Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 5, No. 1, 2023

Appearance Awared Detector for MOT: An Enhanced ReID Branch for Tracking Memorize

Hongyu Chen

College of Electrical Engineering, Southwest Minzu University, Chengdu 610206, China

Abstract: The traditional ByteTrack approach to multi-target tracking, focusing on simple, effective algorithms. It performs well in short-term multi-target tracking tasks with 80.3 MOTA, 77.3 IDF1 and 63.1 HOTA on MOT17 30 FPS and is currently ranked number one in the MOTChallenge rankings. But for situations where the camera is moving, or for completing target recovery tasks after a brief loss of position and track information, conventional MOT modules often do not perform as well as they should. To make bytetrack perform better when the motion background pattern is more complex, we used the re-identification module and added appearance information to enhance ByteTrack's MOT process. We verified our improvement on MOT17, which achieved higher results than the original on the metrics Precision and Recall, and was also more suitable for more complex scenarios.

Keywords: ByteTrack approach, MOTChallenge rankings, Multi-target tracking, Re-identification module.

1. Introduction

ByteTrack [1] is a state of the art multi-target tracking method proposed by Yifu Zhang from HUST, based on Tracking-by-detection which is an important method to achieve multi-target tracking. The Tracking-by-detection method uses the similarity (position, appearance, motion and other information) to associate the detection frame between frames to obtain the tracking trajectory. However due to the imperfection of 2D detectors in detection results, In order to balance the true and false positive examples to be suitable for complex actual scenarios, most of the current multi-target tracking methods will select a threshold, and only keep the detection results higher than this threshold for correlation, and then obtain the tracking results. The detection frame is discarded directly. ByteTrack, on the other hand, takes full advantage of low-scoring detection results. Using the similarity between the detection frame and the tracking trajectory, while retaining the high-scoring detection results, the background is removed from the low-scoring detection results, and the real objects (difficult samples such as occlusion, blur, etc.) are mined. Specifically, BYTE will Each detection box is divided into two categories: high-scoring box and low-scoring box according to the score, and a total of two matches are performed. For the first time, the high score box is used to match the previous tracking trajectory. The second time uses the low-scoring box and the tracked trajectories that are not matched to the high-scoring box in the first time (for example, objects that are severely occluded in the current frame and cause the score to drop) for matching. For the detection box that does not match the tracking track and the score is high enough, a new tracking track is created for it. For the tracking trajectories that do not match the detection frame, keep 30 frames, and perform matching when they appear again, thereby reducing missed detection and improving the continuity of the trajectory.

Re-identification (ReID) is a technique designed to search for objects in non-overlapping camera views at different locations by matching query images. ReID can partially replace face recognition to find target objects in video sequences, and can be widely used in security, personal positioning, recommendation systems in physical shopping malls and other fields. The ReID task was first proposed at the 2006 CVPR conference. In 2007, the first ReID dataset VIPeR [2], [3] was released, and then the datasets under different scenarios were open sourced, which further promoted the development of ReID. Around 2015, most of the research on ReID was based on features, such as color, HOG features, etc [4], [5]. The subsequent metric learning was to find the best approximation between features, but this method is suitable for complex scene data, especially when lighting, occlusion, resolution, human pose, perspective, clothing, and background with large changes. With the further development and application of deep learning, deep learning has been gradually applied to ReID and achieved a higher recognition performance than previous methods. [6]–[8].

However in the data association part, ByteTrack only uses Kalman Filter to predict the position of the tracking trajectory of the current frame in the next frame, the IoU between the predicted frame and the actual detection frame is used as the similarity between two matches, and then passes to Hungary algorithm to complete the match. But when the motion pattern is considered complex, the probability of misses detection and track breaks will be higher. At the same time, we found that when there is severe occlusion between two objects or some background (such as doors, trees, etc.) is detected again after being occluded for several frames, The id of the detected object will change incorrectly. And because the appearance features are not used, the tracking effect is very dependent on the detection effect. When the detection effect is not good, it will seriously affect the tracking effect.2. Data Processing and Model design

2. Data Processing and Model Design

2.1. Data Processing

We evaluate our proposed improvements on MOT17 [9] and MOT20 [10]. MOT17 and MOT20 are two current mainstream datasets in MOT task, which mainly focus on the task of pedestrian tracking in dense scenes. The MOT 17 dataset contains 7 training set videos and 7 test set videos. While in MOT20 dataset, 8 new dense crowd sequences are

extracted from 3 scenes, reaching an average of 246 pedestrians per frame. These sequences contain indoor, outdoor, day and night scenes.

We use Mosaic and Mixup data augmentation:

1) Mosaic: The Mosaic data enhancement method is proposed in the YOLOV4 paper. The main idea is to randomly crop four pictures, and then stitch them into one picture as training data. This not only enriches the background of the picture, but also increases the batch size in disguise, reducing its dependence on batch size.

2) Mixup: Mixup is a widely used algorithm for mixedclass enhancement of images, which can mix images between different classes to expand the training dataset. Some of the in-process samples after data enhancement is shown in the Figure 1:

Figure 1. Data Augmentation Samples

2.2. Model design

We add a new Id head in YOLOX, which is obtained from the feature map duplicate of the earlier Conv2D. This method allows for effective extraction of the ReID features of the target. This allows the network to add a focus on ReID information while maintaining the original detection performance, and we eventually add a loss of ID to the loss function. Further, the modified network structure is shown in the figure 2.

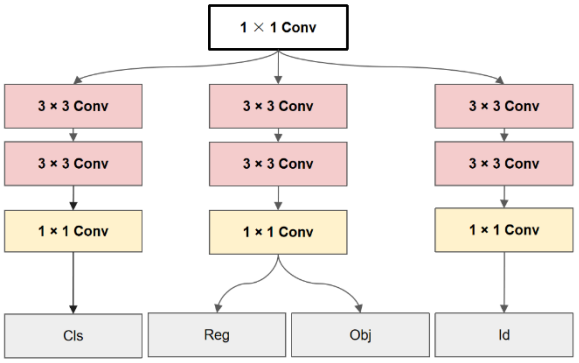

Figure 2. Description of the ID head we added to the model

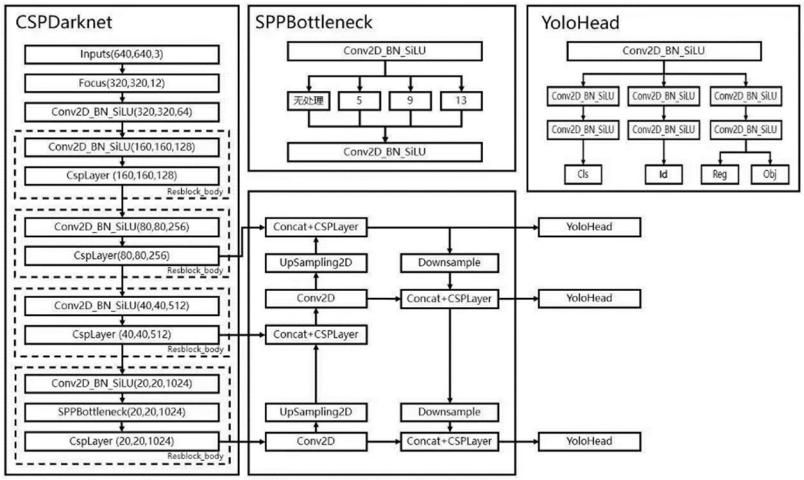

Figure 3. Our Proposed Design for the Whole Model

After that, We use sigmoid focal loss as detection stage loss which remains the same with the origin Bytetracker. The whole loss equation can be described as follows:

$$Loss_{\text{sum}} = W_1 * Loss_{\text{det}} + W_2 * Loss_{\text{reid}} \qquad (1)$$

Where W1 and W2 are weights for detection loss and id loss, we choose W1=0.638 and W2=0.362 as defaults in the following experiments. Thus, we have the whole network defined as Figure 3.

3. Experiment

3.1. Experiment Hyperparameters and Settings

We use Pytorch as our deep learning framework and Python version 3.9.4. Other environment details includes 64-bit Ubuntu18.04 OS and a single NVIDIA GTX 3090 GPU. We train the models for a total of 300 epochs with SGD optimizer and first 5 epochs as warmup on COCO train2017, where the momentum decay factor and weight decay factor are 0.9 and 0.0005,our initial learning rate is set to 0.0004 and is set to decrease for every 40 iterations. In the experiment, we set the high score threshold of detection results as 0.6, the low score threshold as 0.1 while the tracking threshold as 0.7. The IoU threshold in the matching stage is set to 0.2. The temporary storage period of lost objects is 45 frames. The detector uses YOLOX and is pretrained on the COCO dataset.The input image resolution is 1440 × 800 and the shortest side ranges from 576 to 1024 during multiscale training, the data was then augmented using methods shown in the previous section in this paper.

Table 1. Comparison of experimental results on MOT17 and MOT20 data sets

MOT17	With ReID Fusion			Without ReID Fusion			Precision/Recall
	IoU=0.50:0.95	IoU=0.50	IoU=0.75	IoU=0.50:0.95	IoU=0.50	IoU=0.75	
All	0.651	0.851	0.757	0.635	0.827	0.764	Precision
All	0.679	0.874	0.793	0.643	0.852	0.782	Recall
Area	With ReID Fusion			Without ReID Fusion			Precision
	Small	Medium	Large	Small	Medium	Large	
N/A	0.671	0.610	0.712	0.408	0.599	0.690	Precision
N/A	0.579	0.654	0.744	0.451	0.643	0.737	Recall

MOT20 b	With ReID Fusion			Without ReID Fusion			Precision/Recall
	IoU=0.50:0.95	IoU=0.50	IoU=0.75	IoU=0.50:0.95	IoU=0.50	IoU=0.75	
All	0.626	0.877	0.741	0.564	0.828	0.672	Precision
All	0.684	0.821	0.774	0.643	0.786	0.692	Recall
Area	With ReID Fusion			Without ReID Fusion			Precision
	Small	Medium	Large	Small	Medium	Large	
N/A	0.341	0.621	0.706	0.319	0.603	0.701	Precision
N/A	0.441	0.549	0.632	0.423	0.652	0.628	Recall

3.2. Quantitative Comparison Experiment

In this section we show comparison with performance differences between our improved ByteTrack and origin ByteTrack on the MOT17 and MOT20 datasets at TABLE I. As shown in TABLE1, the improved method proposed in this paper has been systematically enhanced on MOT17 and MOT20. In MOT17 the improvement averaged about 2% across the IoU intervals, while in MOT20 we achieved close to 5%. The improvement on Recall is not significant, probably because the original model is already good enough in terms of recall performance, but the added ReID information can significantly improve Precision performance.

4. Conclusion

In this paper, we propose a method to optimize ByteTrack by using reid to extract appearance feature information. We make better use of the appearance feature information by designing the appearance feature extraction network. The network structure of this paper is to add a reid branch based on the original YOLOX Decoupled Head. The Reid branch network extracts the appearance feature information, and at the same time uses the Reid loss function. The experimental results show that the performance of the proposed method on the MOT17, MOT20 has been improved to a certain extent compared with the original ByteTrack, indicating that the method in this paper can work better in complex motion and diverse motion patterns.

Acknowledgment

This work was supported by the Innovative Research Project for Postgraduates of Southwest Minzu University (CX2021061).

References

[1] Y. Zhang, P. Sun, Y. Jiang,et al. “Bytetrack: Multi-object tracking by associating every detection box,” CoRR, vol. abs/2110.06864, 2021. [Online]. Available: https://arxiv.org/abs/2110.06864

[2] D. Gray, and H. Tao.Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features, in Proc. European Conference on Computer Vision (ECCV), 2008.

[3] D. Gray, S. Brennan, and H. Tao, “Evaluating Appearance Models for Recognition, Reacquisition, and Tracking,” Performance Evaluation of Tracking and Surveillance (PETS). IEEE International Workshop on, 2007.

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

[5] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning Discriminative Features with Multiple Granularities for Person Re-Identification. In Proceedings of the 26th ACM international conference on Multimedia. Association for Computing Machinery, New York, NY, USA, 274–282.

[6] He S, Luo H, Wang P, et al. Transreid: Transformer-based object re-identification[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 15013-15022.

[7] Zou Y, Yang X, Yu Z, et al. Joint disentangling and adaptation for cross-domain person re-identification[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer International Publishing, 2020: 87-104.

[8] Zhai Y, Ye Q, Lu S, et al. Multiple expert brainstorming for domain adaptive person re-identification[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. Springer International Publishing, 2020: 594-611.

[9] A. Milan, L. Leal-Taix´e, I. Reid, S. Roth, and K. Schindler.Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831, 2016.

[10] Dendorfer P , Rezatofighi H , Milan A , et al. MOT20: A benchmark for multi object tracking in crowded scenes: arXiv, 10.48550/arXiv.2003.09003[P]. 2020.