TGRMPT: A Head-Shoulder Aided Multi-Person Tracker and a New Large-Scale Dataset for Tour-Guide Robot

Wen Wang, Zheyuan Lin, Min Huang, Shunda Hu, Tianlei Jin, Te Li, Minhong Wan, Jason Gu, Shiqiang Zhu*

Abstract—A service robot serving safely and politely needs to track the surrounding people robustly, especially for Tour-Guide Robot (TGR). However, existing multi-object tracking (MOT) or multi-person tracking (MPT) methods are not applicable to TGR for the following reasons: 1. lacking relevant large-scale datasets; 2. lacking applicable metrics to evaluate trackers. In this work, we target the visual perceptual tasks for TGR and present the TGRDB dataset, a novel large-scale multi-person tracking dataset containing roughly 5.6 hours of annotated videos and over 450 long-term trajectories. Besides, we propose a more applicable metric to evaluate trackers using our dataset. As part of our work, we present TGRMPT, a novel MPT system that incorporates information from head-shoulder and whole body, and achieves state-of-the-art performance. We have released our codes and dataset in https://github.com/wenwenzju/TGRMPT.

I. INTRODUCTION

Robust tracking of surrounding people allow service robots to serve and navigate safely and politely. In this work, we aim at the egocentric perceptual task, multi-person tracking, using a tour-guide robot (TGR). The TGR is a robot that can show people around a site such as museums, castles, and aquariums [1]. In this scenario, targets occlude with each other and step in or out of the camera view frequently. Moreover, humans in a tour group usually dress in the same clothes, making it a difficult task to identify each other. Despite the above challenges, individuals must be tracked and assigned consistent IDs during the whole tour-guide service period.

However, there is a vast gap between the existing MOT datasets and the tour-guide scenario, making existing trackers perform unsatisfactorily. Some examples of such datasets are the well-known MOTChallenge [2] and KITTI [3], targeting applications in video surveillance and autonomous driving, respectively. Furthermore, in these datasets, a person will be assigned a new ID if he/she leaves the camera or is occluded for a prolonged period and then reappears. This is very different from our scenario where consistent IDs are required during the whole tracking period. More recently, an egocentric dataset, JRDB, was presented in [4], which is more relevant to our work. However, this dataset is still inapplicable in our scenario. On the one hand, JRDB was captured in daily life where few interactions between students and robot exist. Thus, cases that humans disappear then reappear again are few. On the other hand, people in JRDB dress casually and almost no one wears the same.

To solve the above problems, we present the TGRDB dataset, a new large-scale dataset for TGR. To capture our dataset, we use a 180° fisheye RGB camera mounted on a moving (sometimes standing) robot shown in Fig. 1. In an indoor tour-guide scenario, the robot shows 5 or 6 individuals around. At first, participants wear their own clothes and complete the first capturing round. Then they are required to change and put on the same clothes, as shown in Fig. 1. More details and statistical analysis are described in III-A. We hope this dataset will drive the progress of research in service robotics, long-term multi-person tracking, and fine-grained person re-identification. Associated with this dataset, a new metric named TGRHOTA is proposed. It is a simplified but more practical version of HOT A [7]. TGRHOTA only considers matches between prediction and ground truth as true positives at the most details and statistical analysis are described in III-A. We hope this dataset will drive the progress of research in service robotics, long-term multi-person tracking, and fine-grained person re-identification. Associated with this dataset, a new metric named TGRHOTA is proposed. It is a simplified but more practical version of HOT A [7]. TGRHOTA only considers matches between prediction and ground truth as true positives at the most details and statistical analysis are described in III-A. We hope this dataset will drive the progress of research in service robotics, long-term multi-person tracking, and fine-grained person re-identification. Associated with this dataset, a new metric named TGRHOTA is proposed. It is a simplified but more practical version of HOT A [7]. TGRHOTA only considers matches between prediction and ground truth as true positives at the most
simply concatenate the two embedding features extracted from the corresponding bounding boxes, resulting in stronger appearance descriptors. Extensive experiments verify the effectiveness of our proposal.

To summarize, our contributions are as follows:

- We release the TGRDB dataset, a first large-scale dataset towards the applications of the tour-guide robot. This dataset not only benefits the domain of service robotics but can also drive the progress of domains related to multi-person tracking and person re-identification.
- We propose a more practical metric, TGRHOTA, to evaluate trackers in the tour-guide scenario. Different from existing metrics, TGRHOTA punishes trackers when a new ID assignment occurs if the target has already been assigned an ID before.
- We propose a novel head-shoulder aided multi-person tracker, named TGRMPT, that leverages best of both information containing in the whole body and head shoulder. Experiments show that our tracker achieves start-of-the-art results.

II. RELATED WORK

A. Tracking Methods

Most MOT algorithms employ tracking-by-detection paradigm [10]–[12], which split tracking task into object detection, embedding feature extraction and data association, while others propose end-to-end models to unify these three isolated sub-tasks [13]–[15]. For more comprehensive reviews on MOT, readers are referred to [16]. This paper mainly focuses on MOT methods in robotic applications.

Tracking surrounding people is one of the primary tasks for robots. To achieve this, methods based on the single sensor like a 2D laser scanner [17], or multi-modal sensors [18]–[21], are proposed. To detect, track and follow people, [17] presents a leg tracker integrating with local occupancy grid maps. Compared to single-sensor-based trackers, multi-sensor-based methods are more popular. To track people around a service robot, [18] and [20] propose to aggregate detections of multiple detectors from color/depth cameras and 2D laser scanners. As for autonomous driving, [19] and [21] present methods that integrate information from 3D laser and color camera to form strong appearance representations. Different from these works, a single fisheye camera is the only sensor that our method employs to track people.

B. Tracking Datasets

Our TGRDB is a multi-person tracking dataset and benchmark containing fine-grained and clothing-inconsistency targets in the tour-guide scenario. [2], [4], [5], [22], [23] are the most relevant datasets to TGRDB. MOT [2], [22] is a well-known multi-object tracking benchmark, and many methods are based on this. JRDB [4] is a novel multi-modal dataset collected from a mobile social JackRabbit. [23] is an image dataset for clothing-change person REID. [5] released a fine-grained REID dataset containing targets with the same clothes. In addition, the automatic driving datasets [3], [24], [25] are also interrelated to our TGRDB, which annotate a large number of pedestrians bounding boxes. To the best of our knowledge, there is no tracking dataset containing fine-grained and clothing-inconsistency targets on the same scale as our TGRDB in the tour-guide scenario.

C. Metrics

Within the last few years, new MOT metrics have been proposed enormously. To remedy the lack of generally applicable metrics in MOT, [26] introduces two novel metrics, the multi-object tracking precision (MOTP) and the multi-object tracking accuracy (MOTA), that intuitively express a tracker’s overall strengths. [27] proposes a new precision-recall measure of performance, IDF1, that treats errors of all types uniformly and emphasizes correct identification over sources of errors. MOTA and IDF1 overemphasize the importance of detection and association separately. To address this, [7] presents a novel MOT evaluation metric, higher-order tracking accuracy (HOTA), which is a unified metric that explicitly balances the effect of detection, association, and localization. Nonetheless, none of the above metrics is applicable in our scenario. Hence we propose a new metric, TGRHOTA, for fair comparison of trackers in TGR.

III. DATASET AND METRIC

A. Dataset

Our TGRDB dataset was collected with a 180° fisheye RGB camera on-board of a tour-guide robot shown in Fig. 1. Equipped with a 360° 2D LiDAR, our TGR can autonomously navigate and show people around. There is a pan-tilt with two degrees of freedom mounted at the neck, so the robot can rotate its head to see what it’s interested in.

Fig. 2: The tour-guide route to capture our TGRDB dataset. There are 4 exhibit points where the robot stays and introduces exhibits to participants. Pictures are captured at these points respectively.

TGRDB dataset was collected in an indoor tour-guide scenario. From the start point, the robot navigates to the first preset exhibit point, stays for a while and introduces the surrounding exhibits to participants, then moves to the next exhibit point, as shown in Fig. 2. Repeat the above steps until it returns to the start point. There are totally four exhibit points and the robot stays for around 22 seconds at each point. During the introduction period, the head of the robot alternatively looks forward or rotates randomly. Participants
Table I: Statistical comparisons between TGRDB and existing datasets. K = Thousand, M = Million, min = minutes.

| Dataset | No. of sequences | No. of frames | No. of boxes | No. of IDs | Duration | Clothes |
|---------|-----------------|---------------|--------------|------------|----------|---------|
| MOT17 [2] | 14 | 34K | 290K | - | - | Casual only |
| MOT20 [22] | 8 | 13K | 1.6M | - | - | Casual only |
| KITTI [13] | 22 | 175K | 6K | - | - | Casual only |
| BDD100K [24] | 1.6K | 318K | 44K | - | - | Casual only |
| Waymo [25] | 1.15K | 600K | 2.1M | - | - | Casual only |
| JRDB [4] | 54 | 28K | 2.4M | - | 64min | Casual only |
| Ours(Train) | 50 | 281K | 1.8M | 51 | 188min | Casual & Same |
| Ours(Test) | 40 | 232K | 1.4M | 40 | 150min | Casual & Same |
| Ours(Total) | 90 | 513K | 3.2M | 91 | 338min | Casual & Same |

B. Metric

An applicable metric is significant when comparing different trackers. Existing metrics, such as HOTA [7], MOTA [26] and IDF1 [27], treat all matched prediction (pr) and ground-truth (gt) pairs as true positives (TPs), which may be reasonable in video surveillance or autonomous driving. However, they are not applicable in TGR for the requirement of consistent IDs during the whole tour-guide period. In retrospect, HOTA [7] counts TPs (True Positive Associations), FNs (False Negative Associations), and FPs (False Positive Associations) for each TP and defines association score as follows:

$$\text{ASS}_A = \frac{1}{|TP|} \sum_{c \in TP} |TPA(c)| + |FNA(c)| + |FPA(c)|$$

where all matched pr and gt pairs are considered as TP of interest. In other words, as shown in Fig. 3, TP of interest is selected in matches within all dashed areas. However, in TGR, we only consider TPs in the green area, i.e., matches at most previous frames. Formally, we define the set of TPs in TGR as follows:

$$TP' = \{tp_t \in TP | t < t', tp_t \equiv tp_{t'} \lor tp_{t'} \not\equiv tp_t\}$$

where $t$ is the time index, and $\equiv$ denotes that two TPs have the same prID and gtID, while $\not\equiv$ denotes two TPs have different prID and gtID. Thus, the association score in TGR can be written as:

$$\text{ASS}_A' = \frac{1}{|TP'|} \sum_{c \in TP'} |TPA(c)| + |FNA(c)| + |FPA(c)|$$

The detection score, DetA, and the calculation of the final HOTA score, is the same as in [7], except that we use ASSA' instead of ASSA.

Fig. 4 shows an example where HOT A gives the two trackers the same scores which is counterproductive for our TGR scenario where tracker B performs much better than tracker A. Treating all TPs equally even if they have been assigned another ID at a previous frame, HOT A is not applicable in our scenario. On the contrary, our TGRHOTA gives much more reasonable scores.

IV. TGRMPT: HEAD-SHOULDER AIDED MULTI-PERSON TRACKING

Fig. 5 depicts our proposed TGRMPT tracker. By sequentially applied whole body (wb)/head-shoulder (hs) detectors and wb/hs feature extractors, deep appearance signatures describing global whole-body and local head-shoulder are generated. Concatenating these two types of features results in the final strong descriptor that contains both global and local information. The followed Hungarian algorithm produces the final robust tracking result.

A. Detector

Detection is the core component of the existing MOT or MPT systems. In consider of system speed, we deploy YOLOv5s [28] as our detector, wb/hs detection network is fine-tuned on our dataset using wb/hs annotations. The corresponding outputs are denoted as $D^{wb}$ and $D^{hs}$, respectively. Defining IoU between $d^{wb}$ and

$$\text{IoU} = \frac{|d^{wb} \cap d^{hs}|}{|d^{wb} \cup d^{hs}|}$$

as the overlap between $d^{wb}$ and $d^{hs}$, the TGRHOTA takes the combination of scores $\text{Det}^{wb} = \text{det}^{wb}(d^{wb})$ and $\text{Det}^{hs} = \text{det}^{hs}(d^{hs})$. The following algorithm is the foundation of our TGRHOTA.
form the final appearance descriptors denoted as \( wb \) match detection computing the pairwise distance between current detections \( S \) and latest them to trajectories at previous frames. To do so, we keep the one trajectory \( D \) to produce the final distance, \( \tau \) values, \( \tau \) threshold. Any association that has a distance greater than the preset \( j \) is treated as unmatched.

\[
d^{wb}_{i} = \frac{|d^{wb}_{1} \cap d^{wb}_{i}|}{|d^{wb}_{1}|}
\]

we use Hungarian algorithm to match \( wb \) and \( hs \) detections, resulting in matched pairs denoted as \( D^{match} = \{(d^{wb}_{1}, d^{wb}_{2}, \ldots, d^{wb}_{j}, d^{wb}_{k})\} \). We denote those \( wb \) detections that are not matched to any \( hs \) detections as \( D^{wb-} = \{d^{wb}_{1}, \ldots, d^{wb-}_{j} \} \), and we discard those \( hs \) detections that are not matched to any \( wb \) detections.

### B. Appearance Descriptor

Appearance descriptors are used to measure the similarities between detections at current frame and history trajectories. Due to the trade-off between speed and performance, we leverage ResNet18 [29] as our feature extracting network. Due to the trade-off between speed and performance, we leverage ResNet18 [29] as our feature extracting network. We explore different values of hyperparameters, \( \alpha \) and \( \tau \), which may impact the final performance dramatically. We fix the distance threshold \( \tau \) and set it to 0.5 at first, and change the age threshold \( \alpha \) to explore the impact. In this stage, we employ the whole body branch only. The results are shown in Table II. Intuitively, \( \alpha \) impacts more on data association. We can see this from ASSA and IDF1 scores. In the TGR scenario, it is required that people's IDs remain consistent even after long-term tracking. Hungarian algorithm is employed to get the final association results and trajectories are updated using the corresponding associated detections. Any association that has a distance greater than the preset threshold \( \tau \) is treated as unmatched.

\[
s_{j} = \frac{1}{|F^{fused}_{i}|} \sum_{f \in F^{fused}_{i}} s(f, j)
\]

The origin DeepSORT is designed for video surveillance and is not well fitted to our tour-guide scenario. In the data association process of DeepSORT, matching cascade is introduced to give higher priority to trajectories that miss targets for less time. Meanwhile, a so-called gating mechanism is applied and only detections near the predicted locations of trajectories are considered. These two tricks may cause more track fragments and worsen the tracking performance in our scenario as frequent occlusion and long-term missing exist in our dataset. Consequently, we abandon these two tricks in our tracker. Another trick to handle challenges in TGR is to assign a relatively great value to the age threshold, \( \alpha \). When a trajectory misses for consecutive \( \alpha \) frames, it will be deleted.

### V. Experiments

To evaluate on our TGRDB dataset, the widely used metrics, HOTA [7], MOTA [26] and IDF1 [27] are adopted. Besides, we report results evaluated using the proposed TGRHOTA metric as well. For a more comprehensive analysis, we split the test dataset into casual-clothes and same-clothes sub-datasets, and conduct experiments respectively. But we don’t split the training dataset.

#### A. Ablation Studies

1) Hyper-Parameters: Different values of hyperparameters, \( \alpha \) and \( \tau \), may impact the final performance dramatically. We fix the distance threshold \( \tau \) and set it to 0.5 at first, and change the age threshold \( \alpha \) to explore the impact. In this stage, we employ the whole body branch only. The results are shown in Table II. Intuitively, \( \alpha \) impacts more on data association. We can see this from ASSA and IDF1 scores. In the TGR scenario, it is required that people’s IDs remain consistent even after long-term tracking. Thus the greater \( \alpha \) should give the better performance. This is verified by the results in Table II, where the infinite \( \alpha \) gives the best performance. This means that during the period of tour-guide, no tracks will be deleted, which gives chance to the tracker to find back the target even if he/she has disappeared for a long time.

We then fix \( \alpha = \infty \) and change the distance threshold \( \tau \) to explore the impact. In data association, there are two ways,
TABLE II: The impact of age threshold $\alpha$ when distance threshold $\tau = 0.5$.

|                  | (a) Casual Clothes | (b) Same Clothes |
|------------------|--------------------|------------------|
|                  | $\alpha$          |                  |
|                  | 30                 | 100              | 500              | $\infty$ |
| HOTA$^\uparrow$  | 33.6               | 36.0             | 37.8             | 41.2     |
| ASSA$^\downarrow$| 0.15               | 0.13             | 0.12             | 0.11     |
| MOA$^\uparrow$   | 86.2               | 86.3             | 86.0             | 85.5     |
| IDF1$^\uparrow$  | 27.4               | 29.6             | 31.2             | 30.4     |

Fig. 6: The impact of different distance threshold $\tau$ values using the two matching distance calculation ways.

(mean or min) as described in IV-C, to compute the distance between detection and trajectory. We show their results$^1$ in Fig. 6, where we conclude that the averaged distance value gives much better performance. The reason is apparent: the reserved $P$ galleries in each trajectory may contain noisy samples and the minimum distance value suffers much from these noises.

The above experiments on hyper-parameters indicate that $\alpha = \infty$ and $\tau = 0.85$ give the best performance. We will fix them in the following experiments and explore other aspects of our method.

2) Addition of Head-Shoulder: Head-shoulder contains non-negligible features, such as haircut, glasses, complexion, etc., which can provide discriminative cues for person re-identification. We conduct comparative experiments on the employment of: 1) $wb$, the whole body only; 2) $hs$, the head shoulder only; 3) $wb+hs$, both whole body and head shoulder. The results are shown in Table III. Undoubtedly, the method that integrates both $wb$ and $hs$ information gives the best performance.

TABLE III: The improvement of performance after adding head-shoulder.

|                  | (a) Casual Clothes | (b) Same Clothes |
|------------------|--------------------|------------------|
|                  | $wb$               | $hs$            | $wb+hs$         | $wb$          | $hs$          | $wb+hs$        |
| HOTA$^\uparrow$  | 71.2               | 53.0            | 72.7            | 67.2          | 51.9          | 68.7           |
| MOA$^\uparrow$   | 87.1               | 54.1            | 87.2            | 87.6          | 56.2          | 87.6           |
| IDF1$^\uparrow$  | 51.2               | 58.8            | 64.2            | 57.2          | 56.3          | 77.9           |

To dig deeper into how head shoulder helps improve the overall performance, we show precision and recall scores of association in Fig. 7. Compared to the $wb$-only method, for casual-clothes sub-dataset, the percentage promotion of precision and recall after involving head shoulder is 2.3% and 2.9%, respectively. Thus, the head shoulder contributes equally to precision and recall. Nevertheless, for same-clothes sub-dataset, the respective promotion is 5.9% and 1.4%. Head shoulder contributes much more to precision. In other words, with the help of head shoulder, the tracker does better in identifying different people even if they wear the same. This is due to the discriminative details contained in head shoulder which are untraceable in whole body.

3) New Metric: We show scores of our proposed TGRHOTA in Table IV. We can see that TGRHOTA scores are higher than HOTA scores. This is a similar situation as tracker B shown in Fig. 4. That is to say, once a target was assigned an ID, our tracker can keep it consistent most of the time.

TABLE IV: The evaluation result of new TGRHOTA metric.

|                  | (a) Casual Clothes | (b) Same Clothes |
|------------------|--------------------|------------------|
|                  | $wb$               | $hs$            | $wb+hs$         | $wb$          | $hs$          | $wb+hs$        |
| TGRHOTA$^\uparrow$ | 74.3               | 55.5            | 75.2            | 70.1          | 54.9          | 72.1           |

4) FPS: On a single thread, our method runs at 16fps end-to-end, using an Nvidia 2080Ti GPU card. Parallel running of $wb$ and $hs$ detectors may speed up the method.

B. Comparison with State-of-the-Art

We compare our proposal with two newest state-of-the-art methods, CenterTrack [13] and ByteTrack [30]. Results are shown in Table V. We train CenterTrack and ByteTrack using our dataset and fine-tune some hyper-parameters. Our method achieves the best performance. We show comparison results for reference only, as we have carefully tailored DeepSORT to fit the tour-guide scenario, especially for data association. Designed for video surveillance, it is foreseeable that CenterTrack and ByteTrack perform poorly without any scenario-oriented modifications.

---

$^1$For simplicity, only results on same-clothes sub-dataset are shown as casual-clothes sub-dataset has similar results.
TABLE V: Comparison with State-of-the-Art.

| Method      | Casual Clothes | Same Clothes |
|-------------|----------------|--------------|
| CenterTrack | 88.1 25.1 32.0 | 39.3 37.7 22.0 29.8 35.5 |
| ByteTrack   | 87.7 26.8 33.1 | 39.7 37.3 23.7 30.8 34.5 |
| Ours        | 87.2 84.2 72.7 | 75.2 87.6 77.9 68.7 72.1 |

VI. CONCLUSION

We release the TGRDB dataset, the first large-scale dataset for the applications of TGR. It was captured using a TGR in an indoor tour-guide scenario. We annotate the dataset with whole-body and head-shoulder bounding boxes, as well as unique IDs for each participant. We believe that TGRDB will help future research in service robotics, long-term multi-person tracking, and find-grained or clothing-inconsistency person re-identification. Along with the dataset, we propose a more practical metric, TGRHOTA, to evaluate trackers in the tour-guide scenario. As part of our work, we propose TGRMPT, a novel head-shoulder aided multi-person tracking system that leverages best of discriminative cues contained in head shoulder which are untraceable in whole body. Extensive experiments have confirmed the significant advantages of our proposal.

ACKNOWLEDGMENT

This work is supported by Ten Thousand Talents Program of Zhejiang Province (Grant No. 2019RS1010), and the Key Research Project of Zhejiang Lab (Grant No. G2021NB0AL03). We would also like to thank Fellow Academicians Jason Gu for leading the Tour-Guide Robot project.

REFERENCES

[1] A. Al-Wazzan, R. Al-Farhan, F. Al-Ali, and M. El-Abd, “Tour-guide robot,” in 2016 International Conference on Industrial Informatics and Computer Systems (CICS). IEEE, 2016, pp. 1–5.

[2] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, “Mot16: A benchmark for multi-object tracking,” arXiv preprint arXiv:1603.00831, 2016.

[3] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012, pp. 3354–3361.

[4] R. Martin-Martin, M. Patel, H. Rezatofighi, A. Shenoi, J. Gwak, E. Frankel, A. Sadeghian, and S. Savarese, “Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[5] J. Yin, A. Wu, and W.-S. Zheng, “Fine-grained person re-identification,” International journal of computer vision, vol. 128, no. 6, pp. 1654–1672, 2020.

[6] F. Wang, Y. Wu, X. Qian, Y. Chen, and Y. Fu, “When person re-identification meets changing clothes,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 830–831.

[7] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe, “Hota: A higher order metric for evaluating multi-object tracking,” International journal of computer vision, vol. 129, no. 2, pp. 548–578, 2021.

[8] B. Xu, L. He, X. Liao, W. Liu, Z. Sun, and T. Mei, “Black re-id: A head-shoulder descriptor for the challenging problem of person re-identification,” in Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 673–681.

[9] N. Wojke, A. Bewley, and D. Paulus, “Simple online and realtime tracking with a deep association metric,” in 2017 IEEE international conference on image processing (ICIP). IEEE, 2017, pp. 3643–3649.

[10] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple online and realtime tracking,” in 2016 IEEE international conference on image processing (ICIP). IEEE, 2016, pp. 3464–3468.

[11] J. Xu, Y. Cao, Z. Zhang, and H. Hu, “Spatial-temporal relation networks for multi-object tracking,” in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 3988–3998.

[12] Y. Xu, Y. Ban, G. Delorme, C. Gan, D. Rus, and X. Alameda-Pineda, “Transcenter: Transformers with dense queries for multiple-object tracking,” arXiv preprint arXiv:2103.15143, 2021.

[13] X. Zhou, V. Koltun, and P. Krähenbühl, “Tracking objects as points,” in European Conference on Computer Vision. Springer, 2020, pp. 474–490.

[14] P. Tokmakov, J. Li, W. Burgard, and A. Gordon, “Learning to track with object permanence,” in Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2019, pp. 10860–10869.

[15] J. Pang, L. Qiu, X. Li, H. Chen, Q. Li, T. Darrell, and F. Yu, “Quasi-dense similarity learning for multiple object tracking,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 164–173.

[16] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, and T.-K. Kim, “Multiple object tracking: A literature review,” Artificial Intelligence, vol. 293, p. 103448, 2021.

[17] A. Leigh, J. Pineau, N. Olmedo, and H. Zhang, “Person tracking and following with 2d laser scanners,” in 2015 IEEE international conference on robotics and automation (ICRA). IEEE, 2015, pp. 726–733.

[18] T. Linder, S. Breuers, B. Leibe, and K. O. Arras, “On multi-modal people tracking from mobile platforms in very crowded and dynamic environments,” in 2016 IEEE international conference on robotics and automation (ICRA). IEEE, 2016, pp. 5512–5519.

[19] A. Shenoi, M. Patel, J. Gwak, P. Goebel, A. Sadeghian, H. Rezatofighi, R. Martin-Martin, and S. Savarese, “Jrmt: A real-time 3d multi-object tracker and a new large-scale dataset,” in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 10 335–10 342.

[20] T. Wengefeld, S. Müller, B. Lewandowski, and H.-M. Gross, “A multi modal people tracker for real time human robot interaction,” in 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). IEEE, 2019, pp. 1422–1423.

[21] H.-k. Chiu, J. Li, R. Ambrus, and J. Bobg, “Probabilistic 3d multi-modal, multi-object tracking for autonomous driving,” in 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021, pp. 14227–14233.

[22] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé, “Mot20: A benchmark for multi object tracking in crowded scenes,” arXiv preprint arXiv:2002.09003, 2020.

[23] Q. Yang, A. Wu, and W.-S. Zheng, “Person re-identification by contour sketch under moderate clothing change,” IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 6, pp. 2029–2046, 2019.

[24] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, “Bdd100k: A diverse driving video dataset with scalable annotation tooling,” arXiv preprint arXiv:1805.04687, vol. 2, no. 5, p. 6, 2018.

[25] P. Sun, H. Kretzschmar, X. Dotzilla, A. Choudar, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. 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, 2020, pp. 2446–2454.

[26] K. Bernardin and R. Stiefelhagen, “Evaluating multiple object tracking performance: the clear mot metrics,” EURASIP Journal on Image and Video Processing, vol. 2008, pp. 1–10, 2008.

[27] E. Ristani, F. Soliera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” in European conference on computer vision. Springer, 2016, pp. 17–35.

[28] G. J. et. al., “ultralytics/yolov5: v6.0 YOLOv5n ‘Nano’ models, Roboforw integration, Tensorflow export, OpenCV DNN support,” Oct. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5563715

[29] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[30] Y. Zhang, P. Sun, Y. Jiang, D. Yu, Z. Yuan, P. Luo, W. Liu, and X. Wang, “Byettrack: Multi-object tracking by associating every detection box,” arXiv preprint arXiv:2110.06664, 2021.