An Improved Association Pipeline for Multi-Person Tracking

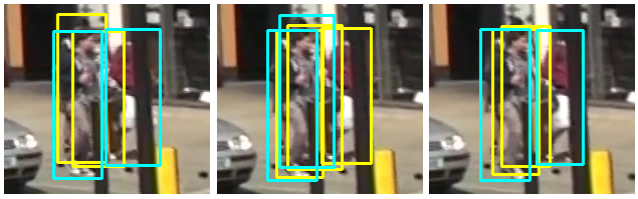
Daniel Stadler[1,2,3] Jürgen Beyerer[2,1,3]
[1]Karlsruhe Institute of Technology [2]Fraunhofer IOSB [3]Fraunhofer Center for Machine Learning
{daniel.stadler,juergen.beyerer}@iosb.fraunhofer.de

Abstract

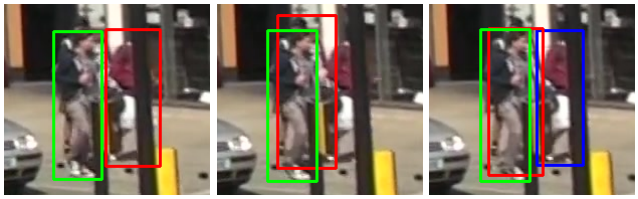
The association task of assigning detections to tracks in multi-person tracking has recently been improved by integration of a second matching stage for low-confident detections that are usually discarded in the tracking process. Despite its success, we find that this two stage matching has some weaknesses. For example, high-confident detections are preferred over low-confident detections in any case, even if the low-confident ones are more accurate. Therefore, a Combined Matching (CM) is proposed which considers all possible assignments simultaneously in a single matching stage and thus improves the association accuracy. Moreover, shortcomings of existing motion and appearance distance combinations are identified and a novel Combined Distance (CD) for motion and appearance information is introduced that significantly outperforms previous fusion approaches. Furthermore, we propose an Occlusion Aware Initialization (OAI) which prevents the start of ghost tracks from duplicate detections under occlusion. The effectiveness of our components is shown with extensive ablative experiments and the competitiveness of our tracker is demonstrated on the MOT17 and MOT20 benchmarks, where the current state-of-the-art is notably surpassed.

1. Introduction

Multi-person tracking (MPT) is a fundamental task in computer vision with application areas including autonomous driving, robotics, and surveillance. Most existing works divide the MPT task into detection and association and solve the two sub-problems independently in a *tracking-by-detection* based approach [1, 3–5, 9, 29, 36, 38, 42]. Typically, only high-confident detections are considered in the tracking process because a lot of false-positives are among the low-confident detections that would introduce *ghost* tracks, *i.e.*, false-positive tracks, when using them for track initialization. ByteTrack [42], however, introduces a second association stage in which low-confident detections are matched to the remaining unassigned tracks that have not been assigned a high-confident detection in the first stage. Since the low-confident detections are only assigned to already tracked targets but are not utilized for track initialization, association performance is improved without the undesired start of ghost tracks. Because of its success, this two stage matching (TSM) has been adopted by many subsequent works [1, 5, 21, 27, 40]. However, we find that the TSM has the following two shortcom-



(a) High-confident (cyan) and low-confident (yellow) detections.



(b) Tracks after two stage matching (baseline).

(c) Tracks after Combined Matching (*ours*).

Figure 1. Improved association with the proposed Combined Matching (CM) approach. (a) High-confident detections and low-confident detections (categorized by detection score) are shown. (b) Previous methods [1,5,21,27,40,42] first match high-confident detections to tracks before matching low-confident detections to remaining unassigned tracks. This leads to an ID switch in the middle frame. (c) In our CM, all detections are considered simultaneously. That allows better fitting low-confident detections to be favored over high-confident ones which improves the association and prevents the ID switch in the example. Note that in both strategies, only high-confident detections are used to start new tracks.

ings. First, high-confident detections are preferred over low-confident ones without taking the matching distances into account. Second, only *active* tracks are leveraged in the second stage such that assignments of low-confident detections to *inactive* tracks, *i.e.*, tracks without assigned detection in the previous frame, are not possible. To solve these problems, we propose a Combined Matching (CM) approach that considers all possible assignments simultaneously in a single matching stage and thus improves the utilization of low-confident detections as well as inactive tracks in the association process. Figure 1 depicts an example sequence, where CM performs better than TSM.

Besides the matching strategy, the distance measures that are used for representing the similarity between detections and tracks highly influence the association performance. As both motion and appearance information is valuable for MPT, some trackers fuse the information and build a combined distance measure [1, 9, 14, 35]. We identify several weaknesses in those approaches and experiment with various metrics for motion distance as well as different strategies for calculating the appearance distance. On the basis of our findings, we introduce a novel Combined Distance (CD) for motion and appearance information that outperforms previous fusion approaches by a large margin.

After the association, unassigned detections are used for track initialization. While usually a confidence threshold is applied such that only high-confident detections start new tracks, the surroundings of a detection are not taken into account for track initialization. However, we argue that unassigned detections that have high overlaps with tracks are most likely duplicate detections and should be removed before track initialization. Therefore, we propose an Occlusion Aware Initialization (OAI) that computes for each unassigned detection the Intersection over Union (IoU) with all tracks and removes it if the maximum IoU exceeds a predefined threshold which reduces the number of ghost tracks.

With the proposed tracking components, we build an advanced framework for MPT termed *ImprAsso* (Improved Association). Our design choices are validated with comprehensive ablative experiments and the superiority of ImprAsso w.r.t. other existing methods is demonstrated on the two MPT benchmarks MOT17 [20] and MOT20 [6], where state-of-the-art results are achieved. The main contributions of our work are summarized in the following:

- A novel Combined Matching (CM) approach is proposed which addresses the weaknesses of the common two stage matching and improves the utilization of low-confident detections in the association.
- We introduce a Combined Distance (CD) for motion and appearance information that significantly outperforms previous fusion methods and investigate various strategies for the calculation of appearance distance.
- A new Occlusion Aware Initialization (OAI) technique is proposed that takes the surroundings of unassigned detections into account and prevents the start of ghost tracks from duplicate detections under occlusion.
- Leveraging the proposed components, we build a sophisticated tracking framework that surpasses the state-of-the-art on the MOT17 and MOT20 datasets.

2. Related Work

We give an overview of related work on distance measures for the association task, matching strategies to assign detections to tracks, and approaches for track initialization.

2.1. Distance Measures

The distance function to determine the similarity between so-far tracked targets and detections from the current time step plays an important role in any tracking-by-detection (TBD) approach. Hence, various distance functions are applied in MPT. Many methods rely only on IoU distance as motion information [3–5, 27, 42] due to its low computational complexity in contrast to appearance information for which a separate convolutional neural network (CNN) is applied [9, 16, 29, 36, 38]. Such a CNN is often adopted from the re-identification community and used to extract high-dimensional features from the detected image patches. Then, the cosine distance of these features is leveraged as appearance distance. There also exist some trackers that build upon a joint detection and embedding network [15, 31, 35, 43]. As this network is trained to perform detection and feature extraction simultaneously, a separate CNN for appearance information becomes obsolete.

For a high association accuracy, it is beneficial to use both motion and appearance information. DeepSORT [36] calculates Mahalanobis distance between Kalman filter [12] predicted track states and detections for motion information instead of IoU. However, the Mahalanobis distance is only used to prevent infeasible assignments and appearance distance is taken for matching. StrongSORT [9], as a further development of DeepSORT, combines Mahalanobis and appearance distance by a weighted sum. The same is done in JDE [35]. In our Combined Distance (CB), we also fuse motion and appearance information with a weighted sum, however, we do not use the Mahalanobis distance. The reason for this is that the Mahalanobis distance gives only rough estimates of the object location when the track state uncertainty is high [36]. Instead, we leverage Distance IoU [44] for motion information which explicitly models the normalized distance between the central points of bounding boxes in addition to the standard IoU formulation. Similarly, the Generalized IoU [22] is combined with appearance information in SimpleTrack [14]. Another fusion mechanism can be found in BoT-SORT [1], where

the minimum of IoU and appearance distance is utilized. We find that this approach does not fully exploit the potential of both motion and appearance information because either the one or the other one is used. Different from previous approaches [9, 14, 35], our CD additionally integrates a minimum IoU requirement. This is important when much weight is given to the appearance information and many targets are in the scene as the risk for confusing different persons with similar appearances is high in crowded scenes.

When calculating the appearance cosine distance between detections and tracks, different strategies can be pursued since for the tracks, features from multiple time steps are used. DeepSORT [36] maintains for each track a feature bank of the assigned detection features from the last time steps and takes the minimum distance of all saved features to the features of a current detection as appearance distance. In [9,35], the track feature is consequently updated in an exponential moving average manner such that only one cosine distance per track-detection pair needs to be calculated. We combine the two approaches by computing a mean feature from the feature bank which then is used as track feature in the cosine distance calculation to the detection features.

2.2. Matching Strategies

Many TBD methods match high-confident detections to tracks in a single stage [3,4,9]. In contrast, DeepSORT [36] applies a matching cascade that favors recently observed tracks. The motivation of the cascaded scheme is that the accuracy of propagated locations of inactive tracks decreases over time. In StrongSORT [9], however, it is shown that such a matching cascade decreases the performance when the tracker gets better because the additional prior constraints limit the association accuracy. ByteTrack [42] proposes a second matching stage in which low-confident detections are associated with unmatched tracks from the first stage. The low-confident detections are not used to start new tracks such that no ghost tracks from low-confident false positive detections are introduced. The authors of ByteTrack show that this two stage matching (TSM) improves the tracking performance when integrated into various frameworks [42]. Consequently, TSM is adopted in many following works [1,5,21,27,40]. Despite its success, we identify two shortcomings of TSM: High-confident detections are preferred over low-confident ones in any case and no assignments of low-confident detections to inactive tracks are possible. To solve these problems, we propose a Combined Matching (CM) strategy that fuses the two matching stages into one matching stage in which all possible assignments are considered simultaneously.

2.3. Track Initialization

In new Transformer [30] based *tracking-by-attention* architectures, detection and tracking are integrated more tightly than in TBD which also changes the concept of track initialization. For example, TrackFormer [19] uses a fixed number of object queries for recognizing new targets besides track queries that carry identity information of already tracked targets. Thus, track initialization is to an extent learnt by the neural network itself, whereas in TBD, a heuristic has to be applied that determines which detections should initialize new tracks after the association.

Naturally, only unmatched detections with high confidence are considered for track initialization such that no or only few ghost tracks stemming from false positive detections are started. Some TBD methods apply a higher confidence threshold for initialization than for association [1, 5, 27, 40, 42], *i.e.*, when a target is detected once with high confidence, detections with lower confidence can be assigned to it in consecutive frames. To suppress false positive detections which occur only in single frames, many methods first start *tentative* tracks that have to be confirmed with assigned detections in consecutive frames before the tentative tracks are activated [1, 3, 5, 9, 42]. Besides detection confidence and continuity, we think that the surroundings of a detection contain important information. Our Occlusion Aware Initialization (OAI) takes this into account by calculating for each unassigned detection the IoU to all track boxes and prevents an initialization if the maximum IoU exceeds an overlap threshold, arguing that the detection is probably a duplicate detection. Note that our OAI also applies a confidence threshold and can additionally be combined with the strategy of starting tentative tracks first.

3. Improved Association

In Section 3.1, we propose Combined Matching (CM) which fuses previously used two matching stages to one stage such that the usage of low-confident detections is improved. A Combined Distance (CD) of motion and appearance information for enhanced association accuracy is presented in Section 3.2. The Occlusion Aware Initialization (OAI) to prevent the start of ghost tracks under occlusion is introduced in Section 3.3 and finally, we give an overview of the proposed tracking framework in Section 3.4.

3.1. Combined Matching

Before the presentation of ByteTrack [42], low-confident detections $\mathcal{D}^{\mathrm{l}}$ with score s below a tracking threshold s_{track} were typically discarded and only high-confident detections $\mathcal{D}^{\mathrm{h}}$ with $s > s_{\mathrm{track}}$ were used in the tracking process. In [42], a two stage matching (TSM) is proposed that first assigns $\mathcal{D}^{\mathrm{h}}$ to tracks $\mathcal{T}$ and afterwards assigns $\mathcal{D}^{\mathrm{l}}$ to the remaining unmatched *active* tracks $\mathcal{T}^{\mathrm{a,u}}$. We find that this TSM hast two shortcomings: First, *inactive* tracks $\mathcal{T}^{\mathrm{i}}$, *i.e.*, tracks without assigned detection in the previous frame, cannot be matched with $\mathcal{D}^{\mathrm{l}}$. Second, $\mathcal{D}^{\mathrm{h}}$ are preferred over $\mathcal{D}^{\mathrm{l}}$ even if there exist $\mathcal{D}^{\mathrm{l}}$ that fit much better than $\mathcal{D}^{\mathrm{h}}$.

We solve these problems with a Combined Matching (CB) approach that considers all possible assignments simultaneously. Distances between tracks and $\mathcal{D}^{\text{h}}$ (first stage in TSM) and between tracks and $\mathcal{D}^{\text{l}}$ (second stage in TSM) are combined in one distance matrix which then is leveraged in a single matching stage. More precisely, let $\mathcal{T} = \{T_1, \dots, T_m\}$ and $\mathcal{D}^{\text{h}} = \{D_1^{\text{h}}, \dots, D_n^{\text{h}}\}$ be the sets of tracks and high-confident detections, respectively. The distance matrix $\mathbf{D}^{\text{h}} = (d_{ij})_{i \in [1,m], j \in [1,n]}$ containing all distances between $\mathcal{T}$ and $\mathcal{D}^{\text{h}}$ is computed leveraging a distance function d^{h}. Equally, $\mathbf{D}^{\text{l}}$ is calculated for $\mathcal{T}$ and $\mathcal{D}^{\text{l}}$ with distance d^{l}. Note that arbitrary distance functions d can be applied for d^{h} and d^{l}. Finally, $\mathbf{D}^{\text{l}}$ is multiplied with a normalization factor β and then concatenated with $\mathbf{D}^{\text{h}}$ to get the combined distance matrix $\mathbf{D}^*$:

$$\mathbf{D}^* = (\mathbf{D}^{\text{h}} \;\; \beta\,\mathbf{D}^{\text{l}}) \tag{1}$$

For a reasonable combination, β should account for different scales of $\mathbf{D}^{\text{h}}$ and $\mathbf{D}^{\text{l}}$ due to different distance functions d^{h} and d^{l}. Moreover, a stricter maximum matching distance $d_{\max}$ should be applied for $\mathcal{D}^{\text{l}}$ as they are more inaccurate on average. To achieve this, we calculate β as the ratio of the maximum distances $d_{\max}^{\text{h}}$ and $d_{\max}^{\text{l}}$ which are adopted from the first and second stage of TSM, respectively:

$$\beta = \frac{d_{\max}^{\text{h}}}{d_{\max}^{\text{l}}} \tag{2}$$

Thus, a maximum distance of $d_{\max}^* = d_{\max}^{\text{h}}$ is applied in the CM. Note that with this choice, the same assignments as in TSM are possible with the desired extension that $\mathcal{D}^{\text{l}}$ can be favored over $\mathcal{D}^{\text{h}}$ and also matched to $\mathcal{T}^{\text{i}}$.

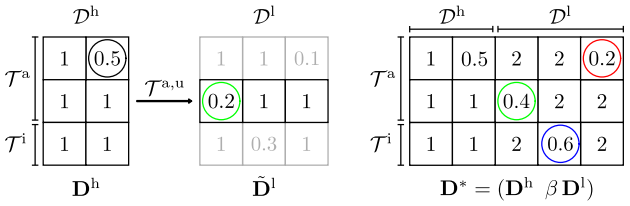
Practically, we first employ TSM and tune the maximum matching distances $d_{\max}^{\text{h}}$ and $d_{\max}^{\text{l}}$. Then, β is calculated with Equation (2) to apply the CM with $\mathbf{D}^*$ from Equation (1). The linear sum assignment problem is solved with the Hungarian method [13] that minimizes the overall costs from the combined distance matrix $\mathbf{D}^*$. Figure 2 illustrates the benefits of our CM compared to the TSM baseline.

3.2. Combined Distance

It has been shown in previous works that fusing motion and appearance information is beneficial in multi-person tracking [1, 9, 35]. In BoT-SORT [1], appearance cosine distance d_{APP} is integrated in the first stage of TSM. It is combined with IoU based motion distance d_{MOT} by taking the minimum of d_{APP} and d_{MOT}. We find that this fusion mechanism is not optimal as either one or the other distance is used but not both. Instead, combining d_{MOT} and d_{APP} as weighted sum with parameter λ is more promising:

$$d_{\text{MOT+APP}} = \lambda d_{\text{MOT}} + (1 - \lambda) d_{\text{APP}} \tag{3}$$

Note that JDE [35] and StrongSORT [9] fuse motion and appearance distance in the same way but use Mahalanobis



(a) TSM. (b) CM ($\beta = 2$).

Figure 2. Toy example of two stage matching (TSM) as in [42] vs. our Combined Matching (CM) approach. (a) In the first stage of TSM, active tracks $\mathcal{T}^{\text{a}}$ and inactive tracks $\mathcal{T}^{\text{i}}$ are matched with high-confident detections $\mathcal{D}^{\text{h}}$ based on $\mathbf{D}^{\text{h}}$. In the second stage, unassigned *active* tracks $\mathcal{T}^{\text{a,u}}$ are matched with low-confident detections $\mathcal{D}^{\text{l}}$ based on $\tilde{\mathbf{D}}^{\text{l}}$. Unconsidered assignments are indicated in gray. (b) In contrast, our CM considers all possible assignments simultaneously. Instead of preferring $\mathcal{D}^{\text{h}}$ over $\mathcal{D}^{\text{l}}$ at any cost, we introduce a normalization factor β that is multiplied with the distances $\mathbf{D}^{\text{l}}$ from low-confident detections $\mathcal{D}^{\text{l}}$. This allows better fitting detections from $\mathcal{D}^{\text{l}}$ to be matched (red circle) instead of more inaccurate detections from $\mathcal{D}^{\text{h}}$ (black circle in (a)). Moreover, assignments of $\mathcal{D}^{\text{l}}$ to *inactive* tracks $\mathcal{T}^{\text{i}}$ (blue circle) are possible.

distance for motion information. Since the Mahalanobis distance is only a rough estimation of the object location if the state uncertainty is high [36], IoU based distances are a better choice for motion information as we will see in the experimental evaluation. Besides IoU and Mahalanobis distance, we also experiment with the Distance IoU (DIoU) from [44] for d_{MOT}, which explicitly models the normalized distance between the central points of bounding boxes.

One shortcoming of $d_{\text{MOT+APP}}$ from Equation (3) in tracking on sequences with high-frame rates is that it allows matches of non-overlapping boxes if both λ and d_{APP} are small. Indeed, we find $\lambda = 0.2$ to be a good choice on the utilized dataset and observe wrong matches when persons with similar appearance occur, which frequently happens in scenes with a large number of targets. Therefore, we integrate a minimum IoU requirement of $o_{\min}$ into Equation (3) which yields the proposed Combined Distance d_{CD}:

$$d_{\text{CD}} = \begin{cases} \lambda d_{\text{MOT}} + (1 - \lambda) d_{\text{APP}} & \text{if IoU} > o_{\min} \\ d_{\max} + \epsilon & \text{otherwise} \end{cases} \tag{4}$$

Here, $d_{\max}$ denotes the maximum allowed distance for matching and ϵ is a very small value, *e.g.*, $1e^{-5}$.

While only the current track state contributes to the motion distance d_{MOT}, the appearance distance d_{APP} integrates information from the past. Various strategies for computing d_{APP} can be pursued and we experiment with three different approaches: First, a feature bank is built for each track with the features of the last n_{feat} assigned detections and the minimum distance of these features to the features of a current detection is chosen [36]. Second, only one track feature is maintained that is updated in an expo-

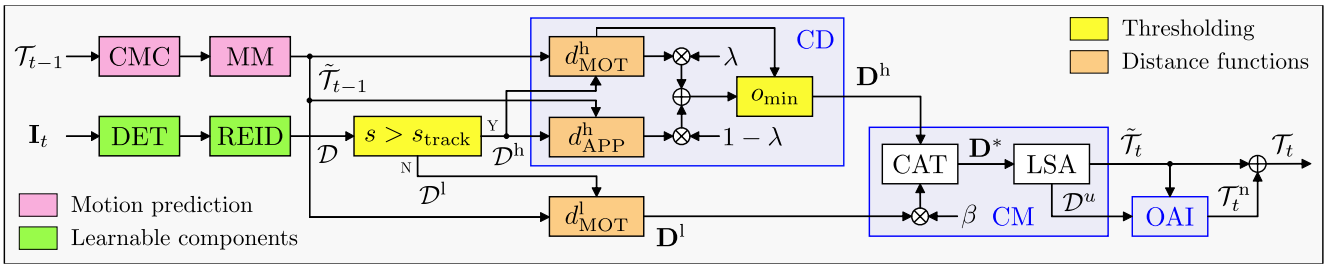


Figure 3. Overview of our tracking framework (own contributions in blue). Tracks from the previous frame $\mathcal{T}_{t-1}$ are predicted with camera motion compensation (CMC) and motion model (MM) yielding $\tilde{\mathcal{T}}_{t-1}$, while the detector (DET) is applied on the image $\mathbf{I}_t$ and a re-identification model (REID) extracts appearance features from the detected image regions. Those are saved next to the bounding boxes in the set of detections $\mathcal{D}$. For high-confident detections $\mathcal{D}^{\text{h}}$ with score $s > s_{\text{track}}$ and low-confident detections $\mathcal{D}^{\text{l}}$ with $s < s_{\text{track}}$, different distances d to the predicted tracks $\tilde{\mathcal{T}}_{t-1}$ are calculated. The proposed Combined Distance (CD) of motion ($d^{\text{h}}_{\text{MOT}}$) and appearance distance ($d^{\text{h}}_{\text{APP}}$) from Equation (4) is leveraged for $\mathcal{D}^{\text{h}}$, while only motion distance ($d^{\text{l}}_{\text{MOT}} = 1 - \text{IoU}$) is used for $\mathcal{D}^{\text{l}}$. The resulting two distance matrices $\mathbf{D}^{\text{h}}$ and $\mathbf{D}^{\text{l}}$ are fused to $\mathbf{D}^*$ according to Equation (1) in our Combined Matching (CM), where CAT denotes concatenation. After solving the linear sum assignment (LSA) problem, the unmatched detections $\mathcal{D}^{\text{u}}$ are compared with updated tracks $\tilde{\mathcal{T}}_t$ in our Occlusion Aware Initialization (OAI) module. Finally, new tracks $\mathcal{T}^{\text{n}}_t$ are added to the updated tracks yielding the final set of tracks $\mathcal{T}_t$.

nential moving average (EMA) manner [9, 35]. Third, we propose to build a feature bank as in [36] but compute a mean track feature by averaging the features of the feature bank and then use this mean track feature for comparison to the features of the current detections.

3.3. Occlusion Aware Initialization

To prevent the initialization of ghost tracks from single false positive detections (FP), some approaches [1,3,5,9,42] first start *tentative* tracks from unmatched detections $\{D^{\text{u}}\}$ that have to be assigned a detection in n_{init} consecutive frames before activation. Despite removing FP, this strategy introduces $n_{\text{init}} - 1$ false negatives for each correct track because tentative tracks do not contribute to the online tracking results. We argue that most of the FP in multi-person tracking are duplicate detections when a decent detector is used, especially in crowded scenes where it is difficult for the detector to reason about object boundaries and duplicate detections can easily occur. Therefore, we propose an Occlusion Aware Initialization (OAI) technique that is capable of identifying duplicate detections and discards them from starting new tracks. For each unmatched detection D^{u}, the IoU with all updated tracks $\{\tilde{T}\}$ is calculated and if the maximum IoU exceeds the overlap threshold o_{max}, the detection is removed. The OAI is illustrated for two updated tracks and an unmatched duplicate detection in Figure 4.

3.4. Proposed Tracking Framework

The interplay of the proposed tracking components (Sections 3.1 to 3.3) is shown in Figure 3, where an overview of our tracking framework is given. For an accurate modelling of target and camera movement, we apply the NSA Kalman filter [8] and the camera motion compensation from [28], respectively. Following the current state-of-the-art (SOTA)

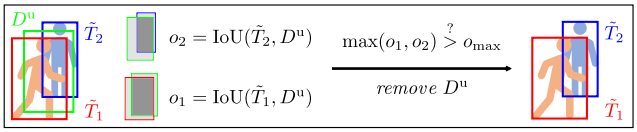


Figure 4. Illustration of Occlusion Aware Initialization (OAI). Typically, unmatched detections like D^{u} initialize new tracks independent from its surroundings. In contrast, we calculate the IoU between unmatched detections and updated tracks $\{\tilde{T}\}$ and compare the maximum IoU with a predefined threshold o_{max}. If the overlap with an already tracked target is too high, the detection D^{u} is removed arguing that it is probably a duplicate detection.

methods in multi-person tracking [1, 5, 9, 21, 27, 33, 34, 42], YOLOX [11] is deployed as detector. The re-identification model is adopted from BoT-SORT [1] because this tracker is most similar to our approach among the SOTA methods. To further enhance the final tracking results, the Gaussian Smoothed Interpolation (GSI) from [9] is leveraged.

4. Experiments

We first specify implementation details in Section 4.1. Then, we briefly describe the utilized datasets and evaluation measures in Section 4.2. The results of our ablative experiments are analyzed in Section 4.3, before a comparison with the SOTA is made in Section 4.4.

4.1. Implementation Details

We use a YOLOX [11] detector in our tracking-by-detection framework which was trained on the detection datasets CrowdHuman [25], CityPersons [41], and ETH [10], as well as the tracking datasets MOT17 [20] and MOT20 [6]. For ablative experiments on MOT17 train and evaluation on the test sets, we adopt different model

weights provided by the authors of ByteTrack [42]. Unless otherwise stated, the input resolution of the images is (1400×800). The minimum score s_{min} and maximum IoU o_{NMS} for filtering detections in the non-maximum suppression (NMS) are $s_{\text{min}} = 0.1$ and $o_{\text{NMS}} = 0.7$, respectively. The detections are separated into $\mathcal{D}^{\text{h}}$ and $\mathcal{D}^{\text{l}}$ with $s_{\text{track}} = 0.6$ and a threshold of $s_{\text{init}} = 0.7$ is applied for initialization of tracks in the ablative experiments. For OAI, the maximum overlap o_{max} is empirically set to 0.55. The best configuration of CD from Equation (4) uses DIoU for d_{MOT}, $\lambda = 0.2$, $o_{\text{min}} = 0.1$, and a maximum distance $d^{\text{h}}_{\text{max}} = 0.65$. For calculation of d_{APP}, the size of the feature bank is $n_{\text{feat}} = 15$. The maximum IoU distance for low-confident detections is empirically set to $d^{\text{l}}_{\text{max}} = 0.19$ which yields $\beta = 3.42$ according to Equation (2). To enable re-activation after occlusion, inactive tracks are kept for 35 frames without assigned detection before termination. The interpolation parameter τ in GSI [9] is empirically set to 12.

4.2. Evaluation Datasets and Metrics

We follow the common practice to divide the MOT17 train split and use its second half as validation set for ablative experiments [1, 9, 24, 28, 32, 37, 42, 45]. Both train and test split contain 7 sequences for multi-person tracking with complex scenarios including many occlusions, camera motion, and scenes at day and night. The MOT20 dataset is, despite containing no camera motion, even more challenging comprising very crowded scenes (127 vs. 21 persons per image [26]) – 4 for train and test each. Comparison with the SOTA on the test sets is done by submitting the tracking results to the evaluation server (*https://motchallenge.net/*) because the annotations are not publicly available.

With TrackEval [17], we compute HOTA [18] as our main evaluation measure for tracking performance as it takes association, detection, and localization accuracy equally into account. In addition, we report MOTA [2] and IDF1 [23], as well as the number of false positives (FP), false negatives (FN), and identity switches (IDSW).

4.3. Ablation Study

We first conduct experiments with different association measures including our Combined Distance (CD) and various strategies for calculating appearance distance. Then, our Combined Matching (CM) is compared with two stage matching and the importance of matching low-confident detections to inactive tracks (L2I) is investigated. Afterwards, our Occlusion Aware Initialization (OAI) is explored. Finally, we study the impact of our tracking components on the final performance and examine different interpolation methods to post-process our tracking results.

Distance measures. To analyze the effectiveness of different motion- and appearance-based distances, we run baseline experiments with only one matching stage. The results are depicted in Table 1. For a fair comparison, the maximum matching distance d_{max} is tuned for each configuration separately. When using only motion distance d_{MOT}, DIoU achieves slightly better results than the standard IoU, while Mahalanobis distance (Mahal) performs much worse which confirms our claim that IoU based distances are the better choice. Combining appearance distance d_{APP} with IoU by taking the minimum as in BoT-SORT [1], only a small gain of 0.1 HOTA is reached although the cosine distance (Cos) is additionally masked to prevent wrong assignments. Fusing d_{APP} with DIoU using a weighted sum as in Equation (3) and setting $\lambda = 0.2$, HOTA is significantly enhanced by 0.9 points showing the high potential of appearance information when integrated correctly. Note that $\lambda = 0.2$ means that 4 times the weight is given to d_{APP} compared to d_{MOT}. Enforcing a minimum IoU requirement o_{min} in our CD from Equation (4) yields further decent improvements (+0.4 HOTA). Our CD outperforms previous fusion approaches from BoT-SORT [1] and JDE [35] by 1.3 and 1.1 HOTA, respectively.

Table 1. Different distance combinations (one stage matching).

d_{MOT}	d_{APP}	Comb.	λ	o_{min}	HOTA	MOTA	IDF1
Mahal	-	-	-	-	65.4	75.4	76.4
IoU	-	-	-	-	68.9	77.1	81.8
DIoU	-	-	-	-	69.0	77.1	82.0
IoU	M. Cos *	Min	-	-	69.0	77.4	81.8
Mahal	Cos **	Eq. (3)	0.02	-	69.2	76.7	82.6
IoU	Cos	Eq. (3)	0.5	-	69.2	77.4	82.3
DIoU	Cos	Eq. (3)	0.5	-	69.4	**77.5**	82.6
DIoU	Cos	Eq. (3)	0.2	-	69.9	77.2	83.4
DIoU	Cos	Eq. (4)	0.2	0.1	**70.3**	77.3	**84.1**

* Masked cosine distance from BoT-SORT [1]. ** JDE [35].

Table 2. Calculation of appearance distance (one stage matching).

Method	Min dist. [36]	Mean dist.	Mean feat.	EMA [35]
HOTA	69.5	69.9	**70.3**	**70.3**

Appearance distance calculation. Besides the way how d_{APP} and d_{MOT} are combined, how d_{APP} is calculated is also important. Various strategies are examined in Table 2. Using the *minimum* cosine distance from all features of a feature bank as done in DeepSORT [36] gives the worst results. Taking the *mean* distance performs better. However, instead of averaging the distances, it is beneficial to average the features and calculate the cosine distance w.r.t. the mean feature. The same is done within the EMA technique despite that an exponential moving average in place of a simple average is used. Not surprisingly, similar results are obtained. Since the simple mean feature performs slightly better in our final tracker (+0.1 HOTA w.r.t. EMA), we use it in all other experiments.

Table 3. Two stage matching (TSM) vs. Comb. Matching (CM).

	HOTA	MOTA	IDF1		HOTA	MOTA	IDF1
TSM	71.2	78.3	85.3	CM	**71.5**	**78.5**	**85.8**

Table 4. Matching of low-confident detections to inactive tracks (L2I). ByteTrack is analyzed with and without CMC. Slightly different results w.r.t. the papers stem from re-implementation.

Tracker	CMC	L2I	HOTA	MOTA	IDF1
ByteTrack [42]	✗	✗	67.9	77.8	80.1
ByteTrack [42]	✗	✓	67.9	78.1	79.9
ByteTrack [42]	✓	✗	69.0	78.3	82.2
ByteTrack [42]	✓	✓	69.3	**78.7**	82.6
BoT-SORT [1]	✓	✗	69.5	78.2	82.9
BoT-SORT [1]	✓	✓	70.2	78.6	83.3
Ours	✓	✗	71.1	78.2	84.9
Ours	✓	✓	**71.5**	78.5	**85.8**

Table 5. Ablation of our Occlusion Aware Initialization (OAI).

OAI	n_{init}	$o_{\max}$	HOTA	MOTA	IDF1
✗	1	-	71.1	78.3	85.2
✗	2	-	71.1	78.3	85.2
✗	3	-	71.0	78.2	85.2
✓	1	0.5	71.4	**78.5**	**85.9**
✓	1	0.55	**71.5**	**78.5**	85.8
✓	1	0.6	71.4	**78.5**	85.7
✓	2	0.55	71.4	**78.5**	85.8

Matching strategy. As already discussed in Section 3.1, our Combined Matching (CM) has the advantage of considering all possible assignments simultaneously which is not the case in two stage matching (TSM). The resulting improved tracking performance is shown quantitatively in Table 3 and qualitatively in Figure 1, where tracking results are visualized both for TSM and CM.

Low-confident detections to inactive tracks. Another benefit of our CM is that it allows low-confident detections $\mathcal{D}^{\text{l}}$ to be matched with inactive tracks (L2I). We apply L2I in our tracking framework and also in ByteTrack [42] and BoT-SORT [1] by utilizing inactive tracks in the second matching stage. Results are summarized in Table 4. In ByteTrack without camera motion compensation (CMC), L2I yields no improvements, however, when integrating CMC, L2I enhances HOTA by 0.3 points. We hypothesize that this is because inactive tracks have inaccurately predicted locations when camera motion is not compensated and matching those to $\mathcal{D}^{\text{l}}$ – that can also be inaccurate – harms association accuracy. A plus of 0.7 and 0.4 HOTA is obtained for BoT-SORT and our framework, respectively, indicating that L2I improves the association when good motion models are used. A qualitative example, where L2I prevents an IDSW is shown in Figure 5.

(a) Without L2I. (b) With L2I.

Figure 5. Qualitative example of L2I. (a) When not allowing low-confident detections to be matched to inactive tracks (dashed lines), hard-to-detect targets, *e.g.*, due to occlusion, can get lost when the inactive patience is over. As a consequence, a new track is started (blue) when a high-confident detection is available again. (b) Matching low-confident detections to inactive tracks (L2I) improves identity preservation for occluded targets.

(a) Without OAI. (b) With OAI.

Figure 6. Qualitative example of OAI. (a) Without OAI, a duplicate detection missed by NMS starts a new track (red) which causes an ID switch. The inactivated track is indicated with dashed lines. (b) In OAI, the track information is leveraged to remove unmatched detections with high overlaps to existing tracks. Thus, no duplicate track is started and the ID switch is prevented.

Track initialization. After the association, high-confident unmatched detections are considered for track initialization. A common strategy is to first start tentative tracks that become active after n_{init} consecutive assignments. Differently, our OAI allows a detection to start a track only if no overlaps above $o_{\max}$ with other tracks exist. The strategies are compared in Table 5. In our tracker, no improvements are obtained for the baseline initialization, whereas OAI enhances HOTA up to 0.4 points for $o_{\max} = 0.55$. Especially identity preservation is improved with a plus of 0.6 IDF1. A qualitative example, where OAI successfully removes a duplicate detection and as a consequence prevents an IDSW can be found in Figure 6.

Component analysis. To investigate the impact of the proposed tracking modules, we add one component after another and evaluate the tracking performance in Table 6. The first row is a baseline with two stage matching and IoU distance. All of our tracking modules improve the overall performance, especially the association accuracy. Compared to the baseline, HOTA, MOTA, and IDF1 are increased by 2.3, 0.5, and 3.6 points, respectively.

Table 6. Impact of our tracking components.

CD	L2I	CM	OAI	HOTA	MOTA	IDF1
✗	✗	✗	✗	69.2	78.0	82.2
✓	✗	✗	✗	70.7	78.3	84.6
✓	✓	✗	✗	70.8	78.3	85.1
✓	✓	✓	✗	71.1	78.3	85.2
✓	✓	✓	✓	**71.5**	**78.5**	**85.8**

Table 7. Different interpolation variants as post-processing.

Type	Max gap	Min length	HOTA	MOTA	IDF1
-	-	-	71.5	78.5	85.8
LI	✗	✗	73.2	81.1	87.2
LI	✓	✗	73.3	81.1	87.3
LI	✓	✓	73.4	81.4	87.5
GSI	✓	✓	**73.7**	**81.7**	**87.7**

Table 8. State-of-the-art methods on MOT17.

Method	MOTA	IDF1	HOTA	FP	FN	IDSW
RTU++ [33]	79.5	79.1	63.9	29508	84618	1302
StrongSORT [9]	79.6	79.5	64.4	27876	86205	1194
SAT [34]	80.0	79.8	64.4	25125	86505	1356
ByteTrack [42]	80.3	77.3	63.1	25491	83721	2196
QuoVadis [7]	80.3	77.7	63.1	25491	83721	2103
FOR [21]	80.4	77.7	63.6	28674	79452	2298
BoT-SORT [1]	80.5	80.2	65.0	**22521**	86037	1212
BYTEv2 [27]	80.6	78.9	63.6	35208	73224	1239
C-BIoU [40]	81.1	79.7	64.1	23136	82011	1455
ImprAsso	**82.2**	**82.1**	**66.4**	26727	**72666**	**924**

Track interpolation. To further improve the results, we evaluate two interpolation approaches for fragmented tracks as post-processing: a simple linear interpolation (LI) and the Gaussian Smoothed Interpolation (GSI) from [9]. In addition, two constraints are applied. Gaps with length larger than 30 frames and tracks which comprise less than 30 detections are not interpolated. The results are listed in Table 7. Even a simple LI boosts HOTA by 1.7 points indicating that our tracker is capable of successfully bridging a lot of occlusions. The two constraints for maximum gap length and minimum track length further improve HOTA by 0.1 points each. Finally, GSI yields another plus of 0.3 HOTA.

4.4. Comparison with the State-of-the-Art

We propose different tracking components to improve the association accuracy in multi-person tracking which is why we term our tracker *ImprAsso* (Improved Association). A comparison of ImprAsso with the state-of-the-art (SOTA) methods on MOT17 is given in Table 8. Note that we follow recent approaches [1, 5, 42] and apply various thresholds for track initialization s_{init} among the sequences, while setting $s_{\text{track}} = s_{\text{init}} - 0.1$. For reproducibility, we specify this parameter in Table 9. On MOT17, ImprAsso surpasses previous approaches significantly with a gain of 1.1 MOTA, 2.4 IDF1, and 2.3 HOTA w.r.t. to the second best entry C-BIoU [40]. Moreover, the least number of IDSW is obtained indicating a superior association performance.

Table 9. s_{init} on MOT17 (left) and MOT20 (right) test.

01	03	06	07	08	12	14	04	06	07	08
0.8	0.75	0.75	0.7	0.7	0.8	0.55	0.7	0.4	0.7	0.4

Table 10. State-of-the-art methods on MOT20.

Method	MOTA	IDF1	HOTA	FP	FN	IDSW
SAT [34]	75.0	76.6	62.6	**15549**	113136	**816**
OC-SORT [5]	75.7	76.3	62.4	19067	105894	942
RTU++ [33]	76.5	76.8	62.8	19247	101290	971
FOR [21]	76.8	76.4	61.4	27112	91254	1443
BYTEv2 [27]	77.3	75.6	61.4	22867	93409	1082
ReMOT [39]	77.4	73.1	61.2	28351	86659	1789
ByteTrack [42]	77.8	75.2	61.3	26249	87594	1223
QuoVadis [7]	77.8	75.7	61.5	26249	87594	1187
BoT-SORT [1]	77.8	77.5	63.3	24638	88863	1313
ImprAsso	**78.6**	**78.8**	**64.6**	27064	**82715**	992

On MOT20, we adopt the input resolution from [1,5,42], that is $(1600, 896)$ for sequences 04 and 07 and $(1920, 736)$ for 06 and 08. We also adopt the initialization thresholds s_{init} which are given in Table 9 for reproducibility. Results of the SOTA methods on the challenging MOT20 dataset are summarized in Table 10. ImprAsso outperforms all approaches including the previous best method BoT-SORT [1] with improvements of 0.8 MOTA, 1.3 IDF1, and 1.3 HOTA.

5. Conclusion

In this work, we propose several tracking modules to improve the association accuracy in multi-person tracking. A novel combined distance of motion and appearance information is introduced that significantly outperforms previous fusion approaches and various strategies for calculating motion and appearance distance are explored. Moreover, it is shown that leveraging inactive tracks in the second matching stage can enhance the performance of different trackers and that our combined matching approach improves the utilization of low-confident detections compared to the basic two stage matching. Despite that, a new track initialization technique is proposed which prevents the start of ghost tracks from duplicate detections under occlusion. The effectiveness of our components is shown with extensive ablative experiments and putting all together, our tracker surpasses the state-of-the-art on the MOT17 and MOT20 benchmarks.

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[2] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP J. Image Video Process.*, 2008, 2008. 6

[3] Alex Bewley, ZongYuan Ge, Lionel Ott, Fabio Tozeto Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE Int. Conf. Image Process.*, pages 3464–3468, 2016. 1, 2, 3, 5

[4] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *IEEE Int. Conf. Adv. Video Sign. Surveillance*, 2017. 1, 2, 3

[5] Jinkun Cao, Xinshuo Weng, Rawal Khirodkar, Jiangmiao Pang, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. *arXiv preprint arXiv:2203.14360*, 2022. 1, 2, 3, 5, 8

[6] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 2, 5

[7] Patrick Dendorfer, Vladimir Yugay, Aljoša Ošep, and Laura Leal-Taixé. Quo vadis: Is trajectory forecasting the key towards long-term multi-object tracking? In *Adv. Neural Inform. Process. Syst.*, 2022. 8

[8] Yunhao Du, Junfeng Wan, Yanyun Zhao, Binyu Zhang, Zhihang Tong, and Junhao Dong. Giaotracker: A comprehensive framework for mcmot with global information and optimizing strategies in visdrone 2021. In *Int. Conf. Comput. Vis. Workshp.*, pages 2809–2819, 2021. 5

[9] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Trans. Multimedia*, 2023. 1, 2, 3, 4, 5, 6, 8

[10] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. A mobile vision system for robust multi-person tracking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2008. 5

[11] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 5

[12] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *J. Basic Engineering*, 82(1):35–45, 1960. 2

[13] Harold William Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. 4

[14] Jiaxin Li, Yan Ding, and Hualiang Wei. Simpletrack: Rethinking and improving the jde approach for multi-object tracking. *Sensors*, 22(15), 2022. 2, 3

[15] Chao Liang, Zhipeng Zhang, Yi Lu, Xue Zhou, Bing Li, Xiyong Ye, and Jianxiao Zou. Rethinking the competition between detection and reid in multiobject tracking. *IEEE Trans. Image Process.*, 31:3182–3196, 2020. 2

[16] Qiankun Liu, Qi Chu, Bin Liu, and Nenghai Yu. Gsm: Graph similarity model for multi-object tracking. In *IJCAI*, pages 530–536, 2020. 2

[17] Jonathon Luiten and Arne Hoffhues. Trackeval. https://github.com/JonathonLuiten/TrackEval, 2020. 6

[18] Jonathon Luiten, Aljoša Ošep, 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.*, 129(2):548–578, 2021. 6

[19] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8844–8854, 2022. 3

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

[21] Mohammad Hossein Nasseri, Mohammadreza Babaee, Hadi Moradi, and Reshad Hosseini. Online relational tracking with camera motion suppression. *J. Vis. Commun. Image Represent.*, 90, 2023. 1, 3, 5, 8

[22] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 658–666, 2019. 2

[23] 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. Worksh.*, pages 17–35, 2016. 6

[24] Chaobing Shan, Chunbo Wei, Bing Deng, Jianqiang Huang, Xiansheng Hua, Xiaoliang Cheng, and Kewei Liang. Tracklets predicting based adaptive graph tracking. *arXiv preprint arXiv:2010.09015*, 2020. 6

[25] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 5

[26] Daniel Stadler and Jürgen Beyerer. On the performance of crowd-specific detectors in multi-pedestrian tracking. In *IEEE Int. Conf. Adv. Video Sign. Surveillance*, 2021. 6

[27] Daniel Stadler and Jürgen Beyerer. Bytev2: Associating more detection boxes under occlusion for improved multi-person tracking. In *Int. Conf. Pattern Recog. Workshp.*, 2022. 1, 2, 3, 5, 8

[28] Daniel Stadler and Jürgen Beyerer. Modelling ambiguous assignments for multi-person tracking in crowds. In *IEEE Wint. Conf. Applications Comp. Vis. Works.*, pages 133–142, 2022. 5, 6

[29] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3701–3710, 2017. 1, 2

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, pages 5998–6008, 2017. 3

[31] Paul Voigtlaender, Michael Krause, Aljoša Ošep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2

[32] Qiang Wang, Yun Zheng, Pan Pan, and Yinghui Xu. Multiple object tracking with correlation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3876–3886, 2021. 6

[33] Shuai Wang, Hao Sheng, Da Yang, Yang Zhang, Yubin Wu, and Sizhe Wang. Extendable multiple nodes recurrent tracking framework with rtu++. *IEEE Trans. Image Process.*, 31:5257–5271, 2022. 5, 8

[34] Shuai Wang, Da Yang, Yubin Wu, Yang Liu, and Hao Sheng. Tracking game: Self-adaptative agent based multi-object tracking. In *ACM Int. Conf. Multimedia*, pages 1964–1972, 2022. 5, 8

[35] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *Eur. Conf. Comput. Vis.*, pages 107–122, 2020. 2, 3, 4, 5, 6

[36] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *IEEE Int. Conf. Image Process.*, pages 3645–3649, 2017. 1, 2, 3, 4, 5, 6

[37] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12352–12361, 2021. 6

[38] Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. Spatial-temporal relation networks for multi-object tracking. In *Int. Conf. Comput. Vis.*, pages 3987–3997, 2019. 1, 2

[39] Fan Yang, Xin Chang, Sakriani Sakti, Yang Wu, and Satoshi Nakamura. Remot: A model-agnostic refinement for multiple object tracking. *Image Vis. Comp.*, 106, 2021. 8

[40] Fan Yang, Shigeyuki Odashima, Shoichi Masui, and Shan Jiang. Hard to track objects with irregular motions and similar appearances? Make it easier by buffering the matching space. In *IEEE Winter Conf. Applications Comput. Vis.*, pages 4799–4808, 2023. 1, 3, 8

[41] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4457–4465, 2017. 5

[42] 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.*, pages 1–21, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[43] 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.*, 129:3069–3087, 2021. 2

[44] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *AAAI Conf. Artific. Intel.*, pages 12993–13000, 2020. 2, 4

[45] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *Eur. Conf. Comput. Vis.*, pages 474–490, 2020. 6