ByteTrack: Multi-Object Tracking by Associating Every Detection Box

Yifu Zhang[1], Peize Sun[2], Yi Jiang[3], Dongdong Yu[3], Fucheng Weng[1], Zehuan Yuan[3], Ping Luo[2], Wenyu Liu[1], and Xinggang Wang[1†]

[1] Huazhong University of Science and Technology
[2] The University of Hong Kong
[3] ByteDance Inc.

A Bounding box annotations

We note MOT17 [15] requires the bounding boxes [29] covering the whole body, even though the object is occluded or partly out of the image. However, the default implementation of YOLOX clips the detection boxes inside the image area. To avoid the wrong detection results around the image boundary, we modify YOLOX in terms of data pre-processing and label assignment. We do not clip the bounding boxes inside the image during the data pre-processing and data augmentation procedure. We only delete the boxes which are fully outside the image after data augmentation. In the SimOTA label assignment strategy, the positive samples need to be around the center of the object, while the center of the whole body boxes may lie out of the image, so we clip the center of the object inside the image.

MOT20 [7], HiEve [14] and BDD100K clip the bounding box annotations inside the image in and thus we just use the original setting of YOLOX.

B Tracking performance of light models

We compare BYTE and DeepSORT [22] using light detection models. We use YOLOX [10] with different backbones as our detector. All models are trained on CrowdHuman and the half training set of MOT17. The input image size is 1088×608 and the shortest side ranges from 384 to 832 during multi-scale training. The results are shown in Table 1. We can see that BYTE brings stable improvements on MOTA and IDF1 compared to DeepSORT, which indicates that BYTE is robust to detection performance. It is worth noting that when using YOLOX-Nano as backbone, BYTE brings 3 points higher MOTA than DeepSORT, which makes it more appealing in real applications.

C Ablation Studies on ByteTrack

Speed v.s. accuracy. We evaluate the speed and accuracy of ByteTrack using different size of input images during inference. All experiments use the same multi-scale training. The results are shown in Table 2. The input size during inference ranges from 512×928 to 800×1440. The running time of the detector ranges from 17.9 ms to 30.0 ms and the

Backbone	Params	GFLOPs	Tracker	MOTA↑	IDF1↑	IDs↓
YOLOX-M	25.3 M	118.7	DeepSORT	74.5	76.2	197
YOLOX-M	25.3 M	118.7	BYTE	75.3	77.5	200
YOLOX-S	8.9 M	43.0	DeepSORT	69.6	71.5	205
YOLOX-S	8.9 M	43.0	BYTE	71.1	73.6	224
YOLOX-Tiny	5.0 M	24.5	DeepSORT	68.6	72.0	224
YOLOX-Tiny	5.0 M	24.5	BYTE	70.5	72.1	222
YOLOX-Nano	0.9 M	4.0	DeepSORT	61.4	66.8	212
YOLOX-Nano	0.9 M	4.0	BYTE	64.4	68.4	161

Table 1. Comparison of BYTE and DeepSORT using light detection models on the MOT17 validation set.

Input size	MOTA↑	IDF1↑	IDs↓	Time (ms)
512×928	75.0	77.6	200	**17.9+4.0**
608×1088	75.6	76.4	212	21.8+4.0
736×1280	76.2	77.4	188	26.2+4.2
800×1440	**76.6**	**79.3**	**159**	29.6+4.2

Table 2. Comparison of different input sizes on the MOT17 validation set. The total running time is a combination of the detection time and the association time. The best results are shown in **bold**.

association time is all around 4.0 ms. ByteTrack can achieve 75.0 MOTA with 45.7 FPS running speed and 76.6 MOTA with 29.6 FPS running speed, which has advantages in practical applications.

Training data. We evaluate ByteTrack on the half validation set of MOT17 using different combinations of training data. The results are shown in Table 3. When only using the half training set of MOT17, the performance achieves 75.8 MOTA, which already outperforms most methods. This is because we use strong augmentations such as Mosaic [3] and Mixup [25]. When further adding CrowdHuman, Cityperson and ETHZ for training, we can achieve 76.7 MOTA and 79.7 IDF1. The big improvement of IDF1 arises from that the CrowdHuman dataset can boost the detector to recognize occluded person, therefore, making the Kalman Filter generate smoother predictions and enhance the association ability of the tracker.

The experiments on training data suggest that ByteTrack is not data hungry. This is a big advantage for real applications, comparing with previous methods [27,12,21,13] that require more than 7 data sources [15,9,26,23,28,8,18] to achieve high performance.

D Tracklet interpolation

We notice that there are some fully-occluded pedestrians in MOT17, whose visible ratio is 0 in the ground truth annotations. Since it is almost impossible to detect them by visual cues, we obtain these objects by tracklet interpolation.

Training data	Images	MOTA↑	IDF1↑	IDs↓
MOT17	2.7K	75.8	76.5	205
MOT17 + CH	22.0K	76.6	79.3	**159**
MOT17 + CH + CE	26.6K	**76.7**	**79.7**	183

Table 3. Comparison of different training data on the MOT17 validation set. "MOT17" is short for the MOT17 half training set. "CH" is short for the CrowdHuman dataset. "CE" is short for the Cityperson and ETHZ datasets. The best results are shown in **bold**.

Suppose we have a tracklet T, its tracklet box is lost due to occlusion from frame t_1 to t_2. The tracklet box of T at frame t_1 is $B_{t_1} \in \mathbb{R}^4$ which contains the top left and bottom right coordinate of the bounding box. Let B_{t_2} represent the tracklet box of T at frame t_2. We set a hyper-parameter σ representing the max interval we perform tracklet interpolation, which means tracklet interpolation is performed when $t_2 - t_1 \leq \sigma$, . The interpolated box of tracklet T at frame t can be computed as follows:

$$B_t = B_{t_1} + (B_{t_2} - B_{t_1}) \frac{t - t_1}{t_2 - t_1}, \tag{1}$$

where $t_1 < t < t_2$.

As shown in Table 4, tracklet interpolation can improve MOTA from 76.6 to 78.3 and IDF1 from 79.3 to 80.2, when σ is 20. Tracklet interpolation is an effective post-processing method to obtain the boxes of those fully-occluded objects. We use tracklet interpolation in the test sets of MOT17 [15], MOT20 [7] and HiEve [14] under the private detection protocol.

Interval	MOTA↑	IDF1↑	FP↓	FN↓	IDs↓
No	76.6	79.3	**3358**	9081	159
10	77.4	79.7	3638	8403	150
20	**78.3**	**80.2**	3941	7606	**146**
30	78.3	80.2	4237	**7337**	147

Table 4. Comparison of different interpolation intervals on the MOT17 validation set. The best results are shown in **bold**.

E Public detection results on MOTChallenge.

We evaluate ByteTrack on the test set of MOT17 [15] and MOT20 [7] under the public detection protocol. Following the public detection filtering strategy in Tracktor [1] and CenterTrack [29], we only initialize a new trajectory when its IoU with a public detection box is larger than 0.8. We do not use tracklet interpolation under the public detection protocol. As is shown in Table 5, ByteTrack outperforms other methods by

Tracker	MOTA↑	IDF1↑	HOTA↑	FP↓	FN↓	IDs↓
STRN [24]	50.9	56.0	42.6	25295	249365	2397
FAMNet [5]	52.0	48.7	-	14138	253616	3072
Tracktor++v2 [1]	56.3	55.1	44.8	**8866**	235449	1987
MPNTrack [4]	58.8	61.7	49.0	17413	213594	1185
LPC_MOT [6]	59.0	66.8	51.5	23102	206948	**1122**
Lif_T [11]	60.5	65.6	51.1	14966	206619	1189
CenterTrack [29]	61.5	59.6	48.2	14076	200672	2583
TMOH [20]	62.1	62.8	50.4	10951	201195	1897
ArTIST_C [17]	62.3	59.7	48.9	19611	191207	2062
QDTrack [16]	64.6	65.1	-	14103	182998	2652
SiamMOT [19]	65.9	63.3	-	18098	170955	3040
ByteTrack (ours)	**67.4**	**70.0**	**56.1**	9939	**172636**	1331

Table 5. Comparison of the state-of-the-art methods under the "public detector" protocol on MOT17 test set. The best results are shown in **bold**.

Tracker	MOTA↑	IDF1↑	HOTA↑	FP↓	FN↓	IDs↓
SORT [2]	42.7	45.1	36.1	27521	264694	4470
Tracktor++v2 [1]	52.6	52.7	42.1	**6930**	236680	1648
ArTIST_C [17]	53.6	51.0	41.6	7765	230576	1531
LPC_MOT [6]	56.3	62.5	49.0	11726	213056	1562
MPNTrack [4]	57.6	59.1	46.8	16953	201384	1210
TMOH [20]	60.1	61.2	48.9	38043	165899	2342
ByteTrack (ours)	**67.0**	**70.2**	**56.4**	9685	**160303**	**680**

Table 6. Comparison of the state-of-the-art methods under the "public detector" protocol on MOT20 test set. The best results are shown in **bold**.

a large margin on MOT17. For example, it outperforms SiamMOT by 1.5 points on MOTA and 6.7 points on IDF1. Table 6 shows the results on MOT20. ByteTrack also outperforms existing results by a large margin. For example, it outperforms TMOH [20] by 6.9 points on MOTA, 9.0 points on IDF1, 7.5 points on HOTA and reduce the identity switches by three quarters. The results under public detection protocol further indicate the effectiveness of our association method BYTE.

F Visualization results.

We show some visualization results of difficult cases which ByteTrack is able to handle in Figure 1. We select 6 sequences from the half validation set of MOT17 and generate the visualization results using the model with 76.6 MOTA and 79.3 IDF1. The difficult cases include occlusion (*i.e.* MOT17-02, MOT17-04, MOT17-05, MOT17-09, MOT17-13), motion blur (*i.e.* MOT17-10, MOT17-13) and small objects (*i.e.* MOT17-13). The pedestrian in the middle frame with red triangle has low detection score, which is obtained by our association method BYTE. The low score boxes not only decrease the number of missing detection, but also play an important role for long-range associa-

tion. As we can see from all these difficult cases, ByteTrack does not bring any identity switch and preserve the identity effectively.

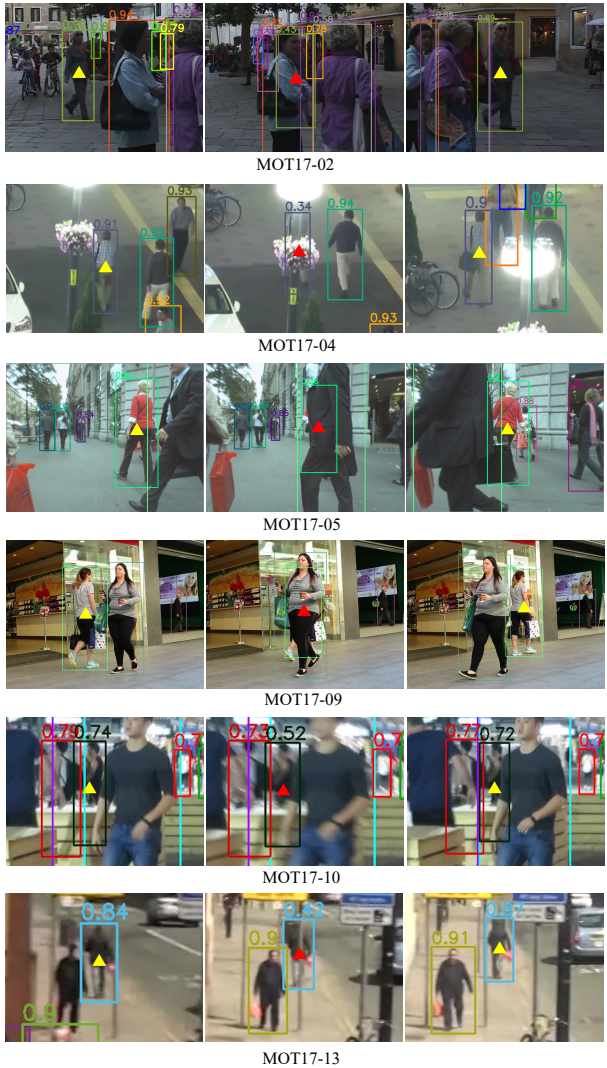

Fig. 1. Visualization results of ByteTrack. We select 6 sequences from the validation set of MOT17 and show the effectiveness of ByteTrack to handle difficult cases such as occlusion and motion blur. The yellow triangle represents the high score box and the red triangle represents the low score box. The same box color represents the same identity.

References

1. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: ICCV. pp. 941–951 (2019) 3, 4
2. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: ICIP. pp. 3464–3468. IEEE (2016) 4
3. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020) 2
4. Brasó, G., Leal-Taixé, L.: Learning a neural solver for multiple object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6247–6257 (2020) 4
5. Chu, P., Ling, H.: Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In: ICCV. pp. 6172–6181 (2019) 4
6. Dai, P., Weng, R., Choi, W., Zhang, C., He, Z., Ding, W.: Learning a proposal classifier for multiple object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2443–2452 (2021) 4
7. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixé, L.: Mot20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003 (2020) 1, 3
8. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: A benchmark. In: CVPR. pp. 304–311. IEEE (2009) 2
9. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: A mobile vision system for robust multi-person tracking. In: CVPR. pp. 1–8. IEEE (2008) 2
10. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021) 1
11. Hornakova, A., Henschel, R., Rosenhahn, B., Swoboda, P.: Lifted disjoint paths with application in multiple object tracking. In: International Conference on Machine Learning. pp. 4364–4375. PMLR (2020) 4
12. Liang, C., Zhang, Z., Lu, Y., Zhou, X., Li, B., Ye, X., Zou, J.: Rethinking the competition between detection and reid in multi-object tracking. arXiv preprint arXiv:2010.12138 (2020) 2
13. Liang, C., Zhang, Z., Zhou, X., Li, B., Lu, Y., Hu, W.: One more check: Making" fake background" be tracked again. arXiv preprint arXiv:2104.09441 (2021) 2
14. Lin, W., Liu, H., Liu, S., Li, Y., Qian, R., Wang, T., Xu, N., Xiong, H., Qi, G.J., Sebe, N.: Human in events: A large-scale benchmark for human-centric video analysis in complex events. arXiv preprint arXiv:2005.04490 (2020) 1, 3
15. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016) 1, 2, 3
16. Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 164–173 (2021) 4
17. Saleh, F., Aliakbarian, S., Rezatofighi, H., Salzmann, M., Gould, S.: Probabilistic tracklet scoring and inpainting for multiple object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14329–14339 (2021) 4
18. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018) 2
19. Shuai, B., Berneshawi, A., Li, X., Modolo, D., Tighe, J.: Siammot: Siamese multi-object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12372–12382 (2021) 4

20. Stadler, D., Beyerer, J.: Improving multiple pedestrian tracking by track management and occlusion handling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10958–10967 (2021) 4
21. Wang, Q., Zheng, Y., Pan, P., Xu, Y.: Multiple object tracking with correlation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3876–3886 (2021) 2
22. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE international conference on image processing (ICIP). pp. 3645–3649. IEEE (2017) 1
23. Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: CVPR. pp. 3415–3424 (2017) 2
24. Xu, J., Cao, Y., Zhang, Z., Hu, H.: Spatial-temporal relation networks for multi-object tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3988–3998 (2019) 4
25. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017) 2
26. Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: CVPR. pp. 3213–3221 (2017) 2
27. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision **129**(11), 3069–3087 (2021) 2
28. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q.: Person re-identification in the wild. In: CVPR. pp. 1367–1376 (2017) 2
29. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. In: European Conference on Computer Vision. pp. 474–490. Springer (2020) 1, 3, 4