# Supplementary Material
# No Train Yet Gain: Towards Generic Multi-Object Tracking in Sports and Beyond

Tomasz Stanczyk[1,2] Seongro Yoon[1,2] Francois Bremond[1,2]
[1] Inria, France [2] Université Côte d'Azur, France
tomasz.stanczyk@inria.fr, seong-ro.yoon@inria.fr, francois.bremond@inria.fr

This supplementary material contains the following appendices as referred in the main paper:

- A More experiments and details with mask-based tracking systems
- B State-of-the-art comparison with transformer-based and other types of method
- C Additional visual examples
- D The running speed and heaviness of mask

## A. More experiments and details with mask-based tracking systems

We evaluate DEVA [3], Grounded SAM 2 [15, 21], and MASA [13] on MOT datasets, saving each bounding box output per frame in MOT format [6].

We conduct additional experiments to thoroughly explore the performance differences between the mask-based tracking systems and our McByte. These include several variants on the MOT17 [19] validation set, as well as experiments on the DanceTrack [22] validation set, analogous to the ones presented in the main paper.

Tab. 1 presents various experimental variants on the MOT17 validation set, where different detectors and parameters are used. The variants marked with ‡ correspond to those discussed in the main paper on SportsMOT [5].

For DEVA, we first run the default settings using the Grounding Dino [15] detector with the "person" prompt and a confidence threshold of 0.35 to accept bounding boxes. Then, we replace it with the YOLOX [10] detector, trained on the MOT17 dataset from our baseline [27]. We test two threshold values, 0.6 and 0.7. In our baseline, initialization of the new tracklets happens for the values 0.1 higher than the high confidence detection threshold. As we consider the default value of 0.6 for the latter (Sec. 4.1 in the main paper), we also experiment with the value of 0.7 with DEVA and other mask based systems.

For Grounded SAM 2 [15, 21], we use the "Video Object Tracking with Continuous ID" version as specified on its GitHub page[1]. Initially, we run it with the original settings, using the Grounding Dino [15] detector with the "person" prompt, a confidence detection threshold of 0.25, and a step value of 20. The step value defines how often detections are processed (e.g., every 20th frame) to create mask tracklets, functioning as the segment length (we refer to tracking objects in segments mentioned in the main paper, Sec. 2.3). We then test an analogous variant with a step value of 100.

Next, we integrate YOLOX detector with weights from our baseline [27] and run variants with step values of 20, 100, and 1 (thus processing detections every frame), using different bounding box allowance thresholds of 0.25, 0.6, and 0.7 (analogous to the DEVA experiments). We also attempt to run a variant with the segment length set to the entire video sequence, but it fails due to excessive GPU memory requirements. Additionally, this setup would only track objects visible in the first frame.

MASA [13] offers several models for inference. We test variants using two different feature backbones: GroundingDINO [15] (GDino) and ResNet-50 [11] (R50). For the GroundingDINO variant, we use the Detic-SwinB detector [16, 30] with the "person" prompt, applying the original detection confidence threshold of 0.2. We also run a similar variant with the YOLOX detector trained on the COCO [14] dataset, as provided by the authors, using a confidence threshold of 0.3, default for this variant.

Further, we incorporate the YOLOX detector with weights from our baseline [27] and test variants with detection confidence thresholds of 0.3, 0.6, and 0.7, analogously to DEVA and Grounded SAM 2. Additionally, we run the ResNet-50 feature variants with the YOLOX COCO model (threshold 0.3) and the baseline-pre-trained weights (thresholds 0.3, 0.6, 0.7).

As shown in Tab. 1, McByte outperforms the referenced mask-based systems, making it more suitable for MOT.

Tab. 2 presents the performance of DEVA, Grounded

[1] https://github.com/IDEA-Research/Grounded-SAM-2

| Details | HOTA | IDF1 | MOTA |
|---|---|---|---|
| DEVA | | | |
| GDino ”person”, th. 0.35 ‡ | **31.8** | **31.3** | **-89.4** |
| YOLOX ByteTrack, th. 0.6 ‡ | 24.7 | 20.4 | -239.7 |
| YOLOX ByteTrack, th. 0.7 | 27.0 | 23.7 | -187.8 |
| Grounded SAM 2 | | | |
| GDino ”person”, th. 0.25, step 20 ‡ | 43.4 | 47.6 | 18.4 |
| GDino ”person”, th. 0.25, step 100 | 44.0 | 49.0 | 15.5 |
| YOLOX ByteTrack, th. 0.25, step 20 | 46.4 | 51.6 | 36.0 |
| YOLOX ByteTrack, th. 0.6, step 20 ‡ | **47.5** | 54.1 | 43.0 |
| YOLOX ByteTrack, th. 0.7, step 20 | 47.4 | 54.1 | 44.3 |
| YOLOX ByteTrack, th. 0.25, step 100 | 46.8 | 54.2 | 30.2 |
| YOLOX ByteTrack, th. 0.6, step 100 | 47.4 | **54.9** | 34.8 |
| YOLOX ByteTrack, th. 0.7, step 100 | 47.4 | **54.9** | 35.9 |
| YOLOX ByteTrack, th. 0.25, step 1 | 43.0 | 43.9 | 36.2 |
| YOLOX ByteTrack, th. 0.6, step 1 | 44.4 | 46.5 | 44.9 |
| YOLOX ByteTrack, th. 0.7, step 1 | 44.3 | 46.7 | **46.5** |
| MASA | | | |
| GDino feat. Detic-SwinB ”person”, th 0.2 | 46.8 | 52.1 | 24.3 |
| GDino feat. YOLOX COCO, th 0.3 | 45.4 | 53.1 | 36.9 |
| GDino feat. YOLOX ByteTrack, th 0.3 | 61.8 | 70.8 | 71.3 |
| GDino feat. YOLOX ByteTrack, th 0.6 | 63.4 | 73.3 | 73.8 |
| GDino feat. YOLOX ByteTrack, th 0.7 | 62.5 | 71.9 | 72.9 |
| R50 feat. YOLOX COCO, th 0.3 ‡ | 45.5 | 53.6 | 36.9 |
| R50 feat. YOLOX ByteTrack, th 0.3 | 62.5 | 72.0 | 71.5 |
| R50 feat. YOLOX ByteTrack, th 0.6 ‡ | **63.5** | **73.6** | **74.0** |
| R50 feat. YOLOX ByteTrack, th 0.7 | 62.6 | 72.3 | 73.0 |
| McByte | | | |
| McByte (ours) | **69.9** | **82.8** | **78.5** |

Table 1. Extended comparison with the other tracking methods using segmentation mask: DEVA [3], Grounded SAM 2 [12, 15] and MASA [13] on MOT17 validation set [19], while changing their parameters. ‡ denotes the variants reported in the main paper and in Tab. 2.

| Method | HOTA | IDF1 | MOTA |
|---|---|---|---|
| DEVA, original settings | 21.9 | 15.8 | -347.1 |
| DEVA, with YOLOX | 20.1 | 13.3 | -423.9 |
| Grounded SAM 2, original settings | 51.3 | 48.0 | 73.5 |
| Grounded SAM 2, with YOLOX | 52.9 | 49.6 | 81.6 |
| MASA, original settings | 38.2 | 34.9 | 71.9 |
| MASA, with YOLOX | 46.0 | 41.1 | 85.6 |
| McByte (ours) | **62.3** | **64.0** | **89.8** |

Table 2. Comparison with the other tracking methods using segmentation mask: DEVA [3], Grounded SAM 2 [12, 15] and MASA [13] on DanceTrack validation set [22]. The reported variants correspond to the variants with ‡ symbol in Tab. 1

SAM 2, and MASA on the DanceTrack [22] validation set. The listed variants correspond to those marked with ‡ in Tab. 1 and are the ones reported in the main paper on SportsMOT.

On DanceTrack, McByte also demonstrates significantly higher performance, reinforcing its effectiveness and suitability for MOT.

## B. State-of-the-art comparison with transformer-based and other types of method

There exist MOT methods outside the tracking-by-detection domain manifesting performance differences, but usually these methods are not directly comparable, because they require a lot of training data and might use other detections. Further, they make certain hypotheses, e.g. global optimization on the whole video. At the same time, these methods might perform visibly worse on some benchmarks as we discuss below. On the contrary, we stress that McByte performs well on all the discussed benchmarks (Secs. 4.3 and 4.4 of the main paper). McByte is a tracking-by-detection approach, which is the main focus of our work. For an additional reference, though, we also list performance of the transformer-based, global optimization, and joint detection and tracking methods.

Tabs. 3 to 5 show extended comparison including other

| Method | HOTA | IDF1 | MOTA |
|---|---|---|---|
| Transformer-based | | | |
| MeMOTR [8] | 70.0 | 71.4 | 91.5 |
| MOTIP [9] | 71.9 | 75.0 | 92.9 |
| Joint detection and tracking | | | |
| FairMOT [26] | 49.3 | 53.5 | 86.4 |
| CenterTrack [29] | 62.7 | 60.0 | 90.8 |
| Tracking-by-detection | | | |
| ByteTrack [27] | 64.1 | 71.4 | 95.9 |
| MixSort-Byte [5] | 65.7 | 74.1 | 96.2 |
| OC-SORT [1] | 73.7 | 74.0 | 96.5 |
| MixSort-OC [5] | 74.1 | 74.4 | 96.5 |
| GeneralTrack [20] | 74.1 | 76.4 | 69.8 |
| DiffMOT [17] | 76.2 | 76.1 | 97.1 |
| McByte (ours) | 76.9 | 77.5 | 97.2 |

Table 3. Extended state-of-the-art method comparison on SportsMOT [5] test set.

| Method | HOTA | IDF1 | MOTA |
|---|---|---|---|
| Transformer-based | | | |
| MOTR [25] | 57.8 | 68.6 | 73.4 |
| MeMOTR [8] | 58.8 | 71.5 | 72.8 |
| MOTRv2 [28] | 62.0 | 75.0 | 78.6 |
| MOTIP [9] | 59.2 | 71.2 | 75.5 |
| Global optimization | | | |
| SUSHI [2] | 66.5 | 83.1 | 81.1 |
| Joint detection and tracking | | | |
| FairMOT [26] | 59.3 | 72.3 | 73.7 |
| RelationTrack [24] | 61.0 | 75.8 | 75.6 |
| CenterTrack [29] | 52.2 | 64.7 | 67.8 |
| Tracking-by-detection with parameter tuning per sequence | | | |
| ByteTrack [27] | 63.1 | 77.3 | 80.3 |
| MixSort-Byte [5] | 64.0 | 78.7 | 79.3 |
| StrongSORT++ [7] | 64.4 | 79.5 | 79.6 |
| OC-SORT [1] | 63.2 | 77.5 | 78.0 |
| MixSort-OC [5] | 63.4 | 77.8 | 78.9 |
| Deep OC-SORT [18] | 64.9 | 80.6 | 79.4 |
| Hybrid-SORT [23] | 64.0 | 78.7 | 79.9 |
| Tracking-by-detection without parameter tuning per sequence | | | |
| ByteTrack [2] | 62.8 | 77.1 | 78.9 |
| GeneralTrack [20] | 64.0 | 78.3 | 80.6 |
| DiffMOT [17] | 64.2 | 79.3 | 79.8 |
| McByte (ours) | 64.2 | 79.4 | 80.2 |

Table 4. Extended state-of-the-art method comparison on MOT17 [19] test set.

types of tracking methods based on the result availability. All the tracking-by-detection methods use the same object detector models per dataset.

Tab. 3 presents extended state-of-the-art comparison on

| Method | HOTA | IDF1 | MOTA |
|---|---|---|---|
| Transformer-based | | | |
| MOTR [25] | 54.2 | 51.5 | 79.7 |
| MeMOTR [8] | 63.4 | 65.5 | 85.4 |
| MOTRv2 [28] | 73.4 | 76.0 | 92.1 |
| MOTIP [9] | 67.5 | 72.2 | 90.3 |
| Global optimization | | | |
| SUSHI [2] | 63.3 | 63.4 | 88.7 |
| Joint detection and tracking | | | |
| FairMOT [26] | 39.7 | 40.8 | 82.2 |
| CenterTrack [29] | 41.8 | 35.7 | 86.8 |
| Tracking-by-detection | | | |
| ByteTrack [27] | 47.7 | 53.9 | 89.6 |
| MixSort-Byte [5] | 46.7 | 53.0 | 85.5 |
| OC-SORT [1] | 55.1 | 54.9 | 92.2 |
| StrongSORT++ [7] | 55.6 | 55.2 | 91.1 |
| Hybrid-SORT [23] | 65.7 | 67.4 | 91.8 |
| GeneralTrack[20] | 59.2 | 59.7 | 91.8 |
| DiffMOT [17] | 63.4 | 64.0 | 92.7 |
| McByte (ours) | 67.1 | 68.1 | 92.9 |

Table 5. Extended state-of-the-art method comparison on DanceTrack [22] test set.

SportsMOT [5] test set. In this dataset, the number of subjects can vary as due to abrupt camera motion, subjects can continuously enter and leave the scene. Further, due to the team sport nature, there are many occlusions and blur among the tracked objects. Transformer-based methods cannot handle all the mentioned challenges and perform lower than most of the tracking-by-detection approaches, including ours. Joint detection and tracking methods generalize poorly to this dataset and fall behind the other two types of tracking methods. Our method can handle the challenges present in the sport settings and outperforms all the other methods.

Tab. 4 shows extended state-of-the art comparison on MOT17 [19] test set. Note that analogously to the main paper, we also put the result of ByteTrack [27] not being tuned per sequence as reported in [2] (*"ByteTrack [2]"*). Transformer-based methods perform visibly lower than the tracking-by-detection methods (including ours) as they struggle with the subjects frequently entering and leaving the scene. In contrast, SUSHI [2], which is a powerful global optimization approach, reaches highly satisfying performance. However, it accesses all the video frames at the same time while processing detections and associating the tracklets, which makes it impossible to run in online settings. Current state-of-the-art joint detection and tracking methods generally perform lower than the tracking-by-detection methods. In that paradigm, the detection and association step is performed jointly. In our method, we perform these two steps separately and focus on the association part.

Tab. 5 presents extended state-of-the-art comparison on

DanceTrack [22] test set. As in this dataset the subjects remain mostly at the scene, the transformer-based methods performance is more satisfying. The performance of transformer-based methods can be both higher [9, 28] or lower [8, 25] compared to the the tracking-by-detection methods. For similar reasons, the global optimization method, SUSHI [2] can also perform higher than the other tracking-by-detection methods on this dataset, or lower, e.g. when compared to our method. On this dataset, joint detection and tracking methods also manifest lower performance than the tracking-by-detection methods.

## C. Additional visual examples

We provide full frame inputs and outputs of the examples used in the main paper, see Figs. 1 and 2 in this supplementary material. We also provide a larger version of one figure from the main paper, see Fig. 3.

In the main paper, we discuss that McByte can handle challenging scenarios due to the temporally propagated mask signal used in the controlled manner as an association cue (Sec. 3.3). Fig. 4 in this supplementary material shows another example of our method handling association of ambiguous boxes, improving over the baseline. Fig. 5 shows an example of our method handling longer occlusions in the crowd.

## D. The running speed and heaviness of mask

The running speed of McByte oscillates around 3-5 FPS over the datasets examined [4, 5, 19, 22] on a single A100 GPU. It is more costly compared to the baseline [27] and other derived methods, but McByte is more reliable - it generalizes well on 4 different datasets and we do not tune it per dataset or per sequence. We believe that it is a good tradeoff. Mask-based tracking is a promising concept and we believe it will be further optimized in the community.

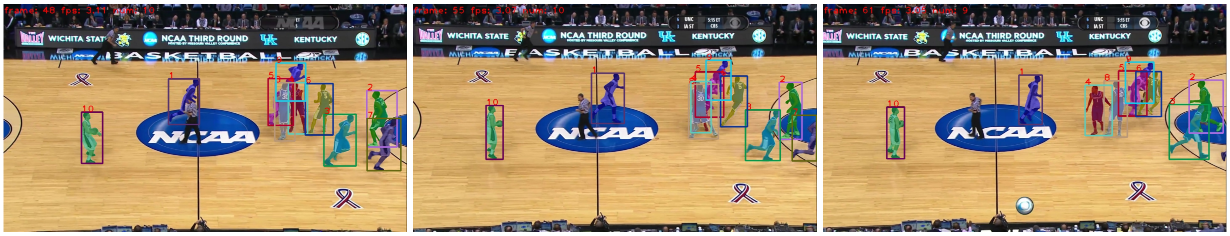


Figure 1. Full output frames corresponding to Fig. 1 from the main paper. Input image data from [5].

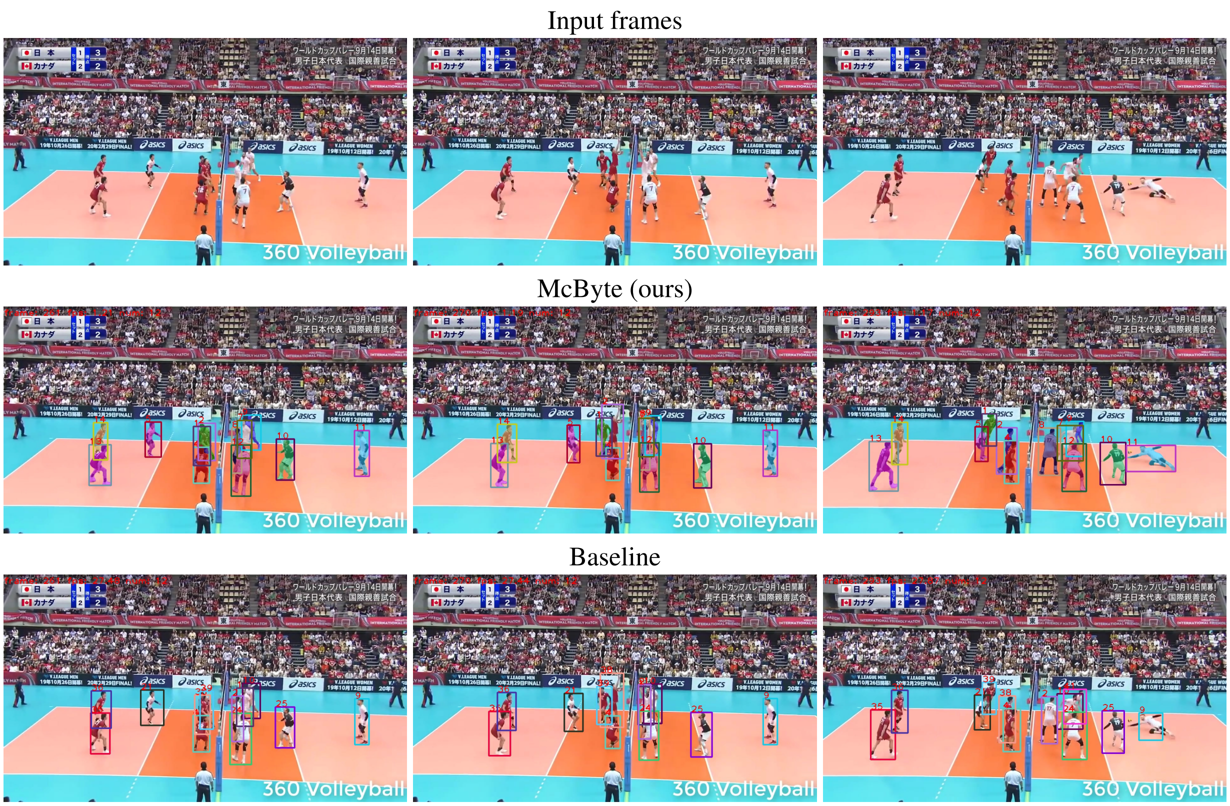
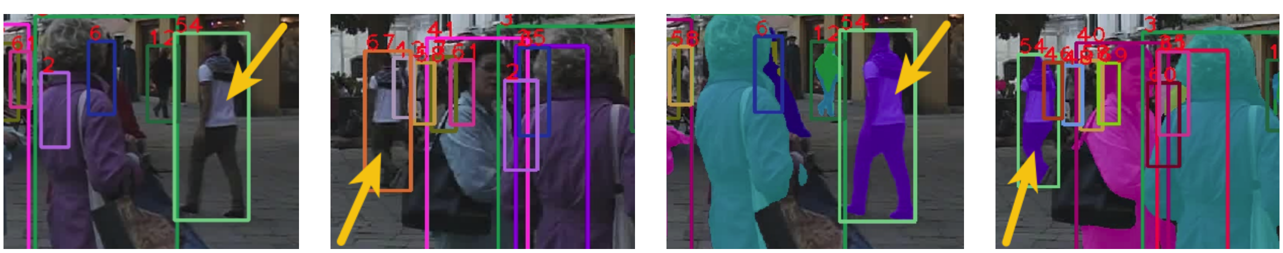



Figure 2. Full input and output frames corresponding to Fig. 4 from the main paper. Input image data from [5].


Figure 3. Larger version of Fig. 5 from the main paper. Input image data from [5].

Frame 459 (baseline) | Frame 475 (baseline) | Frame 459 (McByte) | Frame 475 (McByte)

Figure 4. Visual output comparison between the baseline and McByte. With the temporally propagated mask guidance, McByte can handle the association of an ambiguous set of bounding boxes - see the subjects with IDs 59 and 63 on the output of McByte. Input image data from [19].



Frame 319 (baseline) | Frame 401 (baseline) | Frame 319 (McByte) | Frame 401 (McByte)

Figure 5. Visual output comparison between the baseline and McByte. With the temporally propagated mask guidance, McByte can handle longer occlusion in the crowd - see the subject with ID 54 on the output of McByte. Input image data from [19].

## References


[1] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9686–9696, 2023. 3

[2] Orcun Cetintas, Guillem Brasó, and Laura Leal-Taixé. Unifying short and long-term tracking with graph hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22877–22887, 2023. 3, 4

[3] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 2

[4] Anthony Cioppa, Silvio Giancola, Adrien Deliege, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3502, 2022. 4

[5] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 3, 4, 5

[6] Patrick Dendorfer, Aljossa Ossep, Anton Milan, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixe. Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision*, 129:1–37, 2021. 1

[7] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 2023. 3

[8] Ruopeng Gao and Limin Wang. MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9901–9910, 2023. 3, 4

[9] Ruopeng Gao, Yijun Zhang, and Limin Wang. Multiple object tracking as id prediction, 2024. 3, 4

[10] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE, 2016. 1

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[13] Siyuan Li, Lei Ke, Martin Danelljan, Luigi Piccinelli, Mattia Segu, Luc Van Gool, and Fisher Yu. Matching anything by segmenting anything. *CVPR*, 2024. 1, 2

[14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *CoRR*, abs/1405.0312, 2014. 1

[15] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1, 2

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[17] Weiyi Lv, Yuhang Huang, Ning Zhang, Ruei-Sung Lin, Mei Han, and Dan Zeng. Diffmot: A real-time diffusion-based multiple object tracker with non-linear prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19321–19330, 2024. 3

[18] Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. *arXiv preprint arXiv:2302.11813*, 2023. 3

[19] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, 2016. arXiv: 1603.00831. 1, 2, 3, 4, 6

[20] Zheng Qin, Le Wang, Sanping Zhou, Panpan Fu, Gang Hua, and Wei Tang. Towards generalizable multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19004, 2024. 3

[21] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[22] 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. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3, 4

[23] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6504–6512, 2024. 3

[24] En Yu, Zhuoling Li, Shoudong Han, and Hongwei Wang. Relationtrack: Relation-aware multiple object tracking with decoupled representation. *IEEE Transactions on Multimedia*, 25:2686–2697, 2022. 3

[25] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision (ECCV)*, 2022. 3, 4

[26] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087, 2021. 3

[27] 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. 2022. 1, 3, 4

[28] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pre-trained object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22056–22065, 2023. 3, 4

[29] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[30] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1