applied sciences

MDPI

Article

Non-Maximum Suppression Performs Later in Multi-Object Tracking

Hong Liang, Ting Wu *, Qian Zhang and Hui Zhou

College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China; Liangh@upc.edu.cn (H.L.); zhangqian8266@163.com (Q.Z.); zhinupc@163.com (H.Z.)
* Correspondence: tingw028@163.com

Abstract: Multi-object tracking aims to assign a uniform ID for the same target in continuous frames, which is widely used in autonomous driving, security monitoring, etc. In the previous work, the low-scoring box, which inevitably contained occluded target, was filtered by Non-Maximum Suppression (NMS) in a detection stage with a confidence threshold. In order to track occluded target effectively, in this paper, we propose a method of NMS performing later. The NMS works in tracking rather than the detection stage. More candidate boxes that contain the occluded target are reserved for trajectory matching. In addition, unrelated boxes are discarded according to the Intersection over Union (IOU) between the predicted and detected box. Furthermore, an unsupervised pre-trained person re-identification (ReID) model is applied to improve the domain adaptability. In addition, the bicubic interpolation is used to increase the resolution of low-scoring boxes. Extensive experiments on the MOT17 and MOT20 datasets have proven the effectiveness of tracking occluded targets of the proposed method, which achieves an MOTA of 78.3%.

Keywords: multi-object tracking; deep learning; person re-identification

Citation: Liang, H.; Wu, T.; Zhang, Q.; Zhou, H. Non-Maximum Suppression Performs Later in Multi-Object Tracking. *Appl. Sci.* **2022**, *12*, 3334. https://doi.org/10.3390/app12073334

Academic Editors: Antonio Fernández-Caballero and Luis Javier Garcia Villalba

Received: 3 March 2022
Accepted: 22 March 2022
Published: 25 March 2022

1. Introduction

Multi-object tracking (MOT) is a deep learning task used to distinguish multiple targets appearing in a video with different IDs, and has a wide range of applications in the fields of intelligent monitoring and autonomous driving. The current mainstream MOT methods are based on the tracking by detecting mode, and the tracking effect largely depends on the target detection method used. In some multi-target tracking competitions, the detector is fixed. After inputting video frames into the object detector [1–3], adetection bounding boxes are generated, and the same target between two frames is associated with obtaining the trajectory of the target.

At present, the widely used tracking methods first model the appearance of the detected targets, then calculate the affinity of the target features between frames, and finally match the targets through a matching algorithm. However, usually, appearance modeling will calculate the affinity between each trajectory and the global detection bounding boxes, which is prone to ID Switches. The introduction of motion prediction will make the tracking more robust.

Sometimes, the detector recognizes the background as an object or the object is occluded, and a low-confidence detection bounding box, also known as a low-scoring box, appears. Due to its uncertainty, this type of box can easily disappear in a certain frame, resulting in tracking failure. In order to solve this situation, most of the current practice is to set the confidence threshold of the object detection network, and the detection boxes below this threshold will be filtered out. The problem with this approach is that, when the target is occluded, it cannot be associated with the previous frame. The matching problem of low-scoring boxes is crucial.

In the MOT method of tracking by detecting mode, tracking and detection are two parts and do not affect each other. However, the tracking task is the main task, and the

detection task needs to serve the main tracking task. Although more and more one-stage MOT methods that integrate detection and tracking have been proposed, the accuracy is not optimal. In general work, NMS [4] in two-stage MOT is applied to the detection stage, first filtering low-scoring boxes, then removing redundant detection boxes and retaining the best one as the target box to match trajectories. However, while retaining the high-scoring frame, NMS may filter out the real low-scoring boxes due to occlusion, so that the filtered real frame cannot match the trajectory, resulting in tracking failure. Therefore, this paper improves the two-stage MOT method, and performs NMS in tracking instead of detection, which effectively avoids the boxes that need to be tracked from being filtered out.

The current method is fair for high-scoring boxes and low-scoring boxes, and both directly use them together as target boxes to extract features. However, low-scoring boxes usually have smaller resolution and poor quality before feature extraction. In order to deal with the challenge of MOT, this paper proposes a MOT method that introduces bicubic interpolation. The low-scoring box has a high probability of being a distant target, its size is small, and the resolution is low after reshape. After interpolation, the resolution of the low-segment box can be improved, which is helpful for extracting its features.

When calculating the affinity, the person ReID network is usually used to extract the target appearance information. Person ReID, as a task in computer vision, is trained on datasets dedicated to person ReID. Therefore, when applyng person ReID to MOT, the domain gap caused by the different scopes has an impact on the matching results. Through unsupervised pre-training on the dataset, the obtained person ReID model effectively reduces the domain gap and has good performance on different datasets.

In conclusion, the main contributions of this paper are as follows:

- In order to avoid good detection boxes being filtered out and reducing the impact of detection tasks on tracking, a method of performing NMS later in tracking is proposed. NMS does not play a role in the detection stage, but plays a role later, serving the tracking and prediction stage, effectively improving the performance of MOT.
- Experimental results show that the method proposed in this paper can effectively solve the occlusion problem.
- The effectiveness of the method proposed in this paper is verified through a large number of experiments, and good results have been achieved on the MOT17 and MOT20 datasets, and the state-of-the-art FP and ML metrics have been achieved.

2. Related Work

MOT has attracted significant interest from researchers in recent years. MOT algorithms can be divided into two categories according to whether the detection and prediction use the same network: one-stage and two-stage MOT.

2.1. One Stage MOT

In order to balance the speed and accuracy of MOT, one-stage MOT methods have been continuously proposed. In JDE [5], object detection and appearance embeddings are allowed to be learned in a shared model. The competition of detection and tracking tasks impairs the learning of task-relevant representations, which impairs the performance of tracking. To address this issue, CSTrack [6] effectively motivates each branch to learn task-relevant representations. Previous work usually treats the ReID task as a secondary task, whose accuracy is severely affected by the main detection task. As a result, the network is biased towards the main detection task, which is unfair to the ReID task, which FairMOT [7] solves simply and effectively. In CorrTrack [8], adding spatiotemporal correlation to [7] makes the tracking results more robust. CenterTrack [9] is borrowed from [3] and implements end-to-end tracking by inputting the heatmap of the previous frame and the current frame.

2.2. Two-Stage MOT

With the development of deep learning, object detection has achieved relatively high accuracy and effect. One-stage MOT methods are not optimal in accuracy. MOT based on detecting largely benefit from the development of target detection, so they are better than one-stage methods in accuracy.A novel end-to-end learning neural network, MATNet [10], leverages motion cues as a bottom-up signal to guide the perception of object appearance. SORT [11] is a typical tracking by detecting the MOT method, using a simple Kalman filter and Hungarian algorithm for tracking, which achieving the most accurate results at that time. This paper refers to ByteTrack [12], which divides the detection boxes into two categories, including high-scoring boxes and low-scoring boxes, and performs matching strategies, respectively, which significantly improves the metrics of multi-target tracking. In TransMOT [13], a sparse weighted graph is used to represent the spatial relationship between objects, encoding and decoding are performed according to transFormer, and an allocation matrix is generated for matching. Ref. [14] propose two-stage methods for pixel-level tracking.

2.3. MOT with Person ReID

When matching the target, since the occluded target reappears, the target needs to be re-identified, so ReID is introduced into the tracking phase. DeepSORT [15] proposed by Nicolai Wojke integrates appearance information in [11], and extracts features according to appearance information to calculate affinity, which effectively improves the problem of ID Switch. The method proposed by Chen et al. [16] uses a large number of person ReID data sets for training, and obtains the corresponding appearance representation of person ReID, which improves the recognition ability of the tracker.

3. Materials and Methods

The process of the method proposed in this paper is shown in Figure 1. The method in this paper is two-stage, including the detection and tracking stage. In the detection stage, the detection boxes are generated by the object detection algorithm [1–3], which is input into the tracking stage as the detection result. The tracking stage predicts the position of the target in the frame according to the tracking result of $t - 1$ frames, and uses the IOU value of detection results and predicted position to perform NMS to obtain the predicted boxes. The prediction boxes are divided into two categories according to the relationship between the confidence and the confidence threshold, including high-scoring boxes and low-scoring boxes. Match the high-scoring boxes first, and then match the unsuccessfully matched trajectories with the low-scoring boxes to obtain the final tracking result.

3.1. Performing NMS Later

In most of the work now, NMS is used in the detection stage. After NMS, some low-scoring or redundant detection boxes can be filtered out, and some high-scoring detection results can be obtained. This paper believes that the application of NMS in the detection stage will make the results more biased towards the detection task, and the tracking in the multi-target tracking task is the main task of the research. Therefore, it is unreasonable to filter out low-scoring and redundant boxes in the detection stage, which would filter out candidate boxes that are useful for tracking. In this paper, we try to postpone NMS, that is, use NMS in the tracking stage instead of NMS in the detection stage, so that the results are more focused on the main tracking task.

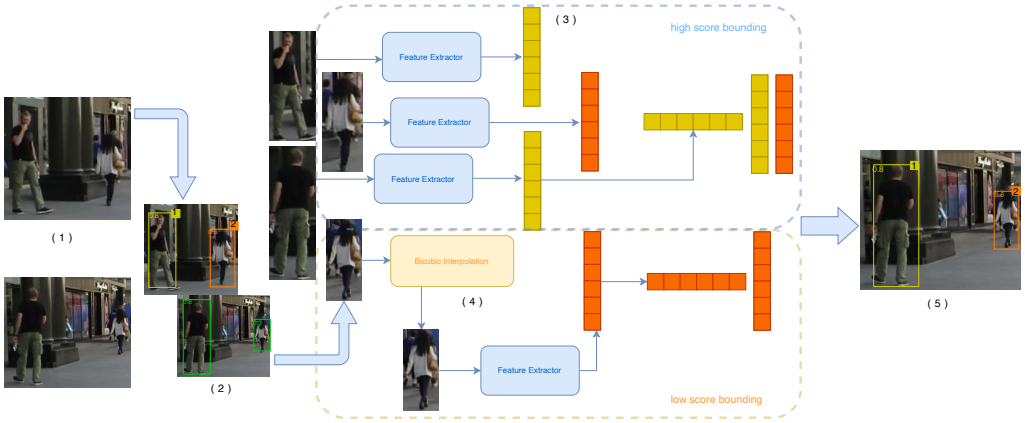

Figure 1. Workflow of our method: Given the video frame (**1**), the picture in the first line is the previous frame of the video, and the picture in the second line is the result of YOLOX detection. The tracking result of the previous frame and the image of this frame are predicted by the Kalman filter and matched with the detection result through IOU (**2**). For the high-scoring boxes in the detection boxes, after extracting its features, calculate the affinity with trajectories of the previous frame, and then match with each track (**3**); For boxes with lower scores, feature extraction is performed on the image after bicubic interpolation, and the affinity with the unmatched trajectories is calculated and then matched (**4**). Finally, the tracked result (**5**) is obtained.

3.1.1. NMS Performs in Tracking

This paper removes the NMS in the detection; therefore, all detection boxes obtained by tracking are used as detection results for tracking. In the tracking stage, after using a Kalman filter to predict the target position, NMS processing is performed according to the IOU value of the calculated prediction result and the real frame. The detection boxes below the Intersection over Union (IoU) threshold will be filtered out, leaving several detection boxes that are close to the predicted box in position, and the ground-truth boxes with the highest confidence are further screened out as the candidate boxes. The workflow of NMS performs in tracking is shown in Figure 2.

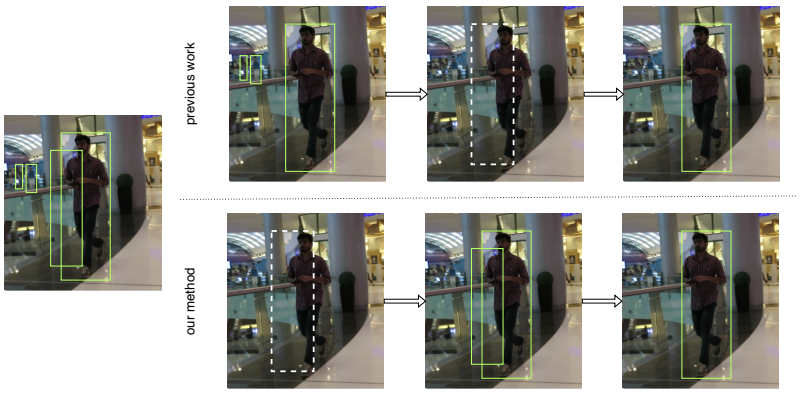

Figure 2. Workflow of NMS performing in tracking and previous work: The first column of pictures on the left is the detection result *a* without NMS performed. The first row of pictures on the right is previous work, the first column of pictures is the gt boxes after performing NMS on *a*, the white box of the second column of pictures is the target position predicted by Kalman filtering, and the third column of pictures is the result of IOU calculation of gt boxes and predictions. The second row of pictures is our method, the first column of pictures is the target position predicted by Kalman filtering, and the second column of pictures is the result of the detection result and the predicted NMS according to the IOU. The third column of pictures is the result of further taking the box with higher confidence as the prediction result.

When a person of interest appears in a frame for the first time, all detection boxes will be retained in the detection stage. In the tracking stage, these detection boxes calculate IOU according to the position predicted by a Kalman filter. The detection box with IOU higher than the threshold will be used as the candidate box, and the ReID model will be used to judge whether it is an existing target. If it is not an existing target, it will filter out the error detection boxes through NMS together with all detection boxes. After NMS, the detection box that is neither an existing target nor an unfiltered one will be regarded as a new target and assigned a new ID.

Input the video frame into the detection algorithm and obtain a series of detection boxes $B_t = \{B_t^1, B_t^2, B_t^3, \ldots, B_t^n\}$. Among them, $B_t^n(x_1, y_1, x_2, y_2)$ represents the ith detection frame of the t frame, and the predicted target position is $P(x_1', y_1', x_2', y_2')$. The IOU value c_i of each detection box and predicted position can be calculated according to the coordinates of the detection box and the predicted position. c_i can be formulated as:

$$c_i = \frac{Intersection(B_t^i, P)}{Union(B_t^i, P)} \tag{1}$$

$$\hat{B}_t = \{B_t^i \in B_t | c_i > \theta\} \tag{2}$$

As (2), according to the calculated IOU value c_i of each detection box and predicted position, the box whose IOU value is less than the IOU threshold θ is filtered out, and the boxes $\hat{B}_t$ that are closer to the predicted position are obtained. Then, perform non-maximum suppression on the boxes in $\hat{B}_t$ as (3), filter out redundant detection boxes, and obtain the final candidate box B_t' as a candidate option for trajectory matching:

$$B_t' = NMS(\hat{B}_t) \tag{3}$$

3.1.2. ID Match

The performance of multi-target tracking largely depends on the accuracy of the detection frame. When the detection frame and the trajectory are ID matched, there are great differences in the resolution and confidence of the detected target frame. Tests on a large number of datasets show that boxes with lower confidence account for a large proportion. Moreover, it is difficult to match the boxes with low confidence, and it is easy to have disordered matching or unsuccessful matching. In the method of this paper, the detection boxes with higher scores are first matched, and then the boxes that are not successfully matched in the first match are matched with the boxes with low scores together. After the two matches are completed, the high-scoring box that has not been successfully matched is regarded as the new target of the frame t as a new track. A low-scoring box that still does not match is considered poor quality detection and is removed.

3.2. Bicubic Interpolation

Applying bicubic interpolation to process the detection box images with low scores to improve the resolution of the target is of great significance for extracting target features. Bicubic interpolation uses the 4×4 pixel area where the target pixel is located for interpolation calculation. The interpolation basis function [17] is:

$$W(x) = \begin{cases} (a+2)|x|^3 - (a+3)|x|^2 + 1, & \text{if } |x| \leq 1 \\ a|x|^3 - 5a|x|^2 + 8a|x| - 4a, & \text{if } 1 < |x| < 2 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

As [17], a is -0.5. When (x, y) is a point in the rectangular subdivision $[x_j, x_{j+1}] \times [y_k, y_{k+1}]$, the bicubic interpolation function is:

$$g(x, y) = \sum_{l=-1}^{2} \sum_{m=-1}^{2} g(x_{j+l}, y_{k+m}) W(x - x_{j+l}) W(y - y_{k+m}) \tag{5}$$

3.3. Unsupervised Pre-Training for ReID

Some works believe that ReID will not greatly improve the accuracy of the MOT challenge, but this paper believes that ReID is indispensable for solving the problem of occlusion and irregular motion.

Person ReID has always played an important role in MOT tasks. After the object detection model obtains the position and size of the target, the patch containing the object will be cropped from the original input image and input into the module, so as to obtain the feature embedding of the object in the image as the discriminating condition of the target trajectory. Although person ReID plays an irreplaceable role in MOT, the existing methods generally integrate the model trained on the person ReID dataset directly into the whole model, which is too rough. Moreover, due to the large domain gap between the person ReID dataset and the MOT dataset, the feature embeddings captured by the person ReID model are less credible.

To solve the above problems, this paper adopts the strategy in [18]. First, the MoCo v2 [19] self-supervised model is used to perform unsupervised pre-training on the large-scale unlabeled dataset LUPerson, and the domain adaptability of the person re-ID model is improved through the sample identity diversity of the large-scale dataset and data augmentation strategies. After pre-training, this paper fuses multiple person ReID datasets such as DukeMTMC [20], Market1501 [21], and MSMT17 [22] to form a labeled hybrid dataset and conducts supervised training here to further improve the performance of the model on the premise of ensuring domain adaptability.

In the past, the person ReID model generally used ResNet-50 [23] as the backbone network. Although the accuracy is high, the huge amount of parameters greatly improves the inference time of the tracking process, which is not conducive to the application of this technology in real scenarios. To this end, this paper uses OSNet [24] to replace the original feature extraction network, which greatly reduces the amount of parameters and computation.

4. Results

4.1. Datesets and Settings

4.1.1. Datasets

In the past few years, many MOT datasets have been proposed and applied. MOT Challenge is the most commonly used and most convincing benchmark for MOT, and MOT16 [25] includes 14 challenge video sequences (7 train, 7 test) from pedestrian videos captured with moving or stationary cameras. MOT17 Det is based on MOT16 video sequences, using new and more accurate ground truth. MOT20 [26] is the Pedestrian Detection Challenge proposed by Patrick Dendorfer and his team in 2020, including eight challenging video sequences (4 train, 4 test). Compared to previous challenges, MOT20 has more crowded pedestrians and is more difficult to track. This paper evaluates more commonly used datasets such as MOT17 Det [25] and MOT20 [26].

4.1.2. Evaluation Metrics

The study benchmarks the proposed method using several standard evaluation metrics, including Multiple Object Tracking Accuracy (MOTA) [27], IDF1 Score [28], Mostly tracked targets (MT), Mostly lost targets (ML), the number of False Positives (FP), the number of False Negatives (FN), and the number of Identity Switch (IDs) [29].

4.1.3. Implementation Details

The paper design framework follows deepSORT [15] structure. The method uses YOLO X as the detector and is trained on the YOLOX-x model pre-trained on the COCO dataset. The research trains the model on public datasets such as Crowdperson [30], KITTI [31] and the train set of MOT17 [25], MOT20 [26]. The trained SGD optimizer is 30 epochs and the batch size is 10. The initial learning rate is 0.001, and, after the tenth epoch,

the learning rate is 0.0005. The study sets the IOU threshold to 0.35 and the low-scoring boxes threshold to 0.55. The MOT network is trained on a Titan RTX GPU.

4.2. Comparison with State-of-the-Art

This paper compares the proposed method with the current state-of-the-art (SOTA) method. These methods include one-stage methods such as FairMOT [7], Corrtracker [8], MOTR [32], TrackFormer [33], CSTrack [6]; and detection-based two-stage methods such as TransMOT [13], and MPNTrack [34].

4.2.1. Comparison on MOT17

As shown in Table 1, the method proposed in this paper is best on MOTA, IDF1, ML, and FN on the MOT17 dataset, and MT is second only to the TransMOT. It is proved that the method in this paper can effectively reduce false negatives and most lost while having a good tracking effect.

Table 1. Comparison with the state-of-the-art methods over the dataset MOT17 test set. The best results are shown in **bold**. The next best result is underlined.

Method	Mode	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓
MOTR [32]	one-stage	67.4	67.0	34.6	24.5	32,355	149,400	1992
CorrTracker [8]	one-stage	76.5	73.6	47.6	12.7	29,808	99,510	3369
FairMOT [7]	one-stage	73.7	72.3	43.2	17.3	27,507	117,477	3303
CSTrack++ [6]	one-stage	70.6	71.8	38.2	17.8	-	-	1071
TrackFormer [33]	one-stage	62.5	60.7	-	-	32,828	174,921	3917
TransMOT [13]	two-stage	76.7	75.1	**51.0**	16.4	36,231	125,665	**1042**
MPNTrack [34]	two-stage	58.8	61.7	28.8	33.5	**17,413**	213,594	1185
Our method	two-stage	**78.3**	**76.1**	50.8	**11.8**	33,754	**93,797**	1435

4.2.2. Comparison on MOT20

As shown in Table 2, the method proposed in this paper is close to the state of the art in MOTA, IDF1, FP, IDs and other metrics on the MOT20 dataset. TransMOT leverages graph transformers to efficiently model the spatial and temporal interactions among numerous objects. In addition, TransMOT proposed a cascade correlation framework to deal with low score detection and long-term occlusion, which further improved the tracking speed and accuracy. However, it requires large computational resources to model in TransMOT. Our two-stage tracker can outperform other methods except TransMOT for tracking effect without a lot of calculations.

Table 2. Comparison with the state-of-the-art methods over the dataset MOT20 test set. The best results are shown in **bold**. The next best result is underlined.

Method	Mode	MOTA↑	IDF1↑	MT↑	ML↓	FP↓	FN↓	IDs↓
CorrTracker [8]	one-stage	65.2	69.1	66.4	8.9	79,429	95,855	5183
FairMOT [7]	one-stage	61.8	67.3	68.8	**7.6**	103,440	88,901	5243
TransMOT [13]	two-stage	**77.5**	**75.2**	**70.7**	9.1	**34,201**	**80,788**	**1615**
Our method	two-stage	75.7	71.4	68.3	9.7	42,134	90,833	2026

4.3. Ablation Study

In order to prove the effectiveness of each module of the method proposed in this paper, ablation experiments were conducted on the MOT20 train. The baseline is the YOLOX [2] detector. IOU stands for using IOU to calculate matching, NMS represents the NMS post-module, UR represents Unsupervised pre-training for ReID, and BI represents Bicubic interpolation module. The research tests the performance of each method on five indicators including MOTA, IDF1, FP, FN, and IDs.

As shown in Table 3, the experimental results verify the effectiveness of each module of the method. Comparing the first and second rows of the table, MOTA increased by 8.8%, IDF1 increased by 0.8%, and both FP and IDs decreased, which proved that NMS post-position played a role and improved the performance of MOT. Comparing the second and third rows of the table, MOTA is significantly improved by 0.3%, which proves that unsupervised pre-training can make the model have the ability to reduce the domain gap and improve the generalization ability of the model on the data set. Comparing the 3rd and 4th rows of the table, all indicators are improved, which proves the effectiveness of Bicubic interpolation.

Table 3. Ablation results on MOT20.

Method	MOTA↑	IDF1↑	FP↓	FN↓	IDs↓
Baseline + IOU	66.3%	70.4%	56,794	102,512	2072
Baseline + NMS + IOU	75.1%	71.2%	**40,654**	96,341	2043
Baseline + NMS + UR	75.4%	71.2%	42,372	98,784	2032
Baseline + NMS + UR + BI	**75.7%**	**71.4%**	42,134	**90,833**	**2026**

4.4. Occlusion Problem

The purpose of postponing NMS to the tracking stage is to prevent the two-stage multi-object tracking method from filtering out the detection boxes useful for tracking in the detection stage and reduce the negative effects of detection on tracking. At the same time, we find that this is also effective for solving the occlusion problem. In multi-object tracking, when objects occluded each other, the occluded object will generate a low-score detection box. The low-score detection box of the occluded target overlaps with the detection frame of the obstacle (people) and is filtered by NMS, which makes the occluded target untraceable. After the NMS is postponed to the tracking stage, the detection box with a low score will be preserved, which effectively solves the problem that the occluded target is lost in tracking. In addition, in ID matching, our method matches the candidate boxes with high score and low score, respectively, which further enhances the matching of low score boxes. Figure 3 shows that our method is effective for the occlusion problem.

As shown in Figure 3, when NMS acts on the detection stage, the object will be lost when it is blocked. However, when NMS is postponed to the tracking stage, the occluded objects will still be tracked.

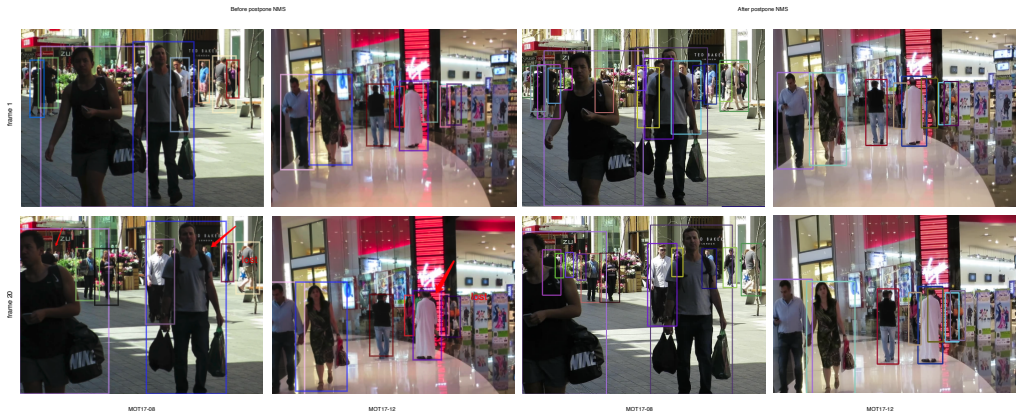

Figure 3. Comparison of results before and after postponing NMS. The position pointed by the red arrow is an untraceable object after being occluded.

5. Discussion

The effectiveness of the method proposed in this paper comes from the processing and matching of low-scoring boxes. As shown in Figure 4, the tracking visualization results verify the effectiveness of the method proposed in this paper. In the sequence of MOT17-07,

the target with ID 7 is still accurately tracked even when it is almost completely occluded, which directly verifies the importance of performing NMS later in the tracking stage. In addition, there are different degrees of occlusion problems in several other sequences, and the method in this paper successfully tracks the occluded targets. In particular, the sequence in MOT20 has a larger number of targets, and the targets are more likely to occlude each other, which is a challenge for tracking. This study is still accurate and robust in such cases. Accuracy and robustness mainly come from two aspects: (1) Motion prediction is used in multi-target tracking. Motion prediction enables the prediction results to be combined with the previous tracking results and avoids comparison with the global target frame when matching the tracking trajectory, which can make the results more accurate and robust; (2) An unsupervised pretrained person ReID network is used. The MOT dataset has different scenes, picture quality, and shooting styles on different sequences, so there is a clear domain gap between different test sequences. Unsupervised pre-training of person ReID network improves the model's ability to narrow the domain gap and has accurate recognition accuracy in different scenarios.

Figure 4. Visualization of tracking results on MOT17 and MOT20 test datasets. Boxes of the same color and numbers in the upper left corner of the box indicate the same tracking target. The method proposed in this paper has visible and effective effects on object density and occlusion.

6. Conclusions

This paper focuses on the challenging occlusion problem of MOT. The proposed method performs NMS later to avoid discarding low-scoring boxes that are useful for tracking, and to sequentially match high-scoring and low-scoring boxes. Bicubic interpolation is used to improve the quality of low-scoring boxes, which effectively improves the matching accuracy of low-scoring boxes. At the same time, by performing unsupervised pre-training on the person ReID network, the ability of the ReID network to reduce the domain gap is improved. The experimental results verify that the method proposed in this paper can effectively solve the occlusion problem while obtaining good tracking results.

Author Contributions: Resources, Q.Z.; data curation, H.Z.; writing—original draft preparation, T.W.; writing—review and editing, T.W.; supervision, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Science Foundation of Shandong Province under Grant No. ZR2020MF005.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to Data not being completely sorted out.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

MOT	Multiple Object Tracking
NMS	Non-Maximum Suppression
IOU	Intersection over Union
ReID	Re-Identification

References

1. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
2. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
3. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
4. Neubeck, A.; Van Gool, L. Efficient non-maximum suppression. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; Volume 3, pp. 850–855.
5. Wang, Z.; Zheng, L.; Liu, Y.; Li, Y.; Wang, S. Towards real-time multi-object tracking. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin, Germany, 2020; Part XI 16, pp. 107–122.
6. 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* **2020**, arXiv:2010.12138.
7. Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; Liu, W. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *Int. J. Comput. Vis.* **2021**, *129*, 3069–3087. [CrossRef]
8. 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, Nashville, TN, USA, 20–25 June 2021; pp. 3876–3886.
9. Zhou, X.; Koltun, V.; Krähenbühl, P. Tracking objects as points. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin, Germany, 2020; pp. 474–490.
10. Zhou, T.; Li, J.; Wang, S.; Tao, R.; Shen, J. Matnet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE Trans. Image Process.* **2020**, *29*, 8326–8338. [CrossRef] [PubMed]
11. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468.
12. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. *arXiv* **2021**, arXiv:2110.06864.
13. Chu, P.; Wang, J.; You, Q.; Ling, H.; Liu, Z. Spatial-temporal graph transformer for multiple object tracking. *arXiv* **2021**, arXiv:2104.00194.
14. Zhou, T.; Li, J.; Li, X.; Shao, L. Target-aware object discovery and association for unsupervised video multi-object segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6985–6994.
15. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649.
16. Chen, L.; Ai, H.; Zhuang, Z.; Shang, C. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
17. Keys, R. Cubic convolution interpolation for digital image processing. *IEEE Trans. Acoust. Speech Signal Process.* **1981**, *29*, 1153–1160. [CrossRef]
18. Fu, D.; Chen, D.; Bao, J.; Yang, H.; Yuan, L.; Zhang, L.; Li, H.; Chen, D. Unsupervised Pre-training for Person Re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14750–14759.
19. Chen, X.; Fan, H.; Girshick, R.; He, K. Improved baselines with momentum contrastive learning. *arXiv* **2020**, arXiv:2003.04297.
20. Zheng, Z.; Zheng, L.; Yang, Y. Unlabeled Samples Generated by GAN Improve the Person Re-Identification Baseline in Vitro. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
21. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision, Las Condes, Chile, 11–18 December 2015; pp. 1116–1124.

22. Wei, L.; Zhang, S.; Gao, W.; Tian, Q. Person transfer gan to bridge domain gap for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 79–88.
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
24. Zhou, K.; Yang, Y.; Cavallaro, A.; Xiang, T. Omni-scale feature learning for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 28 October 2019; pp. 3702–3712.
25. Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; Schindler, K. MOT16: A benchmark for multi-object tracking. *arXiv* **2016**, arXiv:1603.00831.
26. 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* **2020**, arXiv:2003.09003.
27. Kasturi, R.; Goldgof, D.; Soundararajan, P.; Manohar, V.; Garofolo, J.; Bowers, R.; Boonstra, M.; Korzhova, V.; Zhang, J. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 319–336. [CrossRef] [PubMed]
28. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2016; Springer: Berlin, Germany, 2016; pp. 17–35.
29. Li, Y.; Huang, C.; Nevatia, R. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 2953–2960.
30. Shao, S.; Zhao, Z.; Li, B.; Xiao, T.; Yu, G.; Zhang, X.; Sun, J. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv* **2018**, arXiv:1805.00123.
31. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
32. Zeng, F.; Dong, B.; Wang, T.; Chen, C.; Zhang, X.; Wei, Y. MOTR: End-to-End Multiple-Object Tracking with TRansformer. *arXiv* **2021**, arXiv:2105.03247.
33. Meinhardt, T.; Kirillov, A.; Leal-Taixe, L.; Feichtenhofer, C. Trackformer: Multi-object tracking with transformers. *arXiv* **2021**, arXiv:2101.02702.
34. 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, Seattle, WA, USA, 13–19 June 2020; pp. 6247–6257.