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], Xinggang Wang[1†]

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

Abstract

Multi-object tracking (MOT) aims at estimating bounding boxes and identities of objects in videos. Most methods obtain identities by associating detection boxes whose scores are higher than a threshold. The objects with low detection scores, e.g. *occluded objects, are simply thrown away, which brings non-negligible true object missing and fragmented trajectories. To solve this problem, we present a simple, effective and generic association method, tracking by associating almost every detection box instead of only the high score ones. For the low score detection boxes, we utilize their similarities with tracklets to recover true objects and filter out the background detections. When applied to 9 different state-of-the-art trackers, our method achieves consistent improvement on IDF1 score ranging from 1 to 10 points. To put forwards the state-of-the-art performance of MOT, we design a simple and strong tracker, named ByteTrack. For the first time, we achieve 80.3 MOTA, 77.3 IDF1 and 63.1 HOTA on the test set of MOT17 with 30 FPS running speed on a single V100 GPU. ByteTrack also achieves state-of-the-art performance on MOT20, HiEve and BDD100K tracking benchmarks. The source code, pre-trained models with deploy versions and tutorials of applying to other trackers are released at* https://github.com/ifzhang/ByteTrack.

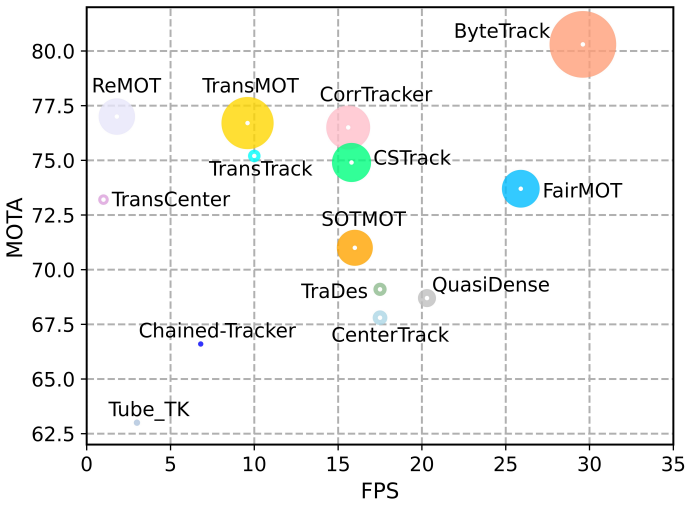


Figure 1. MOTA-IDF1-FPS comparisons of different trackers on the test set of MOT17. The horizontal axis is FPS (running speed), the vertical axis is MOTA, and the radius of circle is IDF1. Our ByteTrack achieves 80.3 MOTA, 77.3 IDF1 on MOT17 test set with 30 FPS running speed, outperforming all previous trackers. Details are given in Table 4.

1. Introduction

Was vernünftig ist, das ist wirklich; und was wirklich ist, das ist vernünftig.

—— *G. W. F. Hegel*

Tracking-by-detection is the most effective paradigm for multi-object tracking (MOT) in current. Due to the complex scenarios in videos, detectors are prone to make imperfect predictions. State-of-the-art MOT methods [1–3, 6, 12, 18, 45, 59, 70, 72, 85] need to deal with true positive / false positive trade-off in detection boxes to eliminate low confidence detection boxes [4, 40]. However, is it the right way to eliminate all low confidence detection boxes? Our answer is NO: as Hegel said "What is reasonable is real; that which is real is reasonable." Low confidence detection boxes sometimes indicate the existence of objects, *e.g.* the occluded objects. Filtering out these objects causes irreversible errors for MOT and brings non-negligible missing detection and fragmented trajectories.

Figure 2 (a) and (b) show this problem. In frame t_1, we initialize three different tracklets as their scores are all higher than 0.5. However, in frame t_2 and frame t_3 when occlusion happens, red tracklet's corresponding detection score becomes lower *i.e.* 0.8 to 0.4 and then 0.4 to

† Corresponding author.

Part of this work was performed while Yifu Zhang worked as an intern at ByteDance.

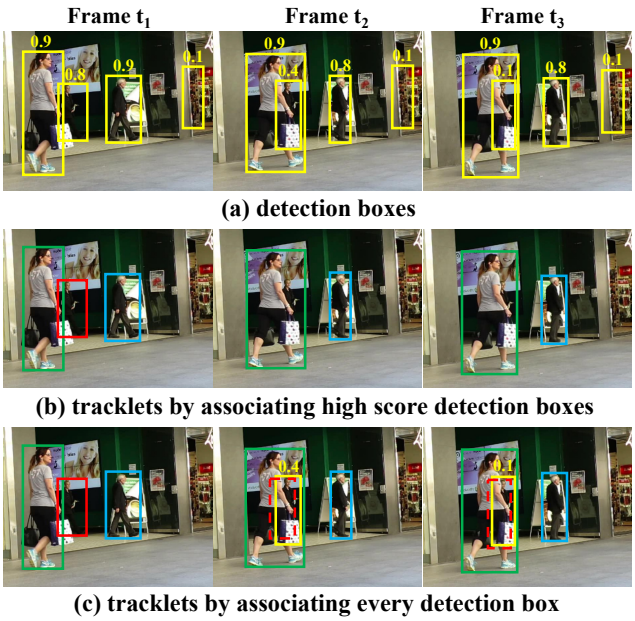


Figure 2. Examples of our method which associates every detection box. (a) shows all the detection boxes with their scores. (b) shows the tracklets obtained by previous methods which associates detection boxes whose scores are higher than a threshold, *i.e*. 0.5. The same box color represents the same identity. (c) shows the tracklets obtained by our method. The dashed boxes represent the predicted box of the previous tracklets using Kalman Filter. The two low score detection boxes are correctly matched to the previous tracklets based on the large IoU.

0.1. These detection boxes are eliminated by the thresholding mechanism and the red tracklet disappears accordingly. Nevertheless, if we take every detection box into consideration, more false positives will be introduced immediately, *e.g*., the most right box in frame t_3 of Figure 2 (a). To the best of our knowledge, very few methods [30, 63] in MOT are able to handle this detection dilemma.

In this paper, we identify that the similarity with tracklets provides a strong cue to distinguish the objects and background in low score detection boxes. As shown in Figure 2 (c), two low score detection boxes are matched to the tracklets by the motion model's predicted boxes, and thus the objects are correctly recovered. At the same time, the background box is removed since it has no matched tracklet.

For making full use of detection boxes from high scores to low ones in the matching process, we present a simple and effective association method BYTE, named for each detection box is a basic unit of the tracklet, as byte in computer program, and our tracking method values every detailed detection box. We first match the high score detection boxes to the tracklets based on motion similarity or appearance similarity. Similar to [6], we adopt Kalman filter [29] to predict the location of the tracklets in the new frame. The similarity can be computed by the IoU or Re-ID feature distance of the predicted box and the detection box. Figure 2 (b) is exactly the results after the first matching. Then, we perform the second matching between the unmatched tracklets, *i.e*. the tracklet in red box, and the low score detection boxes using the same motion similarity. Figure 2 (c) shows the results after the second matching. The occluded person with low detection scores is matched correctly to the previous tracklet and the background (in the right part of the image) is removed.

As the integrating topic of object detection and association, a desirable solution to MOT is never a detector and the following association; besides, well-designed of their junction area is also important. The innovation of BYTE lies in the junction area of detection and association, where low score detection boxes are bridges to boost both of them. Benefiting from this integration innovation, when BYTE is applied to 9 different state-of-the-art trackers, including the Re-ID-based ones [33, 47, 69, 85], motion-based ones [71, 89], chain-based one [48] and attention-based ones [59, 80], notable improvements are achieved on almost all the metrics including MOTA, IDF1 score and ID switches. For example, we increase the MOTA of CenterTrack [89] from 66.1 to 67.4, IDF1 from 64.2 to 74.0 and decrease the IDs from 528 to 144 on the half validation set of MOT17.

Towards pushing forwards the state-of-the-art performance of MOT, we propose a simple and strong tracker, named ByteTrack. We adopt a recent high-performance detector YOLOX [24] to obtain the detection boxes and associate them with our proposed BYTE. On the MOT challenges, ByteTrack ranks 1st on both MOT17 [44] and MOT20 [17], achieving 80.3 MOTA, 77.3 IDF1 and 63.1 HOTA with 30 FPS running speed on V100 GPU on MOT17 and 77.8 MOTA, 75.2 IDF1 and 61.3 HOTA on much more crowded MOT20. ByteTrack also achieves state-of-the-art performance on HiEve [37] and BDD100K [79] tracking benchmarks. We hope the efficiency and simplicity of ByteTrack could make it attractive in real applications such as social computing.

2. Related Work

2.1. Object Detection in MOT

Object detection is one of the most active topics in computer vision and it is the basis of multi-object tracking. The MOT17 dataset [44] provides detection results obtained by popular detectors such as DPM [22], Faster R-CNN [50] and SDP [77]. A large number of methods [3, 9, 12, 14, 28, 74, 91] focus on improving the tracking performance based on these given detection results.

Tracking by detection. With the rapid development of object detection [10, 23, 26, 35, 49, 50, 58, 60], more and more methods begin to utilize more powerful detectors to obtain higher tracking performance. The one-stage object detector RetinaNet [35] begin to be adopted by several methods such as [39, 48]. CenterNet [90] is the most popular detector adopted by most methods [63, 65, 67, 71, 85, 87, 89] for its simplicity and efficiency. The YOLO series detectors [8, 49] are also adopted by a large number of methods [15, 33, 34, 69] for its excellent balance of accuracy and speed. Most of these methods directly use the detection boxes on a single image for tracking.

However, the number of missing detections and very low scoring detections begin to increase when occlusion or motion blur happens in the video sequence, as is pointed out by video object detection methods [41, 62]. Therefore, the information of the previous frames are usually leveraged to enhance the video detection performance.

Detection by tracking. Tracking can also adopted to help obtain more accurate detection boxes. Some methods [12–15, 53, 91] utilize single object tracking (SOT) [5] or Kalman filter [29] to predict the location of the tracklets in the following frame and fuse the predicted boxes with the detection boxes to enhance the detection results. Other methods [34, 86] leverage tracked boxes in the previous frames to enhance feature representation of the following frame. Recently, Transformer-based [20, 38, 64, 66] detectors [11, 92] are adopted by several methods [42, 59, 80] for its strong ability to propagate boxes between frames. Our method also utilize the similarity with tracklets to strength the reliability of detection boxes.

After obtaining the detection boxes by various detectors, most MOT methods [33, 39, 47, 59, 69, 71, 85] only keep the high score detection boxes by a threshold, *i.e.* 0.5, and use those boxes as the input of data association. This is because the low score detection boxes contain many backgrounds which harm the tracking performance. However, we observe that many occluded objects can be correctly detected but have low scores. To reduce missing detections and keep the persistence of trajectories, we keep all the detection boxes and associate across every of them.

2.2. Data Association

Data association is the core of multi-object tracking, which first computes the similarity between tracklets and detection boxes and leverage different strategies to match them according to the similarity.

Similarity metrics. Location, motion and appearance are useful cues for association. SORT [6] combines location and motion cues in a very simple way. It first adopts Kalman filter [29] to predict the location of the tracklets in the new frame and then computes the IoU between the detection boxes and the predicted boxes as the similarity. Some recent methods [59, 71, 89] design networks to learn object motions and achieve more robust results in cases of large camera motion or low frame rate. Location and motion similarity are accurate in the short-range matching. Appearance similarity are helpful in the long-range matching. An object can be re-identified using appearance similarity after being occluded for a long period of time. Appearance similarity can be measured by the cosine similarity of the Re-ID features. DeepSORT [70] adopts a stand-alone Re-ID model to extract appearance features from the detection boxes. Recently, joint detection and Re-ID models [33, 39, 47, 69, 84, 85] becomes more and more popular because of their simplicity and efficiency.

Matching strategy. After similarity computation, matching strategy assigns identities to the objects. This can be done by Hungarian Algorithm [31] or greedy assignment [89]. SORT [6] matches the detection boxes to the tracklets by once matching. DeepSORT [70] proposes a cascaded matching strategy which first matches the detection boxes to the most recent tracklets and then to the lost ones. MOTDT [12] first utilizes appearance similarity to match and then utilize the IoU similarity to match the unmatched tracklets. QDTrack [47] turns the appearance similarity into probability by a bi-directional softmax operation and adopts a nearest neighbor search to accomplish matching. Attention mechanism [64] can directly propagate boxes between frames and perform association implicitly. Recent methods such as [42, 80] propose track queries to find the location of the tracked objects in the following frames. The matching is implicitly performed in the attention interaction process without using Hungarian Algorithm.

All these methods focus on how to design better association methods. However, we argue that the way detection boxes are utilized determines the upper bound of data association and we focus on how to make full use of detection boxes from high scores to low ones in the matching process.

3. BYTE

We propose a simple, effective and generic data association method, BYTE. Different from previous methods [33, 47, 69, 85] which only keep the high score detection boxes, we keep almost every detection box and separate them into high score ones and low score ones. We first associate the high score detection boxes to the tracklets. Some tracklets get unmatched because they do not match to an appropriate high score detection box, which usually happens when occlusion, motion blur or size changing occurs. We then associate the low score detection boxes and these unmatched tracklets to recover the objects in low score detection boxes and filter out background, simultaneously. The

Algorithm 1: Pseudo-code of BYTE.

Input: A video sequence V; object detector Det; detection score threshold τ

Output: Tracks $\mathcal{T}$ of the video

Initialization: $\mathcal{T} \leftarrow \emptyset$
for *frame* f_k *in* V **do**
  /* Figure 2(a) */
  /* predict detection boxes & scores */
  $\mathcal{D}_k \leftarrow \texttt{Det}(f_k)$
  $\mathcal{D}_{high} \leftarrow \emptyset$
  $\mathcal{D}_{low} \leftarrow \emptyset$
  for d *in* $\mathcal{D}_k$ **do**
    if $d.score > \tau$ **then**
      $\mathcal{D}_{high} \leftarrow \mathcal{D}_{high} \cup \{d\}$
    end
    else
      $\mathcal{D}_{low} \leftarrow \mathcal{D}_{low} \cup \{d\}$
    end
  end
  /* predict new locations of tracks */
  for t *in* $\mathcal{T}$ **do**
    $t \leftarrow \texttt{KalmanFilter}(t)$
  end
  /* Figure 2(b) */
  /* first association */
  Associate $\mathcal{T}$ and $\mathcal{D}_{high}$ using Similarity#1
  $\mathcal{D}_{remain} \leftarrow$ remaining object boxes from $\mathcal{D}_{high}$
  $\mathcal{T}_{remain} \leftarrow$ remaining tracks from $\mathcal{T}$
  /* Figure 2(c) */
  /* second association */
  Associate $\mathcal{T}_{remain}$ and $\mathcal{D}_{low}$ using similarity#2
  $\mathcal{T}_{re-remain} \leftarrow$ remaining tracks from $\mathcal{T}_{remain}$
  /* delete unmatched tracks */
  $\mathcal{T} \leftarrow \mathcal{T} \setminus \mathcal{T}_{re-remain}$
  /* initialize new tracks */
  for d *in* $\mathcal{D}_{remain}$ **do**
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{d\}$
  end
end
Return: $\mathcal{T}$

Track rebirth [70, 89] is not shown in the algorithm for simplicity. In green is the key of our method.

pseudo-code of BYTE is shown in Algorithm 1.

The input of BYTE is a video sequence V, along with an object detector Det. We also set a detection score threshold τ. The output of BYTE is the tracks $\mathcal{T}$ of the video and each track contains the bounding box and identity of the object in each frame.

For each frame in the video, we predict the detection boxes and scores using the detector Det. We separate all the detection boxes into two parts $\mathcal{D}_{high}$ and $\mathcal{D}_{low}$ according to the detection score threshold τ. For the detection boxes whose scores are higher than τ, we put them into the high score detection boxes $\mathcal{D}_{high}$. For the detection boxes whose scores are lower than τ, we put them into the low score detection boxes $\mathcal{D}_{low}$ (line 3 to 13 in Algorithm 1).

After separating the low score detection boxes and the high score detection boxes, we adopt Kalman filter to predict the new locations in the current frame of each track in $\mathcal{T}$ (line 14 to 16 in Algorithm 1).

The first association is performed between the high score detection boxes $\mathcal{D}_{high}$ and all the tracks $\mathcal{T}$ (including the lost tracks $\mathcal{T}_{lost}$). Similarity#1 can be computed by either by the IoU or the Re-ID feature distances between the detection boxes $\mathcal{D}_{high}$ and the predicted box of tracks $\mathcal{T}$. Then, we adopt Hungarian Algorithm [31] to finish the matching based on the similarity. We keep the unmatched detections in $\mathcal{D}_{remain}$ and the unmatched tracks in $\mathcal{T}_{remain}$ (line 17 to 19 in Algorithm 1).

BYTE is highly flexible and can be compatible to other different association methods. For example, when BYTE is combined with FairMOT [85], Re-ID feature is added into * first association * in Algorithm 1, others are the same. In the experiments, we apply BYTE to 9 different state-of-the-art trackers and achieve notable improvements on almost all the metrics.

The second association is performed between the low score detection boxes $\mathcal{D}_{low}$ and the remaining tracks $\mathcal{T}_{remain}$ after the first association. We keep the unmatched tracks in $\mathcal{T}_{re-remain}$ and just delete all the unmatched low score detection boxes, since we view them as background. (line 20 to 21 in Algorithm 1). We find it important to use IoU alone as the Similarity#2 in the second association because the low score detection boxes usually contains severe occlusion or motion blur and appearance features are not reliable. Thus, when apply BYTE to other Re-ID based trackers [47, 69, 85], we do not adopt appearance similarity in the second association.

After the association, the unmatched tracks will be deleted from the tracklets. We do not list the procedure of track rebirth [12, 70, 89] in Algorithm 1 for simplicity. Actually, it is necessary for the long-range association to preserve the identity of the tracks. For the unmatched tracks $\mathcal{T}_{re-remain}$ after the second association, we put them into $\mathcal{T}_{lost}$. For each track in $\mathcal{T}_{lost}$, only when it exists for more than a certain number of frames, *i.e.* 30, we delete it from the tracks $\mathcal{T}$. Otherwise, we remain the lost tracks $\mathcal{T}_{lost}$ in $\mathcal{T}$(line 22 in Algorithm 1).Finally, we initialize new tracks from the unmatched high score detection boxes $\mathcal{D}_{remain}$ after the first association. (line 23 to 27 in Algorithm 1).The output of each individual frame is the bounding boxes and identities of the tracks $\mathcal{T}$ in the current frame. Note that we do not output the boxes and identities of $\mathcal{T}_{lost}$.

To put forwards the state-of-the-art performance of MOT, we design a simple and strong tracker, named ByteTrack, by equipping the high-performance detector YOLOX [24] with our association method BYTE.

4. Experiments

4.1. Setting

Datasets. We evaluate BYTE and ByteTrack on MOT17 [44] and MOT20 [17] datasets under the "private detection" protocol. Both datasets contain training sets and test sets, without validation sets. For ablation studies, we use the first half of each video in the training set of MOT17 for training and the last half for validation following [89]. We train on the combination of CrowdHuman dataset [55] and MOT17 half training set following [59, 71, 80, 89]. We add Cityperson [82] and ETHZ [21] for training following [33, 69, 85] when testing on the test set of MOT17. We also test ByteTrack on HiEve [37] and BDD100K [79] datasets. HiEve is a large scale human-centric dataset focusing on crowded and complex events. BDD100K is the largest driving video dataset and the dataset splits of the MOT task are 1400 videos for training, 200 videos for validation and 400 videos for testing. It needs to track objects of 8 classes and contains cases of large camera motion.

Metrics. We use the CLEAR metrics [4], including MOTA, FP, FN, IDs, *etc.*, IDF1 [51] and HOTA [40] to evaluate different aspects of the tracking performance. MOTA is computed based on FP, FN and IDs. Considering the amount of FP and FN are larger than IDs, MOTA focuses more on the detection performance. IDF1 evaluates the identity preservation ability and focus more on the association performance. HOTA is a very recently proposed metric which explicitly balances the effect of performing accurate detection, association and localization. For BDD100K dataset, there are some multi-class metrics such as mMOTA and mIDF1. mMOTA / mIDF1 is computed by averaging the MOTA / IDF1 of all the classes.

Implementation details. For BYTE, the default detection score threshold τ is 0.6, unless otherwise specified. For the benchmark evaluation of MOT17, MOT20 and HiEve, we only use IoU as the similarity metrics. In the linear assignment step, if the IoU between the detection box and the tracklet box is smaller than 0.2, the matching will be rejected. For the lost tracklets, we keep it for 30 frames in case it appears again. For BDD100K, we use UniTrack [68] as the Re-ID model. In ablation study, we use FastReID [27] to extract Re-ID features for MOT17.

For ByteTrack, the detector is YOLOX [24] with YOLOX-X as the backbone and COCO-pretrained model [36] as the initialized weights. For MOT17, the training schedule is 80 epochs on the combination of MOT17, CrowdHuman, Cityperson and ETHZ. For MOT20 and HiEve, we only add CrowdHuman as additional training data. For BDD100K, we do not use additional training data and only train 50 epochs. The input image size is 1440 $\times$800 and the shortest side ranges from 576 to 1024 during multi-scale training. The data augmentation includes Mosaic [8] and Mixup [81]. The model is trained on 8 NVIDIA Tesla V100 GPU with batch size of 48. The optimizer is SGD with weight decay of 5×10^{-4} and momentum of 0.9. The initial learning rate is 10^{-3} with 1 epoch warm-up and cosine annealing schedule. The total training time is about 12 hours. Following [24], FPS is measured with FP16-precision [43] and batch size of 1 on a single GPU.

4.2. Ablation Studies on BYTE

Similarity analysis. We choose different types of similarity for the first association and the second association of BYTE. The results are shown in Table 1. We can see that either IoU or Re-ID can be a good choice for `Similarity#1` on MOT17. IoU achieves better MOTA and IDs while Re-ID achieves higher IDF1. On BDD100K, Re-ID achieves much better results than IoU in the first association. This is because BDD100K contains large camera motion and the annotations are in low frame rate, which causes failure of motion cues. It is important to utilize IoU as `Similarity#2` in the second association on both datasets because the low score detection boxes usually contains severe occlusion or motion blur and thus Re-ID features are not reliable. From Table 1 we can find that using IoU as `Similarity#2` increases about 1.0 MOTA compared to Re-ID, which indicates that Re-ID features of the low score detection boxes are not reliable.

Comparisons with other association methods. We compare BYTE with other popular association methods including SORT [6], DeepSORT [70] and MOTDT [12] on the validation set of MOT17 and BDD100K. The results are shown in Table 2.

SORT can be seen as our baseline method because both methods only adopt Kalman filter to predict the object motion. We can find that BYTE improves the MOTA metric of SORT from 74.6 to 76.6, IDF1 from 76.9 to 79.3 and decreases IDs from 291 to 159. This highlights the importance of the low score detection boxes and proves the ability of BYTE to recover object boxes from low score one.

DeepSORT utilizes additional Re-ID models to enhance the long-range association. We surprisingly find BYTE also has additional gains compared with DeepSORT. This suggests a simple Kalman filter can perform long-range association and achieve better IDF1 and IDs when the detection boxes are accurate enough. We note that in severe occlusion cases, Re-ID features are vulnerable and may lead to identity switches, instead, motion model behaves more reliably.

MOTDT integrates motion-guided box propagation results along with detection results to associate unreliable detection results with tracklets. Although sharing the simi-

Similarity#1	Similarity#2	MOT17 MOTA↑	MOT17 IDF1↑	MOT17 IDs↓	BDD100K mMOTA↑	BDD100K mIDF1↑	BDD100K IDs↓
IoU	Re-ID	75.8	77.5	231	39.2	48.3	29172
IoU	IoU	**76.6**	79.3	**159**	39.4	48.9	27902
Re-ID	Re-ID	75.2	78.7	276	45.0	53.4	10425
Re-ID	IoU	76.3	**80.5**	216	**45.5**	**54.8**	**9140**

Table 1. Comparison of different type of similarity metrics used in the first association and the second association of BYTE on MOT17 and BDD100K validation set. The best results are shown in **bold**.

Method	w/ Re-ID	MOT17 MOTA↑	MOT17 IDF1↑	MOT17 IDs↓	BDD100K mMOTA↑	BDD100K mIDF1↑	BDD100K IDs↓	FPS
SORT		74.6	76.9	291	30.9	41.3	10067	**30.1**
DeepSORT	✓	75.4	77.2	239	24.5	38.2	10720	13.5
MOTDT	✓	75.8	77.6	273	26.7	39.8	14520	11.1
BYTE (ours)		**76.6**	79.3	**159**	39.4	48.9	27902	29.6
BYTE (ours)	✓	76.3	**80.5**	216	**45.5**	**54.8**	**9140**	11.8

Table 2. Comparison of different data association methods on MOT17 and BDD100K validation set. The best results are shown in **bold**.

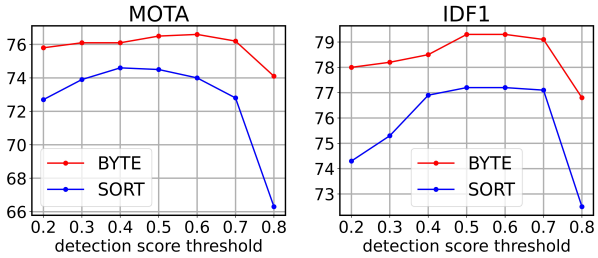


Figure 3. Comparison of the performances of BYTE and SORT under different detection score thresholds. The results are from the validation set of MOT17.

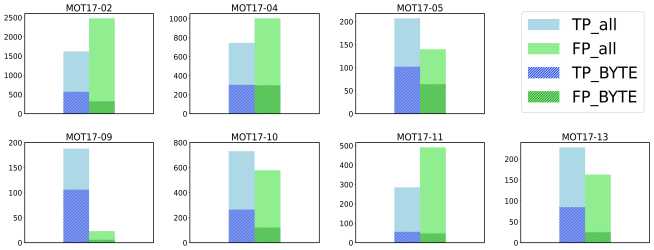


Figure 4. Comparison of the number of TPs and FPs in all low score detection boxes and the low score tracked boxes obtained by BYTE. The results are from the validation set of MOT17.

lar motivation, MOTDT is behind BYTE by a large margin. We explain that MOTDT uses propagated boxes as tracklet boxes, which may lead to locating drifts in tracking. Instead, BYTE uses low-score detection boxes to re-associate those unmatched tracklets, therefore, tracklet boxes are more accurate.

Table 2 also shows the results on BDD100K dataset. BYTE also outperforms other association methods by a large margin. Kalman filter fails in autonomous driving scenes and it is the main reason for the low performance of SORT, DeepSORT and MOTDT. Thus, we do not use Kalman filter on BDD100K. Additional off-the-shelf Re-ID models greatly improve the performance of BYTE on BDD100K.

Robustness to detection score threshold. The detection score threshold τ_{high} is a sensitive hyper-parameter and needs to be carefully tuned in the task of multi-object tracking. We change it from 0.2 to 0.8 and compare the MOTA and IDF1 score of BYTE and SORT. The results are shown in Figure 3. From the results we can see that BYTE is more robust to the detection score threshold than SORT. This is because the second association in BYTE recovers the objects whose scores are lower than τ_{high}, and thus considers almost every detection box regardless of the change of τ_{high}.

Analysis on low score detection boxes. To prove the effectiveness of BYTE, we collect the number of TPs and FPs in the low score boxes obtained by BYTE. We use the half training set of MOT17 and CrowdHuman for training and evaluate on the half validation set of MOT17. First, we keep all the low score detection boxes whose scores range from τ_{low} to τ_{high} and classify the TPs and FPs using ground truth annotations. Then, we select the tracking results obtained by BYTE from low score detection boxes. The results of each sequence are shown in Figure 4. We can see that BYTE obtains notably more TPs than FPs from

the low score detection boxes even though some sequences (*i.e*. MOT17-02) have much more FPs in all the detection boxes. The obtained TPs notably increases MOTA from 74.6 to 76.6 as is shown in Table 2.

Applications on other trackers. We apply BYTE on 9 different state-of-the-arts trackers, including JDE [69], CSTrack [33], FairMOT [85], TraDes [71], QDTrack [47], CenterTrack [89], Chained-Tracker [48], TransTrack [59] and MOTR [80]. Among these trackers, JDE, CSTrack, FairMOT, TraDes adopt a combination of motion and Re-ID similarity. QDTrack adopts Re-ID similarity alone. CenterTrack and TraDes predict the motion similarity by the learned networks. Chained-Tracker adopts the chain structure and outputs the results of two consecutive frames simultaneously and associate in the same frame by IoU. TransTrack and MOTR adopt the attention mechanism to propagate boxes among frames. Their results are shown in the first line of each tracker in Table 3.To evaluate the effectiveness of BYTE, we design two different modes to apply BYTE to these trackers.

- The first mode is to insert BYTE into the original association methods of different trackers, as is shown in the second line of the results of each tracker in Table 3. Take FairMOT [85] for example, after the original association is done, we select all the unmatched tracklets and associate them with the low score detection boxes following the `* second association *` in Algorithm 1. Note that for the low score objects, the Re-ID features are not reliable so we only adopt the IoU between the detection boxes and the tracklet boxes after motion prediction as the similarity. We do not apply the first mode of BYTE to Chained-Tracker because we find it is difficult to implement in the chain structure.

- The second mode is to directly use the detection boxes of these trackers and associate using the whole procedure in Algorithm 1, as is shown in the third line of the results of each tracker in Table 3.

We can see that in both modes, BYTE can bring stable improvements over almost all the metrics including MOTA, IDF1 and IDs. For example, BYTE increases CenterTrack by 1.3 MOTA and 9.8 IDF1, Chained-Tracker by 1.9 MOTA and 5.8 IDF1, TransTrack by 1.2 MOTA and 4.1 IDF1. The results in Table 3 indicate that BYTE has strong generalization ability and can be easily applied to existing trackers to obtain performance gain.

4.3. Benchmark Evaluation

We compare ByteTrack with the state-of-the-art trackers on the test set of MOT17, MOT20 and HiEve under the private detection protocol in Table 4, Table 5 and Table 6, respectively. All the results are directly obtained from the official MOT Challenge evaluation server and the Human in Events server.

Method	Similarity	w/ BYTE	MOTA↑	IDF1↑	IDs↓
JDE [69]	Motion(K) + Re-ID		60.0	63.6	473
	Motion(K) + Re-ID	✓	60.3 (+0.3)	64.1 (+0.5)	418
	Motion(K)	✓	60.6 (+0.6)	66.0 (**+2.4**)	360
CSTrack [33]	Motion(K) + Re-ID		68.0	72.3	325
	Motion(K) + Re-ID	✓	69.2 (**+1.2**)	73.9 (**+1.6**)	285
	Motion(K)	✓	69.3 (**+1.3**)	71.7 (-0.6)	279
FairMOT [85]	Motion(K) + Re-ID		69.1	72.8	299
	Motion(K) + Re-ID	✓	70.4 (**+1.3**)	74.2 (**+1.4**)	232
	Motion(K)	✓	70.3 (**+1.2**)	73.2 (+0.4)	236
TraDes [71]	Motion + Re-ID		68.2	71.7	285
	Motion + Re-ID	✓	68.6 (+0.4)	71.1 (-0.6)	259
	Motion(K)	✓	67.9 (-0.3)	72.0 (+0.3)	178
QuasiDense [47]	Re-ID		67.3	67.8	377
	Motion(K) + Re-ID	✓	67.7 (+0.4)	72.0 (**+4.2**)	281
	Motion(K)	✓	67.9 (+0.6)	70.9 (**+3.1**)	258
CenterTrack [89]	Motion		66.1	64.2	528
	Motion	✓	66.3 (+0.2)	64.8 (+0.6)	334
	Motion(K)	✓	67.4 (**+1.3**)	74.0 (**+9.8**)	144
CTracker [48]	Chain		63.1	60.9	755
	Motion(K)	✓	65.0 (**+1.9**)	66.7 (**+5.8**)	346
TransTrack [59]	Attention		67.1	68.3	254
	Attention	✓	68.6 (**+1.5**)	69.0 (+0.7)	232
	Motion(K)	✓	68.3 (**+1.2**)	72.4 (**+4.1**)	181
MOTR [80]	Attention		64.7	67.2	346
	Attention	✓	64.3 (-0.4)	69.3 (**+2.1**)	263
	Motion(K)	✓	65.7 (**+1.0**)	68.4 (**+1.2**)	260

Table 3. Results of applying BYTE to 9 different state-of-the-art trackers on the MOT17 validation set. "K" is short for Kalman Filter. In green are the improvements of at least **+1.0** point.

MOT17. ByteTrack ranks 1st among all the trackers on the leaderboard of MOT17. Not only does it achieve the best accuracy (*i.e*. 80.3 MOTA, 77.3 IDF1 and 63.1 HOTA), but also runs with highest running speed (30 FPS). It outperforms the second-performance tracker [76] by a large margin (*i.e*. +3.3 MOTA, +5.3 IDF1 and +3.4 HOTA). Also, we use less training data than many high performance methods such as [33, 34, 54, 65, 85] (29K images vs. 73K images). It is worth noting that we only leverage the simplest similarity computation method Kalman filter in the association step compared to other methods [33, 47, 59, 67, 80, 85] which additionally adopt Re-ID similarity or attention mechanisms. All these indicate that ByteTrack is a simple and strong tracker.

MOT20. Compared with MOT17, MOT20 has much more crowded scenarios and occlusion cases. The average number of pedestrians in an image is 170 in the test set of MOT20. ByteTrack also ranks 1st among all the trackers on

https://motchallenge.net
http://humaninevents.org

Tracker	MOTA↑	IDF1↑	HOTA↑	FP↓	FN↓	IDs↓	FPS↑
DAN [61]	52.4	49.5	39.3	25423	234592	8431	<3.9
Tube_TK [46]	63.0	58.6	48.0	27060	177483	4137	3.0
MOTR [80]	65.1	66.4	-	45486	149307	2049	-
CTracker [48]	66.6	57.4	49.0	22284	160491	5529	6.8
CenterTrack [89]	67.8	64.7	52.2	**18498**	160332	3039	17.5
QuasiDense [47]	68.7	66.3	53.9	26589	146643	3378	20.3
TraDes [71]	69.1	63.9	52.7	20892	150060	3555	17.5
MAT [25]	69.5	63.1	53.8	30660	138741	2844	9.0
SOTMOT [87]	71.0	71.9	-	39537	118983	5184	16.0
TransCenter [75]	73.2	62.2	54.5	23112	123738	4614	1.0
GSDT [67]	73.2	66.5	55.2	26397	120666	3891	4.9
Semi-TCL [32]	73.3	73.2	59.8	22944	124980	2790	-
FairMOT [85]	73.7	72.3	59.3	27507	117477	3303	25.9
RelationTrack [78]	73.8	74.7	61.0	27999	118623	**1374**	8.5
PermaTrackPr [63]	73.8	68.9	55.5	28998	115104	3699	11.9
CSTrack [33]	74.9	72.6	59.3	23847	114303	3567	15.8
TransTrack [59]	75.2	63.5	54.1	50157	86442	3603	10.0
FUFET [54]	76.2	68.0	57.9	32796	98475	3237	6.8
SiamMOT [34]	76.3	72.3	-	-	-	-	12.8
CorrTracker [65]	76.5	73.6	60.7	29808	99510	3369	15.6
TransMOT [15]	76.7	75.1	61.7	36231	93150	2346	9.6
ReMOT [76]	77.0	72.0	59.7	33204	93612	2853	1.8
ByteTrack (ours)	**80.3**	**77.3**	**63.1**	25491	**83721**	2196	**29.6**

Table 4. Comparison of the state-of-the-art methods under the "private detector" protocol on MOT17 test set. The best results are shown in **bold**. MOT17 contains rich scenes and half of the sequences are captured with camera motion. **ByteTrack** ranks 1st among all the trackers on the leaderboard of MOT17 and outperforms the second one ReMOT by a large margin on almost all the metrics. It also has the highest running speed among all trackers.

Tracker	MOTA↑	IDF1↑	HOTA↑	FP↓	FN↓	IDs↓	FPS↑
MLT [83]	48.9	54.6	43.2	45660	216803	2187	3.7
FairMOT [85]	61.8	67.3	54.6	103440	88901	5243	13.2
TransCenter [75]	61.9	50.4	-	45895	146347	4653	1.0
TransTrack [59]	65.0	59.4	48.5	27197	150197	3608	7.2
CorrTracker [65]	65.2	69.1	-	79429	95855	5183	8.5
Semi-TCL [32]	65.2	70.1	55.3	61209	114709	4139	-
CSTrack [33]	66.6	68.6	54.0	**25404**	144358	3196	4.5
GSDT [67]	67.1	67.5	53.6	31913	135409	3131	0.9
SiamMOT [34]	67.1	69.1	-	-	-	-	4.3
RelationTrack [78]	67.2	70.5	56.5	61134	104597	4243	2.7
SOTMOT [87]	68.6	71.4	-	57064	101154	4209	8.5
ByteTrack (ours)	**77.8**	**75.2**	**61.3**	26249	**87594**	**1223**	**17.5**

Table 5. Comparison of the state-of-the-art methods under the "private detector" protocol on MOT20 test set. The best results are shown in **bold**. The scenes in MOT20 are much more crowded than those in MOT17. **ByteTrack** ranks 1st among all the trackers on the leaderboard of MOT20 and outperforms the second one SOTMOT by a large margin on all the metrics. It also has the highest running speed among all trackers.

the leaderboard of MOT20 and outperforms other trackers by a large margin on almost all the metrics. For example, it increases MOTA from 68.6 to 77.8, IDF1 from 71.4 to 75.2 and decreases IDs by 71% from 4209 to 1223. It is worth noting that ByteTrack achieves extremely low identity switches, which further indicates that associating every detection boxes is very effective under occlusion cases.

Tracker	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓
DeepSORT [70]	27.1	28.6	8.5%	41.5%	5894	42668	2220
MOTDT [12]	26.1	32.9	8.7%	54.6%	6318	43577	1599
IOUtracker [7]	38.6	38.6	28.3%	27.6%	9640	28993	4153
JDE [69]	33.1	36.0	15.1%	24.1%	9526	33327	3747
FairMOT [85]	35.0	46.7	16.3%	44.2%	6523	37750	**995**
CenterTrack [89]	40.9	45.1	10.8%	32.2%	3208	36414	1568
ByteTrack (Ours)	**61.7**	**63.1**	**38.3%**	**21.6%**	**2822**	**22852**	1031

Table 6. Comparison of the state-of-the-art methods under the "private detector" protocol on HiEve test set. The best results are shown in **bold**. HiEve has more complex events than MOT17 and MOT20. **ByteTrack** ranks 1st among all the trackers on the leaderboard of HiEve and outperforms the second one CenterTrack by a large margin on all the metrics.

Human in Events. Compared with MOT17 and MOT20, HiEve contains more complex events and more diverse camera views. We train ByteTrack on CrowdHuman dataset and the training set of HiEve. ByteTrack also ranks 1st among all the trackers on the leaderboard of HiEve and outperforms other state-of-the-art trackers by a large margin. For example, it increases MOTA from 40.9 to 61.3 and IDF1 from 45.1 to 62.9. The superior results indicate that ByteTrack is robust to complex scenes.

BDD100K. BDD100K is multiple categories tracking dataset in autonomous driving scenes. The challenges include low frame rate and large camera motion. We utilize a simple ResNet-50 ImageNet classification model from UniTrack [68] to extract Re-ID features and compute appearance similarity. ByteTrack ranks first on the leaderboard of BDD100K. It increases mMOTA from 36.6 to 45.5 on the validation set and 35.5 to 40.1 on the test set, which indicates that ByteTrack can also handle the challenges in autonomous driving scenes.

5. Conclusion

We present a simple yet effective data association method BYTE for multi-object tracking. BYTE can be easily applied to existing trackers and achieve consistent improvements. We also propose a strong tracker ByteTrack, which achieves 80.3 MOTA, 77.3 IDF1 and 63.1 HOTA on MOT17 test set with 30 FPS, ranking 1st among all the trackers on the leaderboard. ByteTrack is very robust to occlusion for its accurate detection performance and the help of associating low score detection boxes. It also sheds light on how to make the best use of detection results to enhance multi-object tracking. We hope the high accuracy, fast speed and simplicity of ByteTrack can make it attractive in real applications.

Tracker	split	mMOTA↑	mIDF1↑	MOTA↑	IDF1↑	FN↓	FP↓	IDs↓	MT↑	ML↓
Yu *et al*. [79]	val	25.9	44.5	56.9	66.8	122406	52372	8315	8396	3795
QDTrack [47]	val	36.6	50.8	63.5	**71.5**	108614	46621	**6262**	9481	3034
ByteTrack(Ours)	val	**45.5**	**54.8**	**69.1**	70.4	**92805**	**34998**	9140	**9626**	**3005**
Yu *et al*. [79]	test	26.3	44.7	58.3	68.2	213220	100230	14674	16299	6017
DeepBlueAI	test	31.6	38.7	56.9	56.0	292063	**35401**	25186	10296	12266
madamada	test	33.6	43.0	59.8	55.7	209339	76612	42901	16774	**5004**
QDTrack [47]	test	35.5	52.3	64.3	**72.3**	201041	80054	**10790**	17353	5167
ByteTrack(Ours)	test	**40.1**	**55.8**	**69.6**	71.3	**169073**	63869	15466	**18057**	5107

Table 7. Comparison of the state-of-the-art methods on BDD100K test set. The best results are shown in **bold**. **ByteTrack** ranks 1st among all the trackers on the leaderboard of BDD100K and outperforms the second one QDTrack by a large margin on most metrics.

A. Bounding box annotations

We note MOT17 [44] requires the bounding boxes [89] 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 [17], HiEve [37] 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 [70] using light detection models. We use YOLOX [24] 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 8. 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 9. 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 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.

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 8. 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 9. 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**.

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 10. 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**.

Training data. We evaluate ByteTrack on the half validation set of MOT17 using different combinations of training data. The results are shown in Table 10. 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 [8] and Mixup [81]. When further adding CrowdHu-

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

man, 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 [33, 34, 65, 85] that require more than 7 data sources [19, 21, 44, 55, 73, 82, 88] 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.

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 11, 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 [44], MOT20 [17] and HiEve [37] under the private detection protocol.

E. Public detection results on MOTChallenge

We evaluate ByteTrack on the test set of MOT17 [44] and MOT20 [17] under the public detection protocol. Following the public detection filtering strategy in Tracktor [3] and CenterTrack [89], 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 12, ByteTrack outperforms other methods by a large margin on MOT17. For example, it outperforms SiamMOT by 1.5 points on MOTA and 6.7 points on IDF1. Table 13 shows the results on MOT20. ByteTrack also outperforms existing results by a large margin. For example, it outperforms TMOH [57] 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.

Tracker	MOTA↑	IDF1↑	HOTA↑	FP↓	FN↓	IDs↓
STRN [74]	50.9	56.0	42.6	25295	249365	2397
FAMNet [14]	52.0	48.7	-	14138	253616	3072
Tracktor++v2 [3]	56.3	55.1	44.8	**8866**	235449	1987
MPNTrack [9]	58.8	61.7	49.0	17413	213594	1185
LPC_MOT [16]	59.0	66.8	51.5	23102	206948	**1122**
Lif_T [28]	60.5	65.6	51.1	14966	206619	1189
CenterTrack [89]	61.5	59.6	48.2	14076	200672	2583
TMOH [57]	62.1	62.8	50.4	10951	201195	1897
ArTIST_C [52]	62.3	59.7	48.9	19611	191207	2062
QDTrack [47]	64.6	65.1	-	14103	182998	2652
SiamMOT [56]	65.9	63.3	-	18098	170955	3040
ByteTrack (ours)	**67.4**	**70.0**	**56.1**	9939	**172636**	1331

Table 12. 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 [6]	42.7	45.1	36.1	27521	264694	4470
Tracktor++v2 [3]	52.6	52.7	42.1	**6930**	236680	1648
ArTIST_C [52]	53.6	51.0	41.6	7765	230576	1531
LPC_MOT [16]	56.3	62.5	49.0	11726	213056	1562
MPNTrack [9]	57.6	59.1	46.8	16953	201384	1210
TMOH [57]	60.1	61.2	48.9	38043	165899	2342
ByteTrack (ours)	**67.0**	**70.2**	**56.4**	9685	**160303**	**680**

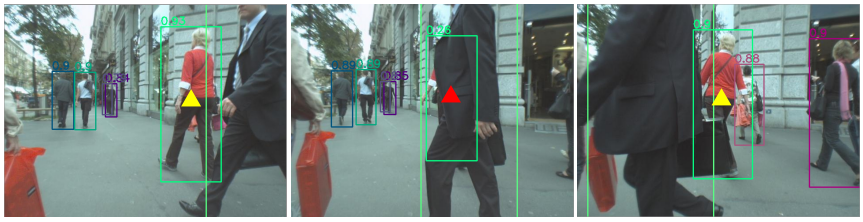
Table 13. Comparison of the state-of-the-art methods under the "public detector" protocol on MOT20 test set. The best results are shown in **bold**.

F. Visualization results.

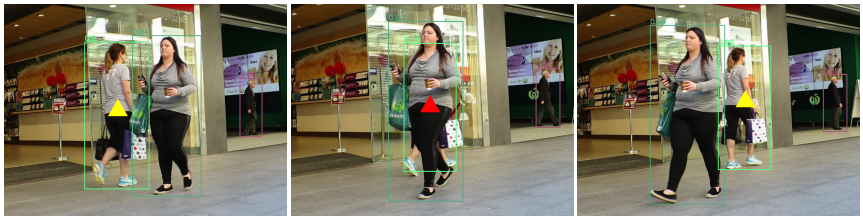
We show some visualization results of difficult cases which ByteTrack is able to handle in Figure 5. The difficult cases include occlusion (*i.e.* MOT17-02, MOT17-04, MOT17-05, MOT17-09, MOT17-13), motion blur (*i.e.* MOT17-10, MOT17-13) and small objects (*i.e.* MOT17-13). The pedestrian in the middle 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 association. As we can see from all these difficult cases, ByteTrack does not bring any identity switch and preserve the identity effectively.

MOT17-02

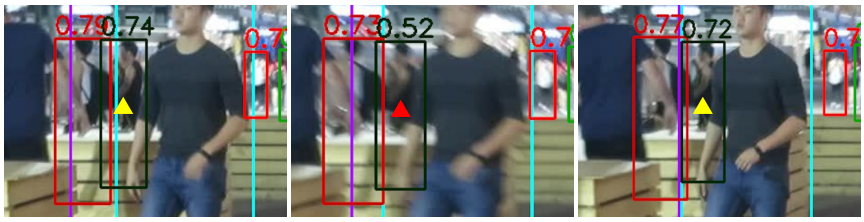
MOT17-04



MOT17-05



MOT17-09



MOT17-10

MOT17-13

Figure 5. 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] S.-H. Bae and K.-J. Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1218–1225, 2014.

[2] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *IEEE transactions on pattern analysis and machine intelligence*, 33(9):1806–1819, 2011.

[3] P. Bergmann, T. Meinhardt, and L. Leal-Taixe. Tracking without bells and whistles. In *ICCV*, pages 941–951, 2019.

[4] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

[5] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016.

[6] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. Simple online and realtime tracking. In *ICIP*, pages 3464–3468. IEEE, 2016.

[7] E. Bochinski, V. Eiselein, and T. Sikora. High-speed tracking-by-detection without using image information. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017.

[8] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[9] G. Brasó and L. Leal-Taixé. Learning a neural solver for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6247–6257, 2020.

[10] Z. Cai and N. Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, pages 6154–6162, 2018.

[11] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[12] L. Chen, H. Ai, Z. Zhuang, and C. Shang. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018.

[13] P. Chu, H. Fan, C. C. Tan, and H. Ling. Online multi-object tracking with instance-aware tracker and dynamic model refreshment. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 161–170. IEEE, 2019.

[14] P. Chu and H. Ling. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In *ICCV*, pages 6172–6181, 2019.

[15] P. Chu, J. Wang, Q. You, H. Ling, and Z. Liu. Transmot: Spatial-temporal graph transformer for multiple object tracking. *arXiv preprint arXiv:2104.00194*, 2021.

[16] P. Dai, R. Weng, W. Choi, C. Zhang, Z. He, and W. Ding. Learning a proposal classifier for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2443–2452, 2021.

[17] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020.

[18] C. Dicle, O. I. Camps, and M. Sznaier. The way they move: Tracking multiple targets with similar appearance. In *Proceedings of the IEEE international conference on computer vision*, pages 2304–2311, 2013.

[19] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, pages 304–311. IEEE, 2009.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[21] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, pages 1–8. IEEE, 2008.

[22] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, pages 1–8. IEEE, 2008.

[23] J. Fu, L. Zong, Y. Li, K. Li, B. Yang, and X. Liu. Model adaption object detection system for robot. In *2020 39th Chinese Control Conference (CCC)*, pages 3659–3664. IEEE, 2020.

[24] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[25] S. Han, P. Huang, H. Wang, E. Yu, D. Liu, X. Pan, and J. Zhao. Mat: Motion-aware multi-object tracking. *arXiv preprint arXiv:2009.04794*, 2020.

[26] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[27] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020.

[28] A. Hornakova, R. Henschel, B. Rosenhahn, and P. Swoboda. Lifted disjoint paths with application in multiple object tracking. In *International Conference on Machine Learning*, pages 4364–4375. PMLR, 2020.

[29] R. E. Kalman. A new approach to linear filtering and prediction problems. *J. Fluids Eng.*, 82(1):35–45, 1960.

[30] T. Khurana, A. Dave, and D. Ramanan. Detecting invisible people. *arXiv preprint arXiv:2012.08419*, 2020.

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

[32] W. Li, Y. Xiong, S. Yang, M. Xu, Y. Wang, and W. Xia. Semi-tcl: Semi-supervised track contrastive representation learning. *arXiv preprint arXiv:2107.02396*, 2021.

[33] C. Liang, Z. Zhang, Y. Lu, X. Zhou, B. Li, X. Ye, and J. Zou. Rethinking the competition between detection and reid in multi-object tracking. *arXiv preprint arXiv:2010.12138*, 2020.

[34] C. Liang, Z. Zhang, X. Zhou, B. Li, Y. Lu, and W. Hu. One more check: Making" fake background" be tracked again. *arXiv preprint arXiv:2104.09441*, 2021.

[35] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[37] W. Lin, H. Liu, S. Liu, Y. Li, R. Qian, T. Wang, N. Xu, H. Xiong, G.-J. Qi, and N. Sebe. Human in events: A large-scale benchmark for human-centric video analysis in complex events. *arXiv preprint arXiv:2005.04490*, 2020.

[38] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[39] Z. Lu, V. Rathod, R. Votel, and J. Huang. Retinatrack: Online single stage joint detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14668–14678, 2020.

[40] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2):548–578, 2021.

[41] H. Luo, W. Xie, X. Wang, and W. Zeng. Detect or track: Towards cost-effective video object detection/tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8803–8810, 2019.

[42] T. Meinhardt, A. Kirillov, L. Leal-Taixe, and C. Feichtenhofer. Trackformer: Multi-object tracking with transformers. *arXiv preprint arXiv:2101.02702*, 2021.

[43] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.

[44] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

[45] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multitarget tracking. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):58–72, 2013.

[46] B. Pang, Y. Li, Y. Zhang, M. Li, and C. Lu. Tubetk: Adopting tubes to track multi-object in a one-step training model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6308–6318, 2020.

[47] J. Pang, L. Qiu, X. Li, H. Chen, Q. Li, T. Darrell, and F. Yu. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 164–173, 2021.

[48] J. Peng, C. Wang, F. Wan, Y. Wu, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In *European Conference on Computer Vision*, pages 145–161. Springer, 2020.

[49] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[50] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[51] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, pages 17–35. Springer, 2016.

[52] F. Saleh, S. Aliakbarian, H. Rezatofighi, M. Salzmann, and S. Gould. Probabilistic tracklet scoring and inpainting for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14329–14339, 2021.

[53] R. Sanchez-Matilla, F. Poiesi, and A. Cavallaro. Online multi-target tracking with strong and weak detections. In *ECCV*, pages 84–99. Springer, 2016.

[54] C. Shan, C. Wei, B. Deng, J. Huang, X.-S. Hua, X. Cheng, and K. Liang. Tracklets predicting based adaptive graph tracking. *arXiv preprint arXiv:2010.09015*, 2020.

[55] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.

[56] B. Shuai, A. Berneshawi, X. Li, D. Modolo, and J. Tighe. Siammot: Siamese multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12372–12382, 2021.

[57] D. Stadler and J. Beyerer. Improving multiple pedestrian tracking by track management and occlusion handling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10958–10967, 2021.

[58] P. Sun, Y. Jiang, E. Xie, W. Shao, Z. Yuan, C. Wang, and P. Luo. What makes for end-to-end object detection? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9934–9944. PMLR, 2021.

[59] P. Sun, Y. Jiang, R. Zhang, E. Xie, J. Cao, X. Hu, T. Kong, Z. Yuan, C. Wang, and P. Luo. Transtrack: Multiple-object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

[60] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14454–14463, 2021.

[61] S. Sun, N. Akhtar, H. Song, A. S. Mian, and M. Shah. Deep affinity network for multiple object tracking. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[62] P. Tang, C. Wang, X. Wang, W. Liu, W. Zeng, and J. Wang. Object detection in videos by high quality object linking. *IEEE transactions on pattern analysis and machine intelligence*, 42(5):1272–1278, 2019.

[63] P. Tokmakov, J. Li, W. Burgard, and A. Gaidon. Learning to track with object permanence. *arXiv preprint arXiv:2103.14258*, 2021.

[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[65] Q. Wang, Y. Zheng, P. Pan, and Y. Xu. Multiple object tracking with correlation learning. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3876–3886, 2021.

[66] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021.

[67] Y. Wang, K. Kitani, and X. Weng. Joint object detection and multi-object tracking with graph neural networks. *arXiv preprint arXiv:2006.13164*, 2020.

[68] Z. Wang, H. Zhao, Y.-L. Li, S. Wang, P. H. Torr, and L. Bertinetto. Do different tracking tasks require different appearance models? *arXiv preprint arXiv:2107.02156*, 2021.

[69] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang. Towards real-time multi-object tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 107–122. Springer, 2020.

[70] N. Wojke, A. Bewley, and D. Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[71] J. Wu, J. Cao, L. Song, Y. Wang, M. Yang, and J. Yuan. Track to detect and segment: An online multi-object tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12352–12361, 2021.

[72] Y. Xiang, A. Alahi, and S. Savarese. Learning to track: Online multi-object tracking by decision making. In *ICCV*, pages 4705–4713, 2015.

[73] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *CVPR*, pages 3415–3424, 2017.

[74] J. Xu, Y. Cao, Z. Zhang, and H. Hu. Spatial-temporal relation networks for multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3988–3998, 2019.

[75] Y. Xu, Y. Ban, G. Delorme, C. Gan, D. Rus, and X. Alameda-Pineda. Transcenter: Transformers with dense queries for multiple-object tracking. *arXiv preprint arXiv:2103.15145*, 2021.

[76] F. Yang, X. Chang, S. Sakti, Y. Wu, and S. Nakamura. Remot: A model-agnostic refinement for multiple object tracking. *Image and Vision Computing*, 106:104091, 2021.

[77] F. Yang, W. Choi, and Y. Lin. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2137, 2016.

[78] E. Yu, Z. Li, S. Han, and H. Wang. Relationtrack: Relation-aware multiple object tracking with decoupled representation. *arXiv preprint arXiv:2105.04322*, 2021.

[79] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[80] F. Zeng, B. Dong, T. Wang, C. Chen, X. Zhang, and Y. Wei. Motr: End-to-end multiple-object tracking with transformer. *arXiv preprint arXiv:2105.03247*, 2021.

[81] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[82] S. Zhang, R. Benenson, and B. Schiele. Citypersons: A diverse dataset for pedestrian detection. In *CVPR*, pages 3213–3221, 2017.

[83] Y. Zhang, H. Sheng, Y. Wu, S. Wang, W. Ke, and Z. Xiong. Multiplex labeling graph for near-online tracking in crowded scenes. *IEEE Internet of Things Journal*, 7(9):7892–7902, 2020.

[84] Y. Zhang, C. Wang, X. Wang, W. Liu, and W. Zeng. Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild. *arXiv preprint arXiv:2108.02452*, 2021.

[85] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *arXiv preprint arXiv:2004.01888*, 2020.

[86] Z. Zhang, D. Cheng, X. Zhu, S. Lin, and J. Dai. Integrated object detection and tracking with tracklet-conditioned detection. *arXiv preprint arXiv:1811.11167*, 2018.

[87] L. Zheng, M. Tang, Y. Chen, G. Zhu, J. Wang, and H. Lu. Improving multiple object tracking with single object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2453–2462, 2021.

[88] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian. Person re-identification in the wild. In *CVPR*, pages 1367–1376, 2017.

[89] X. Zhou, V. Koltun, and P. Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision*, pages 474–490. Springer, 2020.

[90] X. Zhou, D. Wang, and P. Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[91] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, and M.-H. Yang. Online multi-object tracking with dual matching attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 366–382, 2018.

[92] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.