FRoG-MOT: Fast and Robust Generic Multiple-Object Tracking by IoU and Motion-State Associations

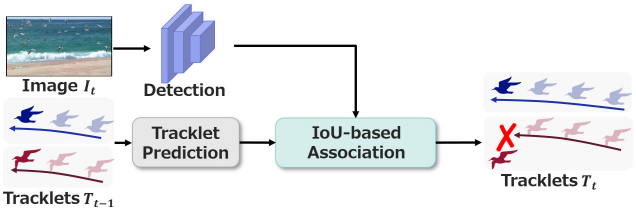
Takuya Ogawa[1] Takashi Shibata[1] Toshinori Hosoi[1]
[1]NEC Corporation
takuya_ogawa@nec.com t.shibata@ieee.org t.hosoi@nec.com

Abstract

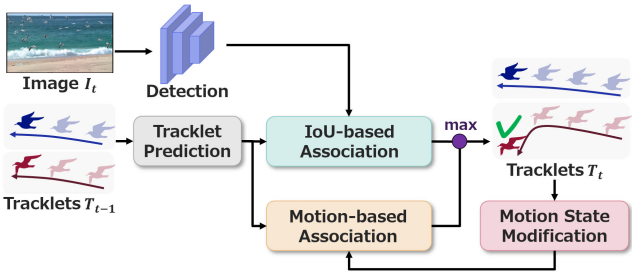
This paper proposes a generic multi-object tracking (MOT) algorithm that is robust to unexpected motion changes for generic objects. Deep learning has dramatically been improving MOT performances. Nevertheless, state-of-the-art tracking algorithms are still sensitive to unexpected motion changes and the generic object target beyond person tracking. This is because standard MOT benchmark datasets such as MOT17 mainly consist of persons in a crowd, often lacking unexpected shape and motion changes; thus, these issues have yet to be focused on. We propose a simple-yet-effective MOT framework that can dynamically improve tracking continuity by associating each target based on adaptively modified motion states. The keys are 1) to represent the target motions using multiple motion states that have weak correlations with each other and 2) to modify those states that have the lowest similarity to past states as outliers. Our approach can improve trajectory continuity and robustness to unexpected motion changes for generic objects. Comprehensive experiments have confirmed that our framework is comparable to existing state-of-the-art methods on a standard dataset and outperforms those algorithms on the GMOT dataset with an overall 2% improvement in IDF1, a measure of tracking continuity.

1. Introduction

Multiple object tracking (MOT) is one of the main streams of computer vision tasks. Various algorithms [5, 43, 52, 53], large-scale datasets [2, 6, 33, 38, 39, 45], and open sources [9, 48] have led to significant performance improvements. Despite this spectacular success, even state-of-the-art methods are sensitive to unexpected motion changes in multiple generic objects, such as crowds of animals. Tracking moving objects for generic objects has various applications, such as animal behavior [31, 36, 37] and video surveillance [4, 25]. The trajectory of these moving generic objects often contains unexpected and sudden motion changes.



(a) Existing tracking-by-detection (only IoU-based association)



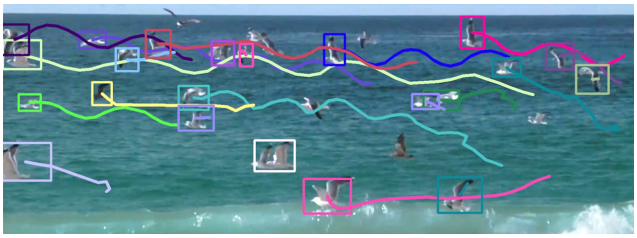
(b) Our framework (IoU-based and motion-based association)

Figure 1. Overviews of the existing tracking-by-detection and our proposed framework. Our framework can improve tracking performance for generic objects by introducing motion-based association and motion-state modification for each target.

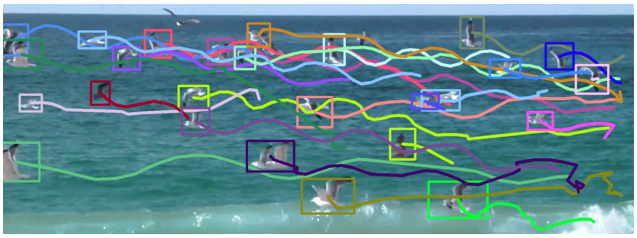
MOT algorithms are usually designed based on the implicit assumption that the shape and motion changes of the tracked object are small. This implicit assumption holds as long as the tracked target is a person in the crowd, as in standard MOT datasets such as MOT17 [33]. In particular, since the trajectory of a person is not significantly disrupted in a crowd, large motion, and unexpected shape changes are restricted. When those motion and shape changes are small, the motions and shapes of each target are easily predictable if these target objects are detected. Therefore, the tracking-by-detection approach [5, 43, 47, 53] that directly uses the strengths of modern object detection algorithms is effective, and various sophisticated tracking-by-detection algorithms (e.g., ByteTrack [52]) have been provided.

The object's shape and motion unexpectedly change in each frame when focusing on multiple generic object tracking, as we will discuss in the analysis of Sec. 3. As a re-

(a) Ground truth



(b) Result by ByteTrack [52]



(c) Result by proposed method

Figure 2. Examples of ByteTrack [52] and our proposed method. While ByteTrack [52] have failed to track many birds as in (b), our proposed method can continuously track birds as in (c).

sult, performance, such as tracking continuity, is significantly degraded in existing tracking algorithms because this implicit assumption does not hold. Figure 2 (b) shows an example of this limitation. The state-of-art tracking algorithm [52] fails to track many birds. There is a critical demand for a new framework that can break this limitation while inheriting the strengths of existing MOT algorithms.

We propose a **F**ast and **Ro**bust **G**eneric **M**ultiple-**O**bject **T**racking (**FRoG-MOT**) that can improve tracking continuity by explicitly associating each target using adaptively modified motion states. The keys are 1) to represent the target motions using multiple motion states that have weak correlations with each other and 2) to modify those states that have the lowest similarity to past states as outliers, as shown in Fig. 1 (b). This simple-yet-effective approach improves trajectory continuity and robustness to unexpected changes in motion and shape, even for generic objects. Indeed, as shown in Fig. 2 (c), the proposed framework can continuously track multiple birds by explicitly using the motion state. Extensive experiments show that our framework outperforms those algorithms in GMOTs involving various objects. Furthermore, on a standard dataset such as MOT17, we show that the performance is comparable to or even better than existing state-of-the-art methods, even though our framework can achieve real-time processing.

In summary, our contributions are: i) We propose a generic MOT algorithm, FRoG-MOT, that can improve tracking continuity by associating each target with adaptively modified motion states. ii) We analyze the standard MOT dataset and generic MOT dataset and reveal challenges to be solved for the generic object tracking task. iii) Comprehensive experiments on MOT17 [33] and GMOT-40 [2] demonstrate the effectiveness of the proposed framework. We prepare GMOT-Split101 based on GMOT-40, which is an evaluation dataset for generic object tracking.

2. Related Works

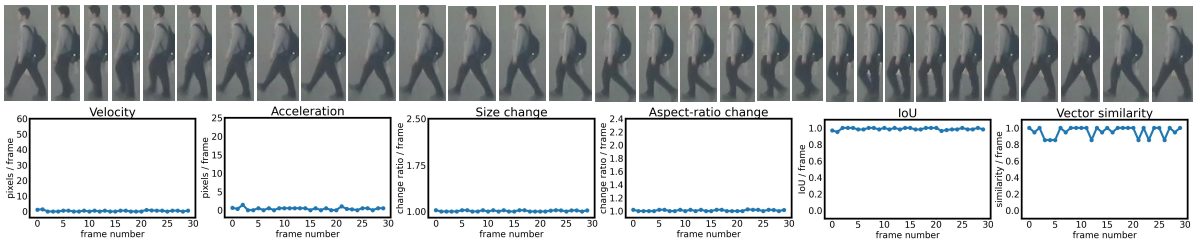
2.1. Multi-Object Tracking

Recent MOT algorithms can be classified into 1) the tracking-by-detection approach and 2) the transformer-based approach.

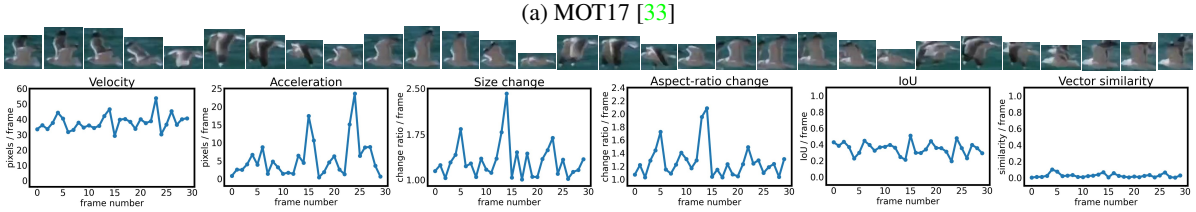
Tracking-by-Detection Approach. The tracking-by-detection approach is one of the mainstreams in MOT algorithms. SORT and its variants [5, 7, 15, 46, 47] combine the Kalman filter [44] and Hungarian [26] for track association. To improve the pre-processing object detection performance, various algorithms have also been proposed that incorporate the Re-ID branch and use object detection, such as Track-RCNN [41], JDE and its variants [21, 30, 40, 43, 53]. CenterTrack [16] utilizes CenterNet [16], an object detection method that captures objects at point origins rather than frames, to improve robustness against object occlusion. The recently proposed ByteTrack [52] uses YOLOX [19] for object detection and only IoU for the multi-stage association. Despite its simplicity, this approach achieves both high speed and high accuracy in the MOT benchmark.

An advantage of the tracking-by-detection approach is that the object detection module is separated from the tracking module, so the model parameters are smaller than the transformer-based one, which makes it easier for real-time processing. While these algorithms perform well on standard datasets such as MOT17 [33], it is still sensitive to unexpected motion changes in generic objects.

Transformer-based Approach. Trackformer [32] and MOTR [50] have been proposed as transformer-based approaches. Trackformer [32] is a transformer-based algorithm that uses background and object queries obtained by DETR [8]. MOTR [50] aims at long-term tracking by adding a novel association mechanism that associates appearance features with location information. The advantage of those methods is that they do not need to be combined with the detection step. Although those network architectures are highly expressive, they are infeasible for real-time processing due to their large number of parameters.



(a) MOT17 [33]



(b) GMOT-40 [2]

Figure 3. Time-series changes in typical target appearance and motion states in MOT17 and GMOT-40. In MOT17, there is little change in each motion state even though the target has moved. In GMOT-40, the motion states of the target changed for each frame.

2.2. Datasets for Multiple Object Tracking

Various benchmark datasets for multiple object tracking have been proposed, roughly classified into two categories: 1) datasets for person tracking and 2) datasets for generic object tracking. In contrast to datasets for image classification [14], segmentation [10, 54], and object detection [17, 28], most of datasets for multiple object tracking are more specific to persons in a crowd.

The MOT Challenge [12, 13, 27, 33] is one of the most standard MOT datasets, and is a benchmark that targets person tracking in the crowd. Among them, the MOT17 [33] dataset is widely used in MOT performance evaluations. Recently, MOT20 [13], which contains many more persons, has been presented. Datasets that focus on more specific human behaviors have also been presented. Sports MOT [11] is a dataset for players playing sports, including basketball, football, and volleyball. Many occlusions between players and high motion speeds characterize it. DanceTrack [38], on the other hand, is a dataset of dance scenes involving multiple players, each wearing the same costume and performing synchronized movements.

Recently, GMOT-40 [2] has been proposed as a dataset focused on generic objects. In GMT-40, each sequence is composed of a crowd of different generic objects. Existing datasets for generic objects [29, 51] contained much fewer objects in each frame than the common MOT dataset, e.g., MOT17 [33]. In contrast, in GMOT-40, an average of 25 generic objects exist in each frame.

3. Challenge of generic MOT

We discuss a challenge in generic object tracking by analyzing standard datasets in multi-object tracking. We compare MOT17, the standard dataset in MOT, and GMOT-40, which contains many generic objects. Table 1 shows the statistical differences in the motion states, e.g., velocity, acceleration, and shape, for tracking targets in MOT17 and GMOT-40. For a more detailed analysis, Figure 3 shows the time-series changes of a target appearance and the corresponding motion states for those datasets.

Table 1. Mean and standard deviation (Std) of motion state and IoU evaluated from their ground truth of MOT17 and GMOT-40.

Mean $\pm$ Std	MOT17 [33]	GMOT-40 [2]
Velocity	3.16 $\pm$6.15	9.35 $\pm$45.7
Acceleration	0.68 $\pm$1.87	6.15 $\pm$62.8
Size change	1.01 $\pm$0.03	1.06 $\pm$0.43
Aspect-ratio change	1.01 $\pm$0.02	1.06 $\pm$0.22
IoU	0.92 $\pm$0.12	0.82 $\pm$0.66
Vector similarity	0.81 $\pm$0.33	0.74 $\pm$0.37

Table 1 shows that the standard deviations of the aspect-ratio, velocity, and acceleration in MOT17 are considerably smaller than those in GMOT-40. In Figure 3, each plot shows velocity, acceleration, size, aspect-ratio change, IoU, and vector similarity corresponding to the ground-truth value for each frame. In MOT17, there is little change in each motion state even though the target has moved. On the other hand, in GMOT-40, the motion states of the target dramatically changed for each frame. Generic multiple-object tracking should be designed to be robust to those sudden motion changes. Those observations also suggest that 1) velocity, acceleration, and shape (size and aspect-ratio) are often dramatically changed for generic objects, as shown in Table 1, and 2) one of those motion states, e.g. acceleration, suddenly changes during tracking, as shown in Fig. 3 (b).

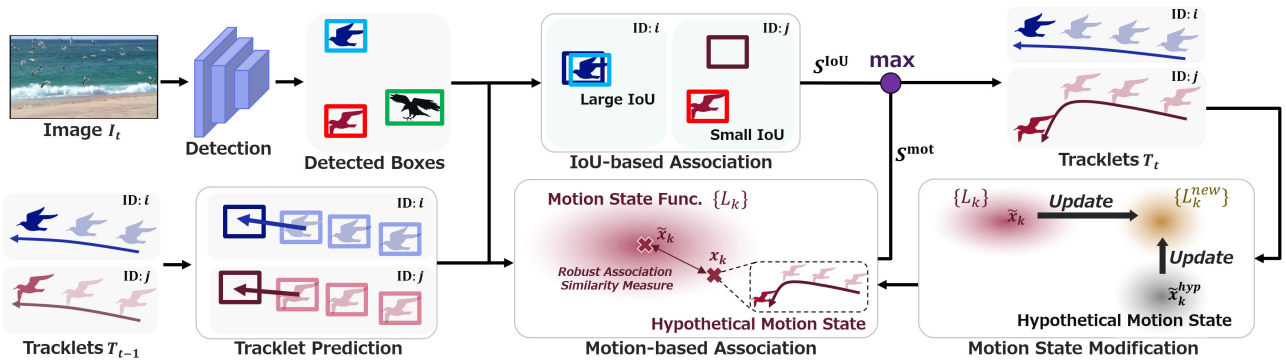


Figure 4. Overview of our proposed method. Our framework consists of detection, tracklet prediction , association using on IoU and motion based association, and motion state modification.

4. Method

The proposed method aims to improve tracking performance for generic objects with unexpected motion changes. The framework of the proposed method consists of object detection, association based on IoU and motion state, and motion state modification, as shown in Figure 4. First, YOLOX [19] is employed to obtain the detection bounding box and the corresponding confidence score. In tracklet prediction, we predict the position of each corresponding bounding box in the next frame from each previous tracklet, using a Kalman filter [22]. The associations using the IoU and the motion state are performed in the association step. These associations are integrated by maximum value calculation to update each tracklet. Finally, the updated tracklets are used to modify the motion state of each tracklet.

The proposed method is inspired by ByteTrack [52]. The main difference from ByteTrack is the improved tracking continuity for generic targets by explicitly incorporating the motion states of each tracklet into the association. In the following, the IoU-based association, the motion-based association, and the motion state modification are described.

4.1. IoU-based Association

The IoU-based association is evaluated from the IoU between the predicted and the detected bounding boxes. Various existing algorithms have employed the IoU-based association because association using IoU is a practical approach for tracking objects with small motion state changes. The IoU-based association is used for the detected bounding boxes with a confidence score above a threshold of 0.3 to associate targets with small motion changes. The IoU-based association score is represented as $S^{\text{IoU}} = \text{IoU}(\phi_{\text{A}}, \phi_{\text{B}})$, where $\text{IoU}(\phi_{\text{A}}, \phi_{\text{B}})$ is the IoU of the predicted bounding box ϕ_A and the detected bounding box ϕ_B.

4.2. Motion-based Association

We introduce motion-based associations and IoU-based associations. First, we represent the motion state of each tracklet using multiple motion states, such as the velocity and acceleration of each target. Next, the motion-based association score is obtained by evaluating the similarity between each motion state for the corresponding trajectory and the hypothesis motion state for each detected object using robust association similarity.

Motion State Representation. In this paper, as shown in Figure 5, we employ the following three values as the motion states: (1) shape, i.e., the aspect-ratio and the size of the bounding box, (2) velocity, i.e., the first-order derivative of position, and (3) acceleration, i.e., the second-order derivative of position. Note that those three motion states change dramatically for the generic object during tracking, as described in Sec. 3. Indeed, as discussed in the ablation studies in Sec. 5, the choice of those three motion states is the most effective one. Let x_k and L_k be k-th motion state and the motion state function for the k-th motion state for each tracklet, respectively. In our proposed method, the motion state function L_k for each tracklet is represented by the Laplace distribution as follows:

$$L_k(x_k; \rho_k, \Lambda_k, \tilde{x}_k) = \exp\Big(-\rho_k \Lambda_k\big(x_k; \tilde{x}_k\big)\Big), \quad (1)$$

where ρ_k, Λ_k, and $\tilde{x}_k$ are the scale for Laplacian distribution, k-th residual function, and the average for the motion state x_k. In this paper, the residual function Λ for the shape, the velocity, and the acceleration are defined as follows:

$$\Lambda_1\big(x_1{=}(r,s); \tilde{x}_1{=}(\tilde{r},\tilde{s})\big) = \max\Big\{\frac{r}{\tilde{r}}, \frac{\tilde{r}}{r}\Big\} \max\Big\{\frac{s}{\tilde{s}}, \frac{\tilde{s}}{s}\Big\} - 1, \quad (2)$$

$$\Lambda_2\big(x_2 = v; \tilde{x}_2 = \tilde{v}\big) = |v - \tilde{v}|, \quad (3)$$

$$\Lambda_3\big(x_3 = a; \tilde{x}_3 = \tilde{a}\big) = |a - \tilde{a}|, \quad (4)$$

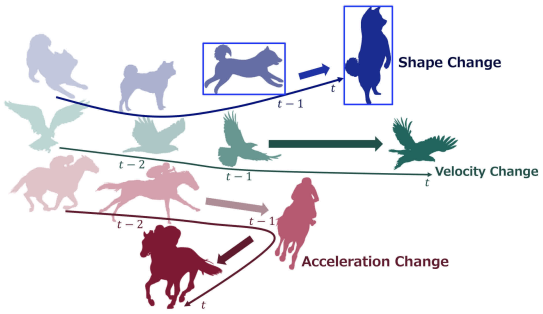


Figure 5. Three motion state variables.

where r, s, v, and a are the aspect-ratio, the size, the velocity, and the acceleration for each tracklet at the $t-1$-th frame, respectively. On the other hand, $\tilde{r}$, $\tilde{s}$, $\tilde{v}$, and $\tilde{a}$ are those values of each tracklet at the t-th frame, respectively. In this paper, the initial values of each scale ρ_k and the average motion state $\tilde{x}_k$ were evaluated using training data set.

Robust Association Similarity Measure. Using the robust association similarity, a motion-based association score S^{mot} is obtained by evaluating the similarity between each motion state x_k for the corresponding trajectory and the hypothesis motion state for each detected object. In our problem setting, it is required to robustly associate motion states that continue to change significantly during tracking. Therefore, the association similarity should be designed to be robust to the corresponding target motion state changes.

The most naive association similarity is to evaluate the product of all the motion state functions $\{L_k\}$. However, this naive design will be sensitive to changes in a small number of motion states. An unexpected change in a single motion state will result in a significant decrease in the association similarity. Our proposed method treats this changed motion state as an outlier and evaluates the association similarity based on robust statistics. Specifically, we introduce a robust similarity measure based on L-estimator so that the similarity does not decrease excessively when a small number of the motion state functions change significantly. Our motion-based association score S^{mot} is given by

$$\begin{aligned} S^{\mathrm{mot}}(\mathbf{x}) &= \lambda \cdot \max_{l} \left\{ \prod_{k\in\{\mathcal{K}\setminus\{l\}\}} L_k(x_k;\rho_k,\Lambda_k,\tilde{x}_k) \right\}, \\ \hat{l} &= \arg\max_{l} \left\{ \prod_{k\in\{\mathcal{K}\setminus\{l\}\}} L_k(x_k;\rho_k,\Lambda_k,\tilde{x}_k) \right\}, \end{aligned} \tag{5}$$

where λ is a parameter to balance the scales of S^{IoU} and S^{mot}, $\{\mathcal{K}\setminus\{l\}\}$ represents the set of all indexes $\mathcal{K}$ from which an index q is removed. We uses three motion states $x_{k=1,2,3}$: shape, velocity, and acceleration. Therefore, our motion-based association score S^{mot} is obtained using the proposed robust correlation similarity, which is equivalent to calculating the score with the highest similarity when two motion states are chosen from those three ones.

$$\begin{aligned} S^{\mathrm{mot}}(\mathbf{x}) = \lambda \cdot \max\big\{ & L_1(x_1;\rho_1,\Lambda_1,\tilde{x}_1)\cdot L_2(x_2;\rho_2,\Lambda_2,\tilde{x}_2) \\ & L_2(x_2;\rho_2,\Lambda_2,\tilde{x}_2)\cdot L_3(x_3;\rho_3,\Lambda_3,\tilde{x}_3), \\ & L_3(x_3;\rho_3,\Lambda_3,\tilde{x}_3)\cdot L_1(x_1;\rho_1,\Lambda_1,\tilde{x}_1)\big\}. \end{aligned}$$

Finally, the IoU-based association score S^{IoU} and the motion-based association score S^{mot} are integrated by maximum value calculation $S = \max(S^{\mathrm{IoU}}, S^{\mathrm{mot}})$ to update each tracklet. In other words, if either S^{IoU} or S^{mot} is higher than the threshold, add a bounding box to the corresponding ID and update the tracklet.

4.3. Motion State Modification

After the associations for each tracklet between frame t and frame $t-1$ are completed and each tracklet is finalized, the motion state for each tracklet is modified. In this step, the motion state x_k that became an outlier in the target association (i.e., the remaining one motion state excluding the two adopted motion states) is the target of this modification. If the current distribution of this outlier motion state and the hypothetical distribution (i.e., the distribution consisting of the most recent frames) are significantly different, the motion state is switched to this hypothetical distribution as follows:

$$L_k^{new}(x_k;\rho_k,\Lambda_k,\tilde{x}_k) = \begin{cases} L_k(x_k;\rho_k^{hyp},\Lambda_k,\tilde{x}_k^{hyp}), & k=\hat{l} \\ L_k(x_k;\rho_k,\Lambda_k,\tilde{x}_k), & k\neq\hat{l} \end{cases} \tag{6}$$

where ρ_k^{hyp} and $\tilde{x}_k^{hyp}$ are the scale and the mean value for the hypothetical motion state.

In the early stages of tracking, there are often short tracklet trajectories (about ten frames), so it is necessary to calculate the distance of distributions from a small sample size. The proposed method measures the distance between the existing and hypothetical distributions based on the effect size in a parametric t-test. The parametric t-test can measure the distance between distributions from a small sample size compared to an approach that explicitly measures the probability distributions of both samples. The effect size r is expressed $r = \sqrt{t_d^2/(t_d^2+df)}$, where t_d and df are the t-value and the degrees of freedom used in the t-test. The effect size r is generally closer to 1, indicating a stronger correlation between the two distributions, and closer to 0, indicating a weaker correlation between the two distributions. The correlation between distributions is generally small when the effect size r is less than 0.3. When the effect size r is less than 0.3, we switch the motion state from the current state to the hypothetical one. As will be shown in our ablation study, our adaptive modification allows tracking more robust than when all motion states are switched.

5. Experiments

5.1. Experimental Setup

Datasets. GMOT-40 [2] and MOT17 [33] were used for our evaluation. The GMOT-40 was used to evaluate MOT methods independent of target type and situation, while the MOT17 was used to compare with various existing MOT algorithms, including state-of-the-art ones. For MOT17, in addition to the MOT17-test evaluation, the MOT17 half-val was used for detailed evaluations.

Although the existing GMOT-40 is a comprehensive and highly useful data set for generic object tracking, unlike most benchmark sets, the data set protocol needs to be organized uniformly. We have prepared a dataset based on GMOT-40, which we call GMOT-Split101, and used this evaluation set for training and testing. In particular, we aligned the length of each test sequence to 101 frames, consisting of one initial frame and 100 frames for tracking, to prevent evaluation bias caused by the non-uniform length of each sequence.

Evaluation Metrics. Our approach focuses on improving tracking continuity. To evaluate the continuity of tracking on GMOT-Split101, we employed IDP, a precision measure for tracking IDs, a similar robustness measure IDR, and their overall measures IDF1 [35]. In the tracking-by-detection approach, including our framework, the detection performance, i.e., precision and recall of detection, are also evaluated because detection performance is also significantly related to tracking performance. In addition to those evaluation measures [3] (i.e., MOTP, MOTA, MT, IDS, Frag) were used in the evaluation for the MOT17-half-val dataset, where MOTP is the multiple object tracking precision, MOTA [24] is multi-object tracking accuracy, MT indicates the ratio of ground-truth trajectories that are covered by a tracking result for at least 80% of their respective lifespan, and IDS and Frag are the numbers of IDs that were swapped and broken, respectively. In MOT17-test, we used IDF1 [35], HOTA [30], MOTA, FP, FN, Recall, Precision, IDS, and Frag as standard evaluation measures, following the evaluation measures in the public benchmark set. FP/FN is the number of false positives and false negatives.

Implementation Details and Environments. The initial value of the motion scale ρ_k and the motion mean value $\tilde{x}_k$ were set to 0.4 and 0 for all motion states. The threshold θ_a for integrated similarity $S = \max(S^{\mathrm{IoU}}, S^{\mathrm{mot}})$ used for association was set to be $\theta_a = 0.6$, and the balance parameter λ was set at 0.5. The training data, e.g. MOT17 and GMOT-Split101, were only used for training our detector, i.e., YOLOX [19]. In particular, the MOT17 half-val was compared under the condition that the various parameters were not changed by the sequence in MOT17 half-val. We compared our proposed method with various tracking-by-detection-based algorithms [1,18,42,49,52,55] and transformer-based ones [32, 50] including state-of-the-arts. We experimented on an Ubuntu 20.04 64-bit PC with Intel(R) Xeon(R) Gold 6154 @ 3.00GHz CPU, Tesla V100 GPU, and 64 GB RAM, PyTorch [34].

Table 2. GMOT-Split101 benchmark results with state-of-the-art methods. The best/Second results are shown in **bold**/underline.

Method	IDF1↑	IDP↑	IDR↑	Rec.↑	Prec.↑	MOTA↑	MOTP↑
CenterTrack [55]	72.8	78.0	68.3	77.6	88.6	65.8	0.21
Trackformer [32]	78.5	**84.0**	73.7	81.1	**92.5**	**73.6**	**0.21**
ByteTrack [52]	78.8	81.3	76.5	83.8	89.2	72.3	0.20
Ours	**80.8**	83.7	**78.2**	**83.9**	89.8	73.2	0.20

5.2. Evaluation Results

Result on GMOT-Split101. To evaluate the MOT method for generic object recognition, comparative evaluations with existing methods are performed using GMOT-Split101[1]. Table 2 shows the evaluation results of the proposed and the existing methods [32, 52, 55]. Compared to CenterTrack [55], the proposed method is superior in all evaluation metrics. Comparing Trackformer [32] and the proposed method, although the proposed method is inferior in Prec., it is also clearly superior in overall performance IDF1 because it significantly outperforms IDR. (As we will describe later, the processing speed of transformer-based methods such as TrackFormer [32] is slow.) We also compared the performance with ByteTrack [52], one of the state-of-the-art tracking-by-detection algorithms. Note that the original code is designed explicitly for MOT benchmarks and person crowds, and it eliminates horizontal tracking results with aspect-ratios w/h greater than 1.6. For fair evaluation on GMOT-Split101, we use the ByteTrack code without using this aspect-ratio restriction. Compared to ByteTrack, the proposed method improved overall, including the evaluation value for ID, indicating that the proposed method is effective for various types of generic targets. In particular, our method outperforms those algorithms with an overall 2% improvement in IDF1. Note that even though our method with the same hyper-parameters is designed for generic object targets, as shown in Table 4, the similar effectiveness of our method is also shown in the result on the DanceTrack dataset [38], a large-scale dataset for a specific target. This is because the proposed method explicitly uses a motion state in its algorithm design, while ByteTrack uses only IoU for the association [2].

Visual Comparisons and Discussion. Figure 6 shows the visual comparisons of our proposed and the existing method. The areas indicated by the yellow and white arrows

[1]The details of our datasets are described in our supplemental material.

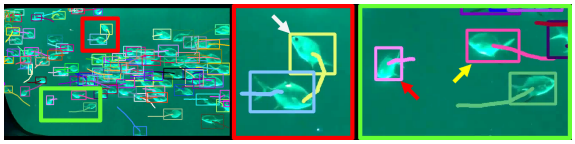
[2]Additional results, analysis of computational efficiency, and detailed discussions of the limitations of our method are also described in Sec. A of our supplemental material.

Table 3. MOT17-val-half benchmark results with state-of-the-art methods. The best/Second results are shown in **bold**/underline.

Method	IDF1↑	IDP↑	IDR↑	Rec.↑	Prec.↑	MT↑	IDS↓	Frag↓	MOTA↑	MOTP↑
CenterTrack [55]	64.2	74.2	56.6	71.6	94.1	41.3	528	588	66.1	**17.9**
Trackformer [32]	71.5	83.4	62.5	73.2	**97.7**	48.7	345	**417**	70.8	14.6
ByteTrack [52]	77.0	81.1	73.2	82.7	91.6	56.0	206	505	74.7	17.1
Ours	**79.2**	**83.9**	**75.0**	**83.2**	93.1	**58.4**	**170**	477	**76.7**	15.5

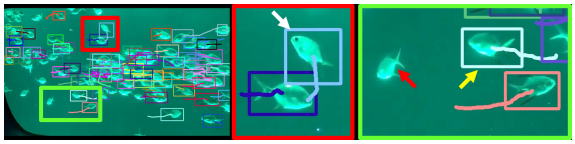
Table 4. Analysis using DanceTrack validation set [38].

Method	IDF1↑	IDP↑	IDR↑	Rec.↑	Prec.↑	MOTA↑	MOTP↑
ByteTrack [52]	46.9	46.1	47.8	94.6	91.2	0.843	**0.18**
Ours	**49.9**	**50.2**	**49.6**	**92.3**	**93.6**	**0.85**	0.178



(a) Ground truth

(b) Result by ByteTrack [52]



(c) Result by proposed method

Figure 6. Examples of existing method and our proposed method. The areas indicated by the yellow and white arrows in (b) and (c) show that the proposed method can track the target fish for a long period, whereas [52] can only track it for a short period.

in Figures 6 (b) and (c) show that the proposed method can track the target fish for a long period, whereas ByteTrack can only track it for a short period. However, as the red arrow indicates, failure to detect the fish makes it difficult to track the object for longer.

In general, if detection fails due to target occlusion, this failure will degrade tracking performance. These limitations are common challenges for tracking-by-detection approaches, and improving object detection performance is also an essential factor. To further analyze the tracking performance without the effects of object detection failure, we evaluated the tracking performance on GMOT-40 using ideal detection results. Table 5 (a) shows that the gain from the conventional method is further increased when detection is ideal.

Note that, interestingly, as shown in Table 5 (b), this advantage is enhanced when the frame rate is small. This huge

Table 5. Additional analysis using ideal detection results.

(a) Result on GMOT-40 [2] with ideal detection

Method	IDF1↑	IDP↑	IDR↑	Rec.↑	Prec.↑	MOTA↑	MOTP↑
ByteTrack [52]	89.3	90.1	88.6	97.3	98.9	95.3	0.095
Ours	**92.0**	**92.3**	**91.8**	**97.9**	**98.4**	**95.7**	**0.098**

(b) Low fps setting evaluation based on IDF1

Method	Videos sampling ratio						
	1	5	10	15	20	25	30
ByteTrack [52]	89.3	72.8	62.1	53.4	47.6	43.6	38.5
Ours	**92.0**	**75.0**	**65.0**	**57.9**	**52.0**	**49.0**	**45.7**
	(+2.7)	(+2.2)	(+2.9)	(+4.5)	(+4.4)	(+5.4)	(+7.2)

drop in ByteTrack is due to the assumption that motion between frames is small, and the proposed method can compensate for this shortcoming[3]. These experiments suggest that improving the detection accuracy further improves the effectiveness of the proposed method.

In contrast to person tracking, generic object tracking involves large geometric transformations such as rotation. The object detection algorithm is known to be sensitive to large geometric transformations [23]. It is essential to improve both object tracking and object detection algorithms.

Result on MOT17. The performance was also evaluated on MOT17 (MOT17-half-val and MOT17-test). Tables 3 and 6 show the evaluation results of the proposed and the existing methods. Note that for processing speed, we referred to the submitted values of the MOT17 test set for a fair comparison. As shown in Table 3, our proposed method outperforms existing methods (e.g., ByteTrack) in terms of evaluation metrics related to tracking continuity (e.g., IDF, IDP, IDR), similar to the evaluation results for GMOT-Split101. This result suggests that our key idea (i.e., integration of motion state and IoU) is also effective for person-tracking tasks. Furthermore, as shown in Table 6, even in the MOT17-test, widely used for evaluation, the proposed method is comparable or superior to the existing methods in several evaluation metrics including IDF1. In particular, compared to ByteTrack, our proposed method adds additional processing (i.e., associating with the motion state and modifying the motion state) [4]. However, our

[3]This shortcoming in ByteTrack has also been described in [20].

[4]Note that MOTA and IDS are highly dependent on detection performance. As FN decreases, IDS and FM increase and MOTA also decreases. As previously discussed in Table 5 (a), tracking performance improves for

Table 6. MOT17-test benchmark results with state-of-the-art methods. The best/Second results are shown in **bold**/underline.

Method	IDF1↑	HOTA↑	MOTA↑	FP↓	FN↓	Rcll↑	Prcn↑	IDS↓	Frag↓	FPS↑
CenterTrack [55]	64.7	52.2	67.8	**18,498**	160,332	-	-	3,039	-	17.5
Trackformer [32]	68.0	57.3	74.1	34,602	108,777	80.7	92.9	2,829	4,221	5.7
STC-Tracker [18]	70.9	59.8	75.8	44,952	87,039	84.6	91.4	4,533	5,721	9.5
GSDT [42]	66.5	55.2	73.2	26,397	120,666	78.6	94.4	3,891	8,604	4.9
MOTR [50]	75.0	62.0	78.6	23,409	94,797	83.2	95.3	2,619	6,231	7.5
ReMOT [49]	72.0	59.7	77.0	33,204	93,612	83.4	93.4	2,853	5,304	1.8
ByteTrack [52]	77.3	63.1	80.3	25,491	83,721	85.2	95.0	2,196	2,277	**29.6**
BoT-SORT [1]	**80.2**	**65.0**	**80.5**	22,521	86,037	84.8	**95.5**	**1,212**	**1,803**	6.8
Ours	77.8	63.5	79.5	28,557	**74,718**	**86.8**	94.4	2,244	2,376	28.2

proposed method is still fast enough to allow for real-time processing, and substantially outperforms the transformer-based methods (TrackFormer [32] and MOTR [50]) in terms of the processing speed. Note that although BoT-SORT [1] achieves comparable or better performance than our proposed method in some evaluation metrics, the proposed method is significantly superior in processing speed.

5.3. Ablation Study and Analysis

Finally, we conducted two ablation studies: 1) the effectiveness of motion state variable selection and 2) the effectiveness of motion state modification.

Effectiveness of Motion State Variable Selection. In the proposed method, three motion states, i.e., shape, velocity, and acceleration, were selected to represent the motion state of each tracked target. We added jerk and velocity directional correlation (called vector in the following) as other candidates of the motion state and evaluated the best motion state selection, including three variables we used in our method. Note that jerk is the first derivative of acceleration (i.e., the third derivative of position), and vector is the cosine similarity S_{cos} of velocity normalized to $L_{cos} = (S_{cos} + 1)/2$. As shown in Table 7, the proposed combination of the three motion states (i.e., shape, velocity, and acceleration) is the most effective. This result indicates that combining these motion states is a simple-yet-effective choice for representing motion states. While acceleration is the most key motion state in motion modification, velocity, and shape are also key motions in the bird sequence shown in Fig. 2. Details are described in our supplemental.

Effectiveness of Motion State Modification. In the proposed method, we select the motion states that are the outlier and adaptively modify these motion states. To evaluate the effectiveness of this adaptive motion-state modification, we compared the performance of the proposed method (i.e., adaptive modification) with two baselines: 1) no modification (i.e., ρ_k and $\tilde{x}_k$ are fixed at their initial values) and 2) complete modification (i.e., all motion states ρ_k and $\tilde{x}_k$) are switched). Table 8 shows that the proposed method outperforms these two naive baselines. Interestingly, when all parameters are modified, all metrics, i.e., IDF1, IDP, IDR, are lower than those of the proposed method. This is because the mandatory modification of motion states that do not require modification results in tracking failures, suggesting that this complete modification strategy overfits the hypothetical motion state. Those results show that our adaptive motion state correction in our framework is effective.

all indicators when detection results are ideal.

Table 7. Comparison results on GMOT-Split101 for the combination use of different motion state of the proposed method.

Set of motion states	IDF1↑	IDP↑	IDR↑
(shape,velocity,vector)	78.7	80.7	76.8
(shape,acceleration,vector)	78.7	80.7	76.7
(shape,jerk,vector)	78.7	80.8	76.7
(shape,jerk,velocity)	80.2	83.4	77.2
Ours	**80.8**	**83.7**	**78.2**

Table 8. Comparison of the results from different modification protocols on GMOT-Split101.

Modification manner	IDF1↑	IDP↑	IDR↑
No modification	78.1	79.8	76.5
Complete modification	79.5	81.9	77.2
Ours (adaptive modification)	**80.8**	**83.7**	**78.2**

6. Conclusion

We proposed the simple-yet-effective generic MOT algorithm called FRoG-MOT that can improve tracking continuity by associating each target with adaptively modified motion state. The keys are 1) to represent the target motions using multiple motion states that have weak correlations with each other and 2) to modify those states that have the lowest similarity to past states as outliers. Our approach can improve tracking continuity and robustness to unexpected motion changes of each object. Comprehensive experiments have confirmed that our framework is comparable to existing state-of-the-art methods on a standard dataset and outperforms those algorithms on the GMOT-Split101 dataset.

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. 6, 8

[2] Hexin Bai, Wensheng Cheng, Peng Chu, Juehuan Liu, Kai Zhang, and Haibin Ling. Gmot-40: A benchmark for generic multiple object tracking. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 6719–6728, 2021. 1, 2, 3, 6, 7

[3] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *J. on Image and Video Process.*, 2008:1–10, 2008. 6

[4] Margrit Betke, Esin Haritaoglu, and Larry S Davis. Real-time multiple vehicle detection and tracking from a moving vehicle. *Mach. Vis.and Appl. (MVA)*, 12:69–83, 2000. 1

[5] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 3464–3468. IEEE, 2016. 1, 2

[6] 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 *IEEE Conf. Comput. Vis. Pattern Recog.(CVPR)*, pages 11621–11631, 2020. 1

[7] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 9686–9696, 2023. 2

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis. (ECCV)*, 2020. 2

[9] MMTracking Contributors. MMTracking: OpenMMLab video perception toolbox and benchmark. https://github.com/open-mmlab/mmtracking, 2020. 1

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016. 3

[11] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. *arXiv preprint arXiv:2304.05170*, 2023. 3

[12] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé. CVPR19 tracking and detection challenge: How crowded can it get? *arXiv:1906.04567 [cs]*, June 2019. arXiv: 1906.04567. 3

[13] 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:2003.09003[cs]*, Mar. 2020. arXiv: 2003.09003. 3

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. of Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 248–255. Ieee, 2009. 3

[15] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Trans. on Multimedia*, 2023. 2

[16] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Int. Conf. Comput. Vis. (ICCV)*, pages 6569–6578, 2019. 2

[17] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis. (IJCV)*, 111(1):98–136, 2015. 3

[18] Amit Galor, Roy Orfaig, and Ben-Zion Bobrovsky. Strongtranscenter: Improved multi-object tracking based on transformers with dense representations. *arXiv preprint arXiv:2210.13570*, 2022. 6, 8

[19] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2, 4, 6

[20] Andreu Girbau, Ferran Marqués, and Shin'ichi Satoh. Multiple object tracking from appearance by hierarchically clustering tracklets. In *Brit. Mach. Vis. Conf. (BMVC)*, 2022. 7

[21] Song Guo, Jingya Wang, Xinchao Wang, and Dacheng Tao. Online multiple object tracking with cross-task synergy. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 8136–8145, 2021. 2

[22] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960. 4

[23] Agastya Kalra, Guy Stoppi, Bradley Brown, Rishav Agarwal, and Achuta Kadambi. Towards rotation invariance in object detection. In *Int. Conf. Comput. Vis. (ICCV)*, 2021. 7

[24] Rangachar Kasturi, Dmitry Goldgof, Padmanabhan Soundararajan, Vasant Manohar, John Garofolo, Rachel Bowers, Matthew Boonstra, Valentina Korzhova, and Jing Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):319–336, 2008. 6

[25] Dieter Koller, Joseph Weber, and Jitendra Malik. Robust multiple car tracking with occlusion reasoning. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 189–196. Springer, 1994. 1

[26] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 2

[27] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942 [cs]*, Apr. 2015. arXiv: 1504.01942. 3

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 740–755. Springer, 2014. 3

[29] Chongyu Liu, Rui Yao, S Hamid Rezatofighi, Ian Reid, and Qinfeng Shi. Model-free tracker for multiple objects using joint appearance and motion inference. *IEEE Transactions on Image Processing*, 29:277–288, 2019. 3

[30] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *Int. J. Comput. Vis. (IJCV)*, 129:548–578, 2021. 2, 6

[31] Wenhan Luo, Tae-Kyun Kim, Bjorn Stenger, Xiaowei Zhao, and Roberto Cipolla. Bi-label propagation for generic multiple object tracking. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1290–1297, 2014. 1

[32] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 8844–8854, 2022. 2, 6, 7, 8

[33] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, Mar. 2016. arXiv: 1603.00831. 1, 2, 3, 6

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 32:8026–8037, 2019. 6

[35] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Eur. Conf. Comput. Vis. Workshops (ECCVW)*, pages 17–35. Springer, 2016. 6

[36] Concetto Spampinato, Yun-Heh Chen-Burger, Gayathri Nadarajan, and Robert B Fisher. Detecting, tracking and counting fish in low quality unconstrained underwater videos. *Int. Conf. Comput. Vis. Theory Appl. (VISAPP)*, 2008(514-519):1, 2008. 1

[37] Concetto Spampinato, Simone Palazzo, Daniela Giordano, Isaak Kavasidis, Fang-Pang Lin, and Yun-Te Lin. Covariance based fish tracking in real-life underwater environment. In *Int. Conf. Comput. Vis. Theory Appl. (VISAPP)*, pages 409–414, 2012. 1

[38] 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 *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 1, 3, 6, 7

[39] 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 *IEEE Conf. Comput. Vis. Pattern Recog.(CVPR)*, pages 2446–2454, 2020. 1

[40] Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon. Learning to track with object permanence. In *Int. Conf. Comput. Vis. (ICCV)*, pages 10860–10869, 2021. 2

[41] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 7942–7951, 2019. 2

[42] Yongxin Wang, Kris Kitani, and Xinshuo Weng. Joint Object Detection and Multi-Object Tracking with Graph Neural Networks. *arXiv:2006.13164*, 2020. 6, 8

[43] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 107–122. Springer, 2020. 1, 2

[44] Greg Welch, Gary Bishop, et al. An introduction to the kalman filter. 1995. 2

[45] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Adv. Neural Inform. Process. Syst. Datasets and Benchmarks Track*, 2021. 1

[46] Nicolai Wojke and Alex Bewley. Deep cosine metric learning for person re-identification. In *2018 IEEE Winter Conf. on Appl. of Comput. Vis. (WACV)*, pages 748–756. IEEE, 2018. 2

[47] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 3645–3649. IEEE, 2017. 1, 2

[48] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1

[49] Fan Yang, Xin Chang, Sakriani Sakti, Yang Wu, and Satoshi Nakamura. Remot: A model-agnostic refinement for multiple object tracking. *Image and Vision Computing*, 106:104091, 2021. 6, 8

[50] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 659–675. Springer, 2022. 2, 6, 8

[51] Lu Zhang and Laurens Van Der Maaten. Preserving structure in model-free tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):756–769, 2013. 3

[52] 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 *Eur. Conf. Comput. Vis. (ECCV)*, pages 1–21. Springer, 2022. 1, 2, 4, 6, 7, 8

[53] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *Int. J. Comput. Vis. (IJCV)*, 129:3069–3087, 2021. 1, 2

[54] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 633–641, 2017. 3

[55] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 474–490. Springer, 2020. 6, 7, 8