 1 shows. For the underlying architecture, the algorithm constructs a multi-modal tracking module for human targets, analyzing environmental and target information; For the upper architecture, a dynamic promoted-tracking module for quadruped robots was built, and a collaborative execution strategy was used to determine the current type of strategy to execute tracking. Based on these, the tracking paradigm can obtain the best learning strategy network according to the scenario of the algorithm application.

A. Multi-modal Tracking Model for Human Target

Fig. 2 (a) illustrates an overview of our **M**ulti-modal **H**uman **Track**ing (MHTrack) model. Using RGB videos and point clouds as dual inputs, MHTrack is proposed to output the spatial position and orientation of a single target human body in three-dimensional space while balancing accuracy and speed. Similar to previous work [8, 18], we use pre-trained models (CLIP [22] and PointMAE [17]) to extract deep features for each video frame and point cloud sequence. After CNN processing, features from untrimmed multimodality are projected by linear fully connected (FC) layers into a common

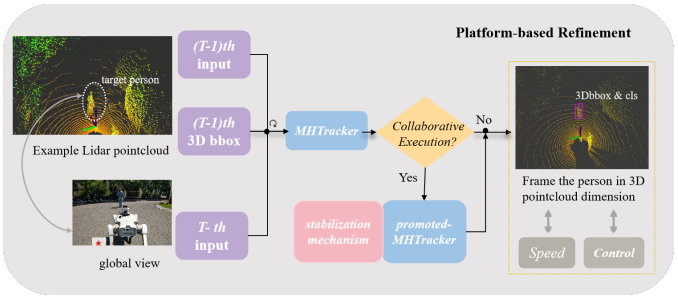

Fig. 1. Overall architecture. Based on different platforms, the proposed paradigm will adaptively switch between the multi-modal human tracking model MHTrack for traditional driving scenes and the promoted MHTrack for robot-centric scenes, while following platform-based refinements.

C-dimensional embedding space. Specially, we define the generated video and point features as $F^p = \{f_i^p\}_{i=1}^K \in \mathbb{R}^{K\times C}$ and $F^v = \{f_{(i,j)}^v\} \in R^{H'\times W'\times C}$, respectively. K denotes the number of point feature groups after point down-sampling and $H' \times W'$ denotes the dimension of the video feature map.

To learn more discriminative features, we propose an Alignment-Guided Self-Attention Module by utilizing the attention mechanism in Transformer to capture long-range contextual information of video and point features. In Fig. 2 (b), the network is hierarchical, consisting of each-modality *feature embedding* and multi-modality *feature interpolation*. For point-modality *feature embedding*, inspired by STNet [10], the encoder consists of L non-local feature embedding modules, which execute self-attention on feature maps at different scales. Specifically, in the l-th layer, we first execute the edge convolution using k-nearest neighbors to aggregate geometric features in the coordinate system, indicated by $E_l^p \in \mathbb{R}^{\frac{K}{2^l}\times C}$. Then, in order to discover long-range details about the point cloud, we use the self-attention on the feature map E_l^p, where $X_l^p \in \mathbb{R}^{\frac{K}{2^l}\times C}$ denotes the position embedding of the initially given BBox in the first frame. The video-modality *feature embedding* follows the same process and generates feature map $E_l^v \in \mathbb{R}^{\frac{H'}{2^l}\times\frac{W'}{2^l}\times C}$ and $X_l^v \in \mathbb{R}^{\frac{H'}{2^l}\times\frac{W'}{2^l}\times C}$. Formally, the attention mechanism is defined as:

$$F_l^p = SelfAttention(E_l^p + X_l^p, E_l^p + X_l^p, E_l^p + X_l^p) \quad (1)$$

$$F_l^v = SelfAttention(E_l^v + X_l^v, E_l^v + X_l^v, E_l^v + X_l^v) \quad (2)$$

In Eq. (1)(2), the query, key, and value are the three inputs utilized from left to right, accordingly. Till now, however, the obtained point-wise feature F_L^p and video-wise feature F_L^p still differ substantially in dimension. To address the multi-modal alignment issue, the adaptive *feature interpolation* is ultilized progressively to transform the low-dimensional point-wise feature to the high-dimensional video-wise features. Driven by previous work [25], the interpolated points $p \in \mathbb{R}^{N\times C}$ are projected into 2D coordinates $\hat{p} \in \mathbb{N}^{H'\times W'\times C}$ based on LiDAR-camera settings and then generate a projected feature map $F^p \in \mathbb{R}^{H'\times W'\times C}$ that have the same dimension as video-wise one. By applying the multi-layer perception χ and concatenation operator $\oplus$, we obtain the point-video-fused feature denoted $F^{pv} = \{f_{(i,j)}^{pv}\} \in \mathbb{R}^{H'\times W'\times C}$, which is written as:

$$F^{pv} = \sum_{i=1}^{H'}\sum_{j=1}^{W'} \chi^p(F_{(i,j)}^p) \oplus \chi^v(F_{(i,j)}^v) \quad (3)$$

As for multi-modal feature aggregation, we propose a 3D Transformer network to yield N^0 highest candidate BBoxes around the target. To make the algorithm not overly dependent on a single modality and to create a correlation map that helps locate the target, a cross-attention aggregation operation is utilized to calculate how similar the template and search areas are. We simply designate the acquired feature maps of the template and search region by $Y_t = F_t^{pv} \in \mathbb{R}^{N_t\times C}$ and $Y_s = F_s^{pv} \in \mathbb{R}^{N_s\times C}$. We create the cross-attention feature $\hat{Y}_s$ by embedding the template Y_t (value) into the search area Y_s (query). Specifically, the cross-attention aggregation operation is formulated as:

$$\hat{Y}_s = CrossAttention(Y_s, Y_t + X_t, Y_t + X_t) \quad (4)$$

We append the positional embedding of the template $X_t \in \mathbb{R}^{N_t\times C}$ to the value $Y_t \in \mathbb{R}^{N_t\times C}$ as the 3D coordinates of the template give the positional relationship of the target. In this way, the potential target in the feature fusion map can be associated with the template. For regressing, the 3D detector [19, 23] is used to generate the target center, target size, yaw angle, and confidence score of N^0 candidate BBoxes.

Furthermore, a human-ware head is designed for 3D BBox filtering and outputting, as Fig. 2 (c) shows. Note that the human category has some certain size priors(width: 10-40cm, height: 1.0-2.0m, length: 20-60cm), so we filter the coarse candidates by immediately discarding the oversized ones, which kind of refinement significantly reduces the time consumption of our network. Finally, with N^b BBoxes remaining, we utilize the Multilayer Perceptions(MLP) head to select the highest confidence score one as the final tracking result. The output prediction of i-th frame will be used to update the next frame's template, until traversing to the last frame.

B. Promoted Human Tracking for Quadruped Robots

The goal of our promoted robotic-centric model is to maximize the probability of bumpy robots and successfully track the target person when violently shaking. A core component of the promotion is the speed-inertia module-assisted stabilization mechanism, which is designed to recover the terminated targets caused by intense movements. We summarize the stabilization process into the five steps. To further improve the performance of the dynamic tracking system on the robot, we proposed a collaborative alternate training strategy to determine between the MHTrack module for general scenarios and the MHTrack-pro module for robot platforms. As Fig. **??** (b) shows, when training MHTrack-pro we fix the parameters of the MHTrack network and update only the promoted one. The same goes for the opposite and the collaborative training will change the target module when the loss is below 60%. Through alternating training, the tracking strategy not only

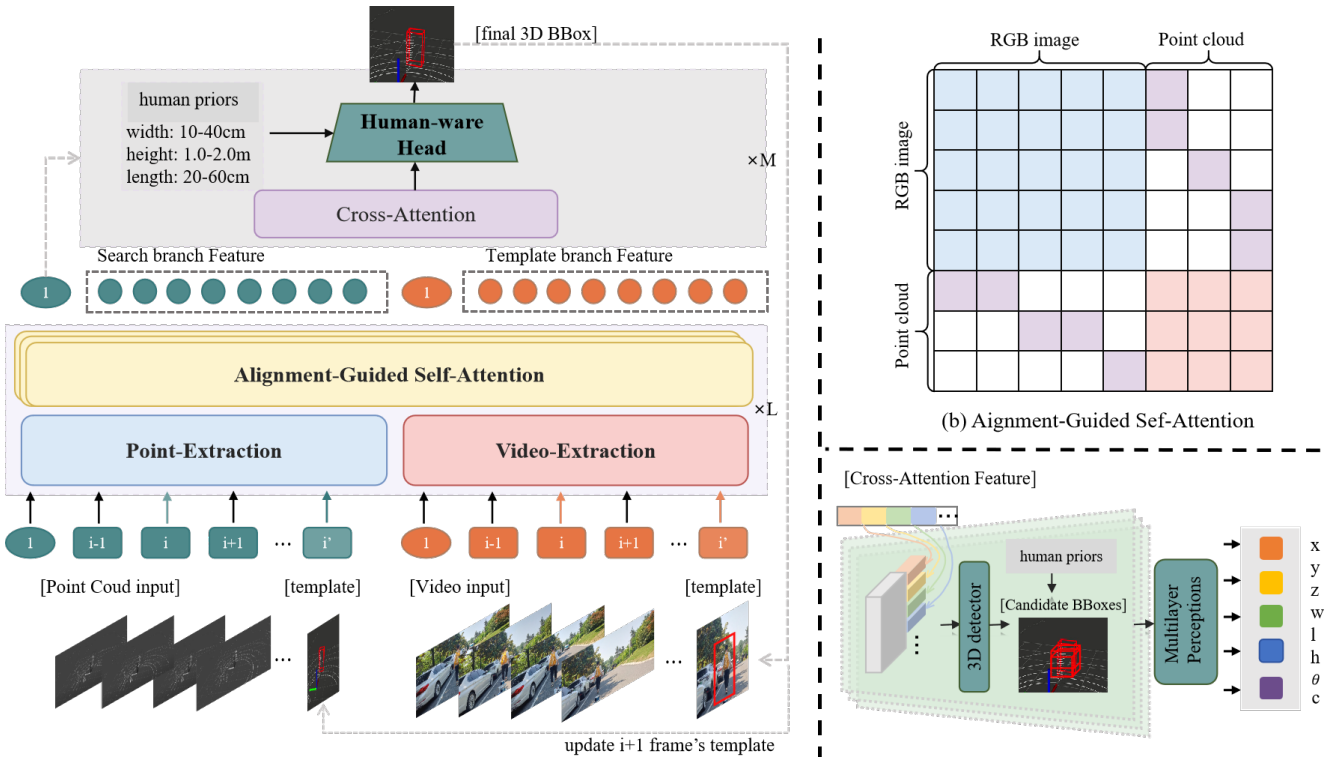

Fig. 2. (a)The MHTrack framework overview. Using N^p point cloud sequences and N^v video frames as input, MHTrack generates multimodal summaries of the key frames. (b) For every video and point pair, an alignment-guided self-attention module is used to align and fuse them. (c) Human-ware Head directly fed the output features of the Cross-Attention decoder for predicting target locations with no additional matcher. Best viewed in color.

adjusts the sample distribution of each module to better match the data distribution according to the real scenarios but also collects more negative samples for constraint during training.

IV. Experiements

In the experimental part, this paper designs two parts, the accuracy evaluation experiment on the open source dataset and the application experiment on the quadruped robot platform.

Experimental environment: All data processing and model training validation were completed based on Fantasy 14. Its hardware configuration is 2G Hz, Intel Core i5, 16GB 3733MHz LPDDR 4X. In the Linux system, the Python version is 3.8, the PyTorch version is 1.4.0, and the CUDA version is 10.0, occupying 4 1080Ti GPU.

Evaluation Protocol. Each scenario is tested through 100 rounds and the results are represented by One Pass Evaluation (OPE) to measure the success rate and precision rate. The definition of success is the Intersection over Union between the 3D predictions and the grounding truth. The definition of precision is the Area Under Curve (AUC) of the distance error between the centers of two boxes within the range of 0 to 2 meters. Furthermore, we applied the Adam optimizer for training. The learning rate was initially 0.002, but after 10 epochs, it decreased fivefold with a batch size of 32.

A. The accuracy evaluation experiment on KITTI

We used KITTI Tracking Datasets (using lidar and camera data) as the experimental dataset, following the data splitting, trajectory generation, and evaluation metrics set in reference [5] for fair comparison. The datasets are divided into a test set, validation set, and training set in a ratio of 3:1:6, in which the test set is used to confirm the model's scalability and generalization capabilities, while the validation set is used to modify the model's hyperparameters.

Table. I gives the comparison results of PVTrack with four classic algorithms for 3D target tracking. The experimental results demonstrate that our method achieves the highest score, in terms of both the success rate (80.22%) and precision rate (80.81%) of 3D human tracking, surpassing existing methods in every category related to people. For the category of *Pedestrians*, it can be seen that under an optimal threshold evaluation standard, the proposed algorithm far exceeds traditional algorithms by more than 20%. For the category of *Cyclists*, the leading gap is slightly lower because *Cyclists* are relatively small in scale and KITTI lacks enough true labels for this target. Besides, our promoted MHTrack in Section III-B surpasses MHTrack in Section III-A by around 6% margin, which proves the effectiveness of the proposed stabilization mechanism to learn useful features even though the evaluation becomes rigorous on datasets.

This paper also focuses on robotic platforms in complex

TABLE I

COMPARISON OF MHTRACK AGAINST STATE-OF-THE-ARTS ON THE KITTI OPEN DATASETS AND ROBOTIC COLLECTION DATASETS.

Datasets		**KITTI open-source datasets**			**Collection dataset from quadruped robot**				
Category		Pedestrian	Cyclist	**Mean***	Flat Outdoor	Flat Indoor	Strong Shaking	Strong Disturbance	**Mean***
Frame Number		*6088*	*308*	*6396*	*1037*	*1059*	*3027*	*739*	*5862*
Success	SC3D [6]	41.29	52.82	47.10	53.82	44.26	27.48	31.11	39.17
	P2B [21]	50.39	62.11	56.25	57.38	55.28	37.19	40.97	47.70
	BAT [34]	51.83	66.28	59.10	63.83	61.38	33.67	38.25	49.28
	PTTR [35]	62.73	70.25	66.49	65.28	67.33	36.33	50.11	54.76
	MHTrack (Ours)	69.38	72.84	71.11	78.26	79.52	58.13	68.13	71.01
	MHTrack-pro(Ours)	**78.17**	**82.27**	**80.22**	**85.51**	**86.29**	**70.28**	**74.25**	**79.80**
	Improvement	↑2.79	↑3.43	↑3.11	↑7.25	↑6.77	↑12.15	↑6.12	↑7.79
Precision	SC3D	20.27	21.35	20.81	57.27	49.25	28.74	32.20	41.87
	P2B	57.82	63.11	60.47	61.74	59.04	33.28	37.19	47.81
	BAT	58.12	64.99	61.57	62.69	66.70	35.72	38.44	50.88
	BAT	61.27	68.39	64.83	67.77	72.32	44.13	47.23	57.86
	MHTrack(Ours)	70.95	74.46	72.70	81.94	81.72	62.67	70.51	74.21
	MHTrack-pro(Ours)	**78.33**	**83.29**	**80.81**	**87.28**	**88.06**	**73.52**	**76.66**	**81.38**
	Improvement	↑2.38	↑3.83	↑3.10	↑5.34	↑6.34	↑10.85	↑6.15	↑7.17

*__Mean__ denotes the average results of counter-categories. The finest and second-best performances are indicated by **Bold** and underline.

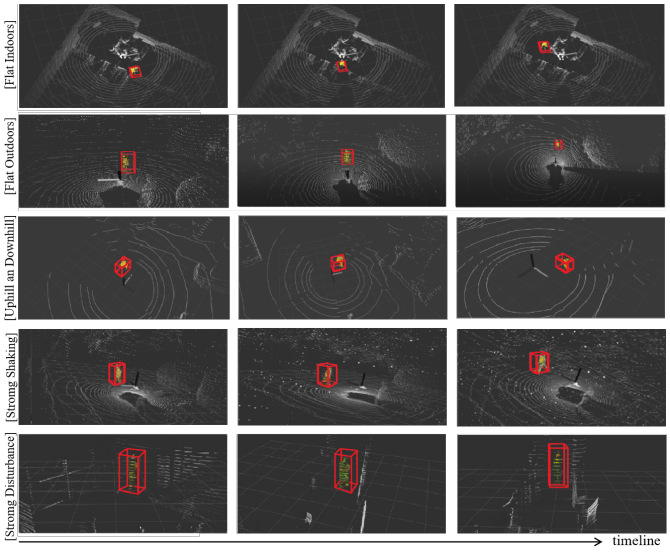

Fig. 3. Visualization results. Our MHTrack can track the ground truth human highlighted in yellow well in different random scenarios by red 3D BBoxes.

environments, using a quadruped robot *Jueying X21* to conduct tracking experiments in more than twenty different indoor and outdoor scenes, including narrow passages, open areas, uphill and downhill slopes, and multi-tree bushes. The examples of scene selection and the tracking effect are shown in Fig. 3. In various scenarios, our proposed MHTrack tracking system can accurately frame the target object, and keep up with the correct object in the form of a 3D BBox within continuous frames, with a duration of more than 20 seconds. In Table. I, we report the extensive comparison results on collection datasets from the jolting quadruped robot.

Our improvement in the anti-shaking performance of the robot platform is significant, achieving 79.80% / 81.38% in terms of success/precision. Regardless of the laser-beam numbers and the terrain changes, our MHTrack-pro surpasses BAT by 20% on average and even more under *Strong Shaking* and *Strong disturbance* scenarios, which confirms its robustness to distractors and appearance changes while showing its tolerance to high-speed moving scenarios.

Ablations. We respectively ablate the point-ware modality input, video-ware modality input, Siamese Transformer Network, and Human-ware Head from the proposed model to understand the components better. Performance deteriorates when any module is removed, even if each module's efficacy varies depending on the dataset. The only exception is the *video-ware modality input* used in the Alignment-Guided Self-Attention Module, which causes a slight drop in KITTI regarding precision. We assume that this happens since KITTI's pedestrian lacks static objects, leading to a biased classifier. Furthermore, the proposed method maintains competitive performance even after module ablation, especially on collection datasets from quadruped robots. This demonstrates the promising possibilities of the robotic-centric paradigm when deployed into dynamic robotic systems.

CONCLUSIONS

In this work, we revisit the 2D and 3D human tracking field and propose a well-designed robotic-centric paradigm, which has been demonstrated to be a great addition to traditional driving scenarios. Particularly, we proposed a novel multi-modal human tracking paradigm MHTrack, and its promoted version. On the one hand, based on multi-modal fusion, the proposed MHTrack can achieve high-precision tracking by framing the target person in the form of a 3D cube by the multi-level Siamese Transformer network and the lateral Human-ware Head. On the other hand, the promoted robotic-centric paradigm with a stabilization mechanism and collaborative training strategy can effectively alleviate the disturbance caused by complex scenes as well as overcome the particularity of the robot platform itself. In the future, we believe that more architecture designs can be guided by the robotic-centric paradigm as a fundamental premise, and it will bring intelligence systems to more high-level applications and more complex scenarios.

REFERENCES

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*, pages 850–865. Springer, 2016.

[2] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[3] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6668–6677, 2020.

[4] Christian Dondrup, Nicola Bellotto, Ferdian Jovan, Marc Hanheide, et al. Real-time multisensor people tracking for human-robot spatial interaction. 2015.

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[6] Silvio Giancola, Jesus Zarzar, and Bernard Ghanem. Leveraging shape completion for 3d siamese tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1359–1368, 2019.

[7] Anfeng He, Chong Luo, Xinmei Tian, and Wenjun Zeng. A twofold siamese network for real-time object tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4834–4843, 2018.

[8] Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14867–14878, 2023.

[9] Le Hui, Lingpeng Wang, Mingmei Cheng, Jin Xie, and Jian Yang. 3d siamese voxel-to-bev tracker for sparse point clouds. *Advances in Neural Information Processing Systems*, 34:28714–28727, 2021.

[10] Le Hui, Lingpeng Wang, Linghua Tang, Kaihao Lan, Jin Xie, and Jian Yang. 3d siamese transformer network for single object tracking on point clouds. In *European Conference on Computer Vision*, pages 293–310. Springer, 2022.

[11] Ugur Kart, Joni-Kristian Kamarainen, and Jiri Matas. How to make an rgbd tracker? In *proceedings of the european conference on computer vision (ECCV) Workshops*, pages 0–0, 2018.

[12] Aleksandr Kim, Aljoša Ošep, and Laura Leal-Taixé. Eagermot: 3d multi-object tracking via sensor fusion. In *2021 IEEE International conference on Robotics and Automation (ICRA)*, pages 11315–11321. IEEE, 2021.

[13] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4282–4291, 2019.

[14] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8971–8980, 2018.

[15] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8971–8980, 2018.

[16] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022.

[17] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*, pages 604–621. Springer, 2022.

[18] AJ Piergiovanni, Vincent Casser, Michael S Ryoo, and Anelia Angelova. 4d-net for learned multi-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15435–15445, 2021.

[19] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019.

[20] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[21] Haozhe Qi, Chen Feng, Zhiguo Cao, Feng Zhao, and Yang Xiao. P2b: Point-to-box network for 3d object tracking in point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6329–6338, 2020.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[23] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019.

[24] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

[25] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8032–8041, 2023.

[26] Torsten Wilhelm, Hans-Joachim Böhme, and Horst-Michael Gross. Sensor fusion for vision and sonar based people tracking on a mobile service robot. In *Proceedings of the International Workshop on Dynamic Perception*, pages 315–320, 2002.

[27] Fei Xie, Chunyu Wang, Guangting Wang, Yue Cao, Wankou Yang, and Wenjun Zeng. Correlation-aware deep tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8751–8760, 2022.

[28] Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12549–12556, 2020.

[29] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European Conference on Computer Vision*, pages 341–357. Springer, 2022.

[30] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

[31] Jin Hyeok Yoo, Yecheol Kim, Jisong Kim, and Jun Won Choi. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 720–736. Springer, 2020.

[32] Jesus Zarzar, Silvio Giancola, and Bernard Ghanem. Efficient tracking proposals using 2d-3d siamese networks on lidar. 2019.

[33] Zhipeng Zhang, Yihao Liu, Xiao Wang, Bing Li, and Weiming Hu. Learn to match: Automatic matching network design for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13339–13348, 2021.

[34] Chaoda Zheng, Xu Yan, Jiantao Gao, Weibing Zhao, Wei Zhang, Zhen Li, and Shuguang Cui. Box-aware feature enhancement for single object tracking on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13199–13208, 2021.

[35] Changqing Zhou, Zhipeng Luo, Yueru Luo, Tianrui Liu, Liang Pan, Zhongang Cai, Haiyu Zhao, and Shijian Lu. Pttr: Relational 3d point cloud object tracking with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8531–8540, 2022.

[36] Hao Zou, Jinhao Cui, Xin Kong, Chujuan Zhang, Yong Liu, Feng Wen, and Wanlong Li. F-siamese tracker: A frustum-based double siamese network for 3d single object tracking. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8133–8139. IEEE, 2020.