Supplementary Material for Focusing on Tracks for Online Multi-Object Tracking

Kyujin Shim Kangwook Ko Yujin Yang Changick Kim
Korea Advanced Institute of Science and Technology (KAIST)
{kjshim1028, kokangook623, ujin.y, changick}@kaist.ac.kr

1. Method

1.1. Distance Function

The distance function $c(T_i, d_j)$ between track T_i and detection result d_j is defined as:

$$c(T_i, d_j) = (c_{HMIoU}(T_i, d_j) + c_{App}(T_i, d_j) + \lambda_1 c_{Conf}(T_i, d_j) + \lambda_2 c_{Ang}(T_i, d_j)) \ / \ 2, \quad (1)$$

where c_{HMIoU}, c_{App}, c_{Conf}, and c_{Ang} are HMIoU, cosine, confidence, and angular distance, respectively, similar to Hybrid-SORT [7], and λ_1 and λ_2 are universally set to 0.2 and 0.1, respectively, for every dataset.

1.2. Flowchart and Pseudocode

For clarity, we provide the overall pipeline of our Track-Track in Fig. 1 and the pseudocode for Track-Perspective-Based Association (TPA) in the algorithm 1. As illustrated in Fig. 1, tracks are categorized as confirmed and unconfirmed tracks. We utilize high-confidence detection results, low-confidence detection results, and high-confidence detection results removed during NMS. TPA matches these detection results with confirmed tracks, as described in the algorithm 1, with an iterative association procedure. Then, unmatched high-confidence detection results are matched with unconfirmed tracks. Finally, matched tracks are updated, and Track-Aware Initialization (TAI) uses these matched tracks along with unmatched detection results to initialize new tracks.

2. Experiments

2.1. Datasets

MOT17 [3] consists of seven sequences each for training and testing, while MOT20 [2] contains four sequences each for training and testing. DanceTrack [6] comprises 40, 25, and 35 sequences for training, validation, and testing, respectively. During the validation with MOT17, we utilized its four training videos for training and the remaining three training videos for validation to avoid object duplication between the training and validation splits.

Algorithm 1 Track-Perspective-Based Association

Require: Existing tracks $\mathcal{T}$; Detection results $\mathcal{D} = \mathcal{D}_{high} \cup \mathcal{D}_{low} \cup \mathcal{D}_{del}$; Kalman Filter `KF`;
Ensure: Matched pairs $\mathcal{M}$

Initialization: $\mathcal{M} \leftarrow \emptyset$
# Predict the next location for each track
for $i \leftarrow 1 : len(\mathcal{T})$ **do**
  $T_i \leftarrow \texttt{KF}(T_i)$
end for

# Cost matrix calculation
for $i \leftarrow 1 : len(\mathcal{T})$ **do**
  for $j \leftarrow 1 : len(\mathcal{D})$ **do**
    if $d_j \in \mathcal{D}_{high}$ **then**
      $\mathbf{C}_{ij} = c(T_i, d_j)$
    else if $d_j \in \mathcal{D}_{low}$ **then**
      $\mathbf{C}_{ij} = c(T_i, d_j) + \tau_p$
    else if $d_j \in \mathcal{D}_{del}$ **then**
      $\mathbf{C}_{ij} = c(T_i, d_j) + \tau_q$
    end if
  end for
end for

$\mathcal{T}' \leftarrow \mathcal{T}, \mathcal{D}' \leftarrow \mathcal{D}, \mathbf{C}' \leftarrow \mathbf{C}$

# Iterative association
while $\max(\mathbf{C}') < \tau_m$ **do**
  $\mathcal{M}' \leftarrow \{(T_i, d_j) | T_i = \text{argmin}_{T_l \in \mathcal{T}'} C'_{lj}, d_j = \text{argmin}_{d_k \in \mathcal{D}'} C'_{ik}, C'_{ij} < \tau_m\}$
  $\mathcal{M} \leftarrow \mathcal{M} \cup \mathcal{M}'$
  $\mathcal{T}' \leftarrow \mathcal{T}' \setminus \{T_i \mid (T_i, d) \in \mathcal{M}'\}$
  $\mathcal{D}' \leftarrow \mathcal{D}' \setminus \{d_j \mid (T, d_j) \in \mathcal{M}'\}$
  $\mathbf{C}'_{ij} = \mathbf{C}_{ij}$ for all $T_i \in \mathcal{T}', d_j \in \mathcal{D}'$
  $\tau_m \leftarrow \tau_m - r$
end while
Return: $\mathcal{M}$

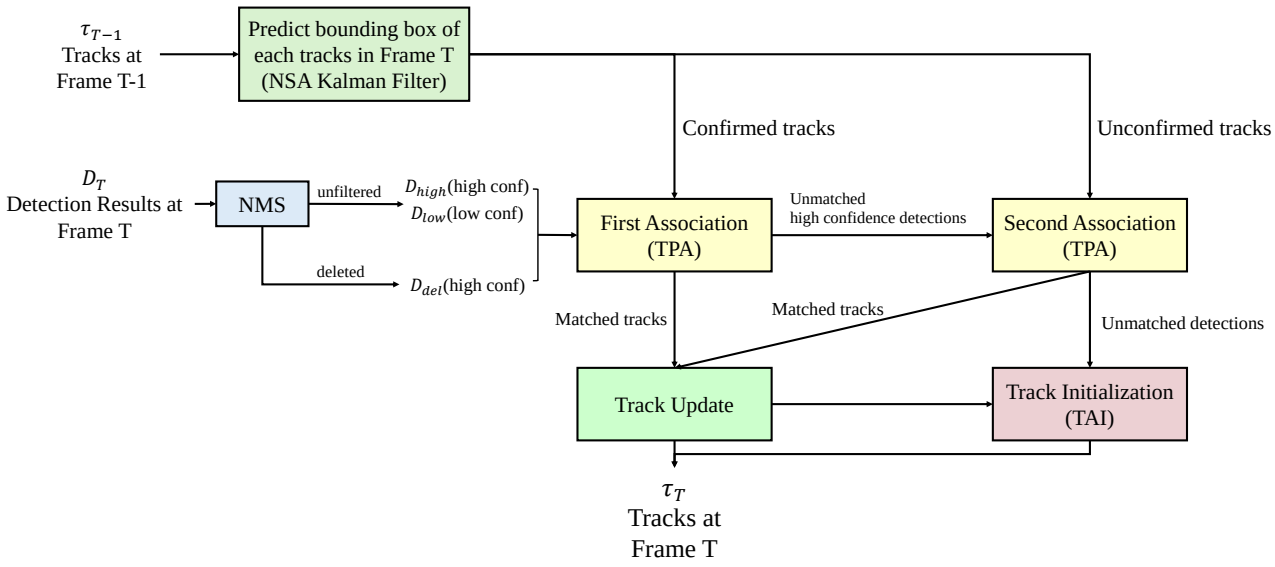


Figure 1. Overall algorithm of our proposed TrackTrack. Given the previous tracks, the bounding box locations for the current frame are predicted using the NSA Kalman filter. The tracks are divided into confirmed and unconfirmed tracks, where confirmed tracks are the ones that tracked three or more frames, and unconfirmed tracks are the ones that tracked less than three frames. We utilize high-confidence detection results, low-confidence detection results, and high-confidence detection results that are deleted during NMS steps. Based on the Track-Perspective Association (TPA), the detection results and confirmed tracks are matched. Then, unmatched high-confidence detection results are matched with unconfirmed tracks, also based on TPA. The matched tracks are updated, and their current location is utilized by Track-Aware Initialization (TAI) along with unmatched detection results to initialize new tracks.

2.2. Implementation Details

The penalty terms τ_p and τ_q of the TPA strategy and the NMS IoU threshold in TAI were set as 0.20, 0.40, and 0.55, respectively. The matching threshold τ_m is set as 0.70, 0.55, and 0.60 for MOT17 [3], MOT20 [2], and DanceTrack [6], respectively, for the test evaluation. The reduction term r is s set as 0.05. The detection confidence threshold for distinguishing high and low-confidence detection results is set similarly to the previous works [5, 8] while detailed settings are shared in our GitHub codes [1].

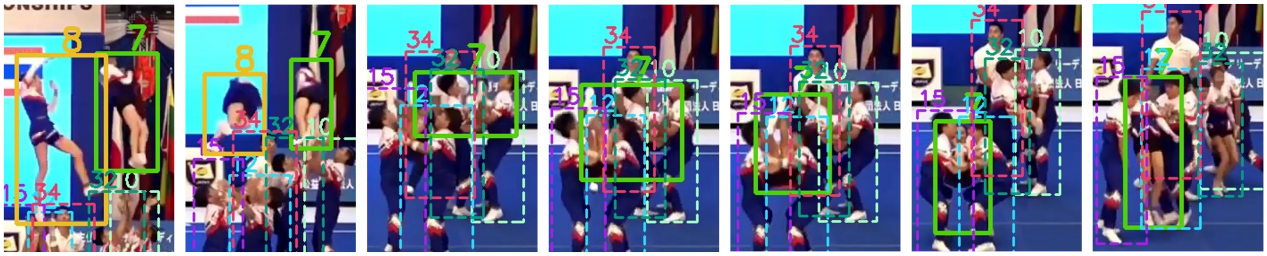
2.3. Qualitative Results

Our TrackTrack demonstrates significant improvements in robustness compared to a baseline tracker using the Hungarian algorithm and multi-stage association scheme, as illustrated in Fig. 2. In challenging scenarios, such as handling detection noise, recovering from occlusions, and resolving overlaps, TrackTrack consistently outperforms the baseline. For instance, our tracker maintains correct identity consistency when the detection noise causes the baseline to swap IDs 10 and 89. The reduced threshold of each iteration suppresses the matching between the detection noise and tracks. Similarly, after an occlusion where the baseline misassigned ID 56 as ID 61, TrackTrack successfully recovered the original identity of the object. Again, with severe overlaps where the baseline switches IDs 6 and 3, our method ensures consistent identity tracking. This is the result of the association considering all detections simultaneously, providing robustness in occlusion. These results highlight the robustness of our TrackTrack, which effectively addresses common challenges in online multi-object tracking through the novel association process.

Despite the improvements of TrackTrack, some failure cases persist. The common reason is that the features of the tracks are not clear due to rapid movement or severe occlusion so it is hard to prevent the ID switch or the effect of detection noises. In Fig. 3 (a), ID 8 is lost, and ID 7 undergoes an identity switch during its rapid movement. The drastic changes in position and posture make detection challenging, leading to low similarity scores with existing tracks. Similarly, in Fig. 3 (b), our tracker loses ID 22 during a severe occlusion. These cases highlight the challenge of maintaining stable associations under extreme motion and occlusion conditions, suggesting the need for further refinements in the tracking algorithm. Also, despite our advanced methods, some detection noises still affect the tracking performance. In Fig. 3 (c), the original track with ID 15 confuses the object due to the detection noises, and another track takes up the role of ID 15. Future research will further explore strategies to effectively suppress detec-

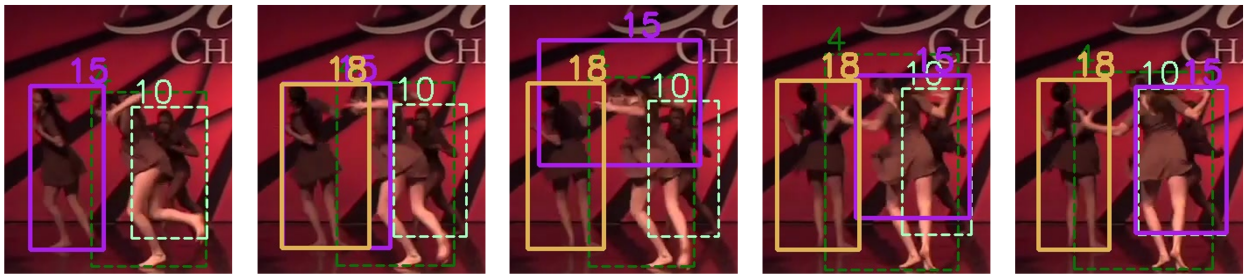
[1] https://github.com/kamkyu94/TrackTrack

Figure 2. Qualitative comparisons between the baseline tracker, which uses the Hungarian algorithm and multi-stage association scheme, and our TrackTrack with the novel assignment and single-stage association methods. In the results from the baseline tracker, (a) ID 10 and 89 are switched because of the detection noise, (b) ID 56 is changed to 61 after occlusion, and (c) ID 6 and 3 are switched after the overlap. In contrast, our TrackTrack shows correct tracking results in every case, demonstrating its robustness against detection noise and occlusions.



(a)

(b)



(c)

Figure 3. Examples of failure cases. In (a), ID 8 is changed to 7 after dynamic movements, and in (b), the person who was designated ID 22 is re-designated as ID 40, and the track ID 22 is no longer following a meaningful target. In (c), the person who was designated ID 15 is re-designated as ID 18, and track ID 15 is following detection noises.

tion noise, enhancing the overall robustness of the tracking system.

2.4. Ablation Studies

We provide additional details for the ablative results shown in the main paper by adding more comparing metrics. Also, we performed ablation studies to examine the influence of the penalty terms τ_p, τ_q, reduction term r, and the IoU threshold during NMS in Track-Aware Initialization (TAI).

2.4.1. Component Ablation

Table 1 shows the detailed results of component ablation. As in the table, each strategy remarkably improves almost every score in both datasets.

2.4.2. Assignment Method

In Table 2, we can confirm the detailed comparison of our proposed assignment method against the traditional Hungarian algorithm. The results demonstrate the superiority of our algorithm.

2.4.3. Association Stage

A detailed ablative result to demonstrate the effectiveness of our joint association scheme within TPA is presented in Table 3. The result indicates that the joint scheme outperforms the multi-stage association strategy in most metrics.

2.4.4. Using Deleted Detection Results

In Table 4, a detailed evaluation result for the impact of utilizing $\mathcal{D}_{\text{del}}$ is shown. The result reveals that incorporation of $\mathcal{D}_{\text{del}}$ enhances tracking performance in most cases.

2.4.5. Penalty Term

An ablative study to assess the impact of the penalty terms τ_p and τ_q in our Track-Perspective-Based Association (TPA) strategy is shown in Table 5. The result reveals that our tracker achieves enough performance on MOT17

		MOT17-val					DanceTrack-val				
TPA	TAI	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑
		67.3	83.2	78.3	69.6	65.5	61.9	65.4	**92.6**	47.7	80.7
✓		68.5	84.6	78.8	72.0	65.5	62.9	66.2	92.5	49.1	**80.9**
	✓	67.4	83.3	78.8	69.6	65.8	62.6	66.6	**92.6**	48.7	**80.9**
✓	✓	**69.1**	**85.1**	**79.5**	**72.7**	**66.0**	**63.3**	**66.8**	92.5	**49.7**	80.8

Table 1. A detailed ablative study for our proposed strategies.

	MOT17-val					DanceTrack-val				
Assignment	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑
Hungarian	68.1	83.9	**79.6**	70.5	**66.2**	61.6	64.8	92.2	47.1	**80.9**
Ours	**69.1**	**85.1**	79.5	**72.7**	66.0	**63.3**	**66.8**	**92.5**	**49.7**	80.8

Table 2. Detailed comparisons between using the Hungarian algorithm and our assignment method that considers local matching precision.

	MOT17-val					DanceTrack-val				
Association	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑
Multi-Stage	68.5	84.2	79.1	71.8	65.9	63.2	**67.5**	**92.5**	**49.8**	80.6
Joint (Ours)	**69.1**	**85.1**	**79.5**	**72.7**	**66.0**	**63.3**	66.8	**92.5**	49.7	**80.8**

Table 3. Detailed comparisons between the multi-stage cascade association of each $\mathcal{D}_{\text{high}}$, $\mathcal{D}_{\text{low}}$, and $\mathcal{D}_{\text{del}}$ similar to the previous works [1, 7, 8] and joint association of all detection results as we proposed.

	MOT17-val					DanceTrack-val				
Use $\mathcal{D}_{\text{del}}$	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑
✗	68.6	85.0	**79.5**	71.9	65.9	62.1	65.1	**92.6**	47.6	**81.3**
✓	**69.1**	**85.1**	**79.5**	**72.7**	**66.0**	**63.3**	**66.8**	92.5	**49.7**	80.8

Table 4. A detailed ablative study for the influence of utilizing the deleted detection results $\mathcal{D}_{\text{del}}$ in our tracking process.

and the highest HOTA score on DanceTrack when we set τ_p and τ_q to 0.20 and 0.40, respectively. Therefore, we use τ_p of 0.20 and τ_q of 0.40 as our default setting for our TrackTrack to optimize the association process by effectively penalizing unlikely matches without overly constraining the association flexibility.

2.4.6. Reduction Term

Table 6 presents an ablative study about the reduction term r in TPA. The result demonstrates that the optimal performance is achieved when r is set to 0.05, resulting in the highest HOTA, IDF1, and AssA scores on both datasets. Thus, we use r of 0.05 as our default setting, effectively reducing the influence of weaker associations while maintaining substantial tracking accuracy and association consistency for each matching iteration.

2.4.7. IoU threshold during NMS in TAI

An ablative study is conducted to assess the influence of different IoU thresholds during non-maximum suppression (NMS) in our Track-Aware Initialization (TAI) strategy. The result, presented in Table 7, indicates that the highest performance on both datasets is observed in most metrics when the IoU threshold is set to 0.55. Therefore, we use the IoU Threshold of 0.55 as the default setting in our TAI since it provides the optimal balance between maintaining accurate track initialization and avoiding excessive suppression of potential detections, thereby maximizing overall tracking accuracy.

2.5. Computational Cost of a Full Pipeline

Table 4 presents the computational efficiency of our TrackTrack compared to other state-of-the-art trackers, Deep OC-SORT [4] and Hybrid-SORT-ReID [7], on the MOT17-val and DanceTrack-val datasets. The results show the frame-per-second (FPS) of the entire pipeline, including object detection, feature extraction, and tracking, for each tracker. As in the table, our TrackTrack achieves the highest FPS on both datasets, indicating superior efficiency. These results confirm that our method achieves robust tracking performance while maintaining high computational efficiency.

		MOT17-val					DanceTrack-val				
τ_p	τ_q	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑
0.15	0.35	69.2	85.1	79.2	73.0	**66.1**	62.7	66.5	92.3	49.4	80.1
0.15	0.40	**69.3**	**85.2**	79.1	**73.1**	**66.1**	63.2	66.7	92.4	**49.9**	80.5
0.15	0.45	68.8	85.0	79.1	72.3	66.0	62.3	66.0	92.5	48.5	80.5
0.20	0.35	69.1	85.1	**79.5**	72.7	66.0	62.1	65.7	92.4	48.1	80.7
0.20	0.40	69.1	85.1	**79.5**	72.7	66.0	**63.3**	**66.8**	92.5	49.7	80.8
0.20	0.45	68.6	85.0	**79.5**	71.9	65.9	61.9	65.0	92.6	47.5	**81.0**
0.25	0.35	68.4	84.3	79.1	71.6	65.8	62.9	66.2	92.5	49.2	80.8
0.25	0.40	68.4	84.3	79.2	71.6	65.8	63.1	**66.8**	92.6	49.6	80.6
0.25	0.45	68.3	84.5	79.2	71.5	65.7	62.4	65.6	**92.7**	48.3	80.9

Table 5. An ablative study for the penalty terms τ_p and τ_q in TPA.

	MOT17-val					DanceTrack-val				
r	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑
0.025	69.0	**85.1**	**79.7**	72.6	**66.1**	62.0	65.3	92.2	48.1	80.2
0.050	**69.1**	**85.1**	79.5	**72.7**	66.0	**63.3**	**66.8**	92.5	**49.7**	80.8
0.075	68.5	84.9	79.5	71.8	65.9	62.0	65.2	**92.7**	47.8	**80.9**

Table 6. An ablative study for the reduction term r in our TPA.

	MOT17-val					DanceTrack-val				
IoU Thr.	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑	HOTA↑	IDF1↑	MOTA↑	AssA↑	DetA↑
0.45	68.8	84.8	79.1	72.5	65.7	63.2	66.8	92.4	49.6	**80.8**
0.50	68.9	84.9	79.2	**72.7**	65.7	**63.3**	**66.9**	**92.5**	**49.8**	**80.8**
0.55	**69.1**	**85.1**	**79.5**	**72.7**	**66.0**	**63.3**	66.8	**92.5**	49.7	**80.8**
0.60	**69.1**	**85.1**	**79.5**	**72.7**	**66.0**	63.2	66.8	**92.5**	49.6	**80.8**
0.65	68.7	84.8	79.3	72.1	65.8	63.2	66.7	**92.5**	49.6	80.9

Table 7. An ablative study for the IoU threshold during NMS in our Track-Aware Initialization (TAI).

	MOT17-val	DanceTrack-val
Tracker	FPS↑	FPS↑
Deep OC-SORT [4]	8.36	14.02
Hybrid-SORT-ReID [7]	7.59	14.23
TrackTrack	9.22	14.53

Table 8. Computational costs of a full pipeline, including object detection, feature extraction, and tracking, of other state-of-the-art methods [4, 7] and our TrackTrack. Each value represents the frame per second (FPS) of the corresponding case.

(The evaluation server is changed due to an internal matter, so the FPS values here are much lower than the results of the main paper.)

2.6. Integration with E2E Detectors

TPA can adapt to end-to-end frameworks, such as DETR, by treating all predicted boxes equally during the association step. For example, if we compare TPA and the multi-stage & Hungarian-based association scheme while using only certain detection results ($D_{high} + D_{low}$ in our setting), HOTA scores are 68.6 and 67.6, respectively, with MOT17-val. The replacement with better detectors would enhance the performances, and our approaches would also show a synergy effect and contribute to the overall performance based on its diverse advantages, such as better assignment over the Hungarian algorithm. We will add the related discussion and the integration results with the end-to-end frameworks in our supplementary material as soon as possible and release it on the GitHub repository [2].

References

[1] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort

[2] https://github.com/kamkyu94/TrackTrack

for robust multi-object tracking. In *CVPR*, pages 9686–9696, 2023. 5

[2] 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:2003.09003*, 2020. 1, 2

[3] Patrick Dendorfer, Aljoša Ošep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *IJCV*, 129:845–881, 2021. 1, 2

[4] Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. In *ICIP*, 2023. 5, 6

[5] Daniel Stadler and Jürgen Beyerer. An improved association pipeline for multi-person tracking. In *CVPRW*, pages 3170–3179, 2023. 2

[6] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *CVPR*, pages 20993–21002, 2022. 1, 2

[7] Mingzhan Yang, Guangxin Han, Bin Yan, Wenhua Zhang, Jinqing Qi, Huchuan Lu, and Dong Wang. Hybrid-sort: Weak cues matter for online multi-object tracking. *AAAI*, 38(7): 6504–6512, 2024. 1, 5, 6

[8] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, pages 1–21, 2022. 2, 5