Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 2, No. 3, 2025

GradsSORT: Research on Multi-Object Tracking Algorithm Based on Pseudo-Gradient

Xiufen Fu *, Haifeng Sang

School of Information Science and Engineering, Shenyang University of Technology, Liaoning, China

* **Corresponding author**: Xiufen Fu (Email: xiufenfu202412@163.com)

Abstract: To address the issue of low tracking accuracy for occluded targets in complex scenes, we propose a multi-object tracking algorithm (GradsSORT) based on pseudo-gradient. First, targets are categorized into different gradient layers according to their degree of occlusion. Then, a hierarchical association strategy is adopted during data matching, prioritizing the association of detection boxes in lower gradient layers. This approach helps remove occluders for occluded targets, making it more effective for tracking tasks in crowded occlusion scenarios. To verify the effectiveness of the proposed algorithm, we implemented improvements based on the HybridSORT tracker and conducted experiments on the public datasets MOT17 and MOT20. The experimental results show that on the MOT20 dataset, the overall tracking performance metrics—HOTA, tracking accuracy (MOTA), and tracking stability (IDF1)—reached 64.8 (+0.5), 76.1 (+0.6), and 79.6 (+1.6), respectively. On the MOT17 dataset, the three key reference metrics, HOTA, MOTA, and IDF1, also improved to 63.5, 78.7, and 78.9, respectively. Our algorithm provides an effective solution for tracking targets in crowded occlusion scenarios.

Keywords: HybridSORT, Multi-object tracking, Pseudo-gradient.

1. Introduction

Multi-Object Tracking (MOT) [1] is an important research direction in the field of computer vision, widely applied in autonomous driving, video surveillance, and robotics. Tracking-by-Detection (TBD) [2] algorithms are non-end-to-end MOT methods, where object detection is first performed on each frame of a video, and then the detection results are used for object tracking (data association). The TBD paradigm decomposes the MOT task into four subtasks: object detection, feature extraction, data association, and trajectory processing. This approach has seen significant development.

It is well known that occlusion is a long-standing challenge in MOT. Previous algorithms have considered both spatial and appearance information. In the object detection stage, many high-performing detectors exist, such as YOLOX [3]. In high-FPS benchmark tests, the time interval between frames is short, and object motion is minimal (bounding box position changes are small), making it approximately linear. This allows spatial position information to serve as an accurate metric for short-term associations. For feature extraction, appearance features such as object color are commonly used. Since the introduction of ReID in DeepSORT [4], subsequent works like BoT-SORT [5] and MOTDT [6] have also incorporated ReID.

In the data association stage, both strong and weak cues are explicitly or implicitly utilized, including spatial information (commonly using IOU) and appearance features. Motion direction (OC-SORT [7]), confidence scores (ByteTrack [8]), and bounding box height are also incorporated. HybridSORT [9] provides a more comprehensive approach by integrating multiple strong and weak signals. These designs are reasonable, as strong cues offer robust instance-level distinction, while weak cues serve as valuable supplements. Finally, trajectory processing techniques improve issues such as duplicate detections of the same object and fragmented trajectories, as seen in StrongSORT [10]. Additionally, SparseTrack [11] decomposes scenes using pseudo-depth to indirectly handle sparse and dense crowds.

This paper takes a target-centric perspective by classifying objects into different gradient layers based on the degree of occlusion at their respective locations, effectively mitigating occlusion issues.

The main contributions of this paper are as follows:

(1) Pseudo-gradient estimation method: A technique to estimate occlusion levels from 2D images, enabling finer sparse decomposition of clustered regions in the scene.

(2) Layered association strategy: Prioritizing association in lower-gradient layers, which facilitates the tracking of occluded objects.

2. Related Work

To some extent, the performance of a tracker depends on its ability to handle occlusions. Recent works have attempted to address the occlusion problem from different perspectives. For example, MotionTrack [12] learns motion pattern trajectories and effectively simulates occluded trajectories by combining them with historical information. MOTR [13] is an end-to-end MOT framework that handles newly emerging targets and employs a multi-frame training approach using detection queries and trajectory queries. Due to the relative independence between queries and sufficient temporal training, MOTR performs well in temporal modeling across occluded time steps. DP-MOT [14] proposes a Subject-Ordered Depth Estimation (SODE) method, which automatically sorts the depth positions of detected objects in a 2D scene in an unsupervised manner. By constructing a pseudo-3D Kalman filter, DP-MOT achieves robust association for occluded targets. BoT-SOR [5] combines Camera Motion Compensation (CMC) and IoU-ReID fusion, integrating motion and appearance cues to achieve accurate tracking of occluded objects. OUTrack [15] employs an unsupervised re-identification module and an occlusion-aware module to predict where occlusions occur,

compensating for missed detections. [16] segments the trajectories of occluded target pairs and recalculates trajectory similarity, effectively associating occluded targets. Although SparseTrack [11] aligns occluded position intervals by performing target set decomposition, this approach can only roughly decompose partial target sets. In this paper, we propose a completely different solution for associating occlusions by using pseudo-gradients to achieve a finer-grained sparse decomposition of the target set.

3. Pseudo-Gradient-Based Multi-Target Tracking Algorithm

3.1. Overall Flowchart of the Algorithm

The positions (x, y, w, h) and confidence scores (s) of the target boxes are obtained from the video frames using a detector. Regardless of whether the boxes have high or low confidence scores, they are classified into three layers—low, medium, and high gradient layers—based on their positional information. Next, association matching is performed. The overall algorithm flowchart is shown in Figure 1.

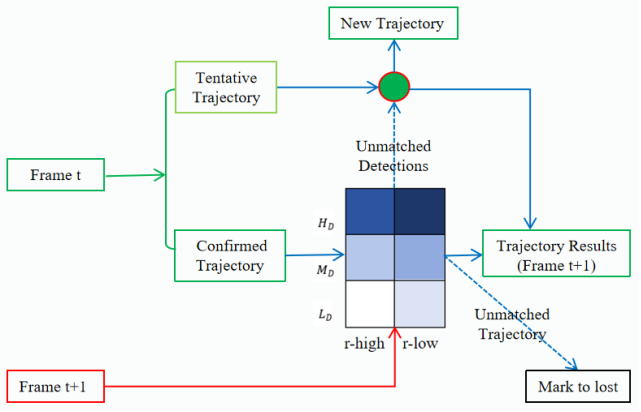

Figure 1. Overall Process Flowchart

For the t+1 frame, detection boxes are obtained using the YOLOX detector. Meanwhile, based on historical trajectories, a Kalman filter predicts the bounding boxes for the t+1 frame using the t frame. The detection boxes and predicted boxes in the t+1 frame is then associated using the Hungarian algorithm. Matched detection boxes are assigned corresponding ID numbers, while unmatched detection boxes are assigned new ID numbers. Unmatched trajectories are temporarily retained for a few frames, typically up to 60 frames, after which they are directly deleted.

3.2. Pseudo-gradient Estimation Method

This method divides targets based on the extent to which they are occluded by other objects. The ratio of the maximum occluded area to the detection box area is defined as the pseudo-gradient value of the target.

$$G_D = S_{IoU-max}/S_D \quad (1)$$

Where G_D is the gradient value of the target, $S_{IoU-max}$ denotes the maximum occluded area of the target bounding box, and S_D represents the area of the target bounding box.

3.3. Pseudo-gradient-based Hierarchical Association

The pseudo-gradient is like contour lines in geography. Based on the gradient hierarchy, targets are grouped and decomposed according to the degree of occlusion. To the best of our knowledge, this is the first method to decompose the target set based on the degree of target clustering. The number of layers can be adjusted based on the actual situation, and in this paper, we agree to divide it into three gradient layers.

An illustration of target set decomposition based on pseudo-gradient is shown in Figure 2. Target 4 and 6 occlude target 8; target 1 occludes target 5; target 3 occludes target 7. As the density value increases, the occlusion of local regions causes the ranking of occluded targets to gradually shift toward the background. The specific steps are as follows: the pseudo-gradient value of each target box is calculated using the pseudo-gradient estimation method from section 2.2, and then the target set is further divided into three layers: low-gradient, medium-gradient, and high-gradient layers, which correspond to the green box, blue box, and red box in the lower-left part of the figure.

In the top left corner of the image is the original detection box of frame t+1, and in the bottom left corner is the scene after pseudo-gradient value layering of frame t+1. The green, blue, and red boxes correspond to the detection boxes in the low, medium, and high gradient layers, respectively.

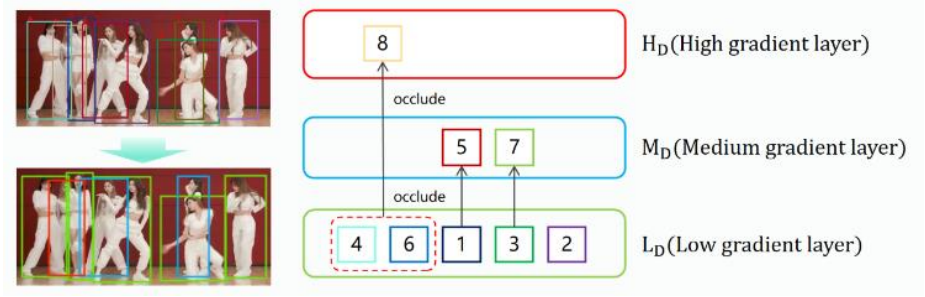

Figure 2. The schematic diagram of target set hierarchical decomposition using pseudo-gradients

In fact, the associative work of low-gradient layers is prioritized, which, to some extent, removes a significant obstruction for the middle and high-gradient layers, indirectly improving the accuracy of association for occluded targets. This can effectively alleviate situations where multiple objects are clustered together, successfully addressing the problem of dense occlusion.

4. Experiments

4.1. Experimental Settings

4.1.1. Datasets

In our experiments, we evaluated our algorithm on two public datasets, MOT17 [17] and MOT20 [18]. MOT17 is a widely-used benchmark dataset containing video sequences captured by static and moving cameras, with motion typically being linear. The dataset is divided into training and

validation sets, with the first half of each video used for training and the second half for validation. MOT20 is more challenging, featuring dense scenes, significant occlusions, and longer video sequences, designed to evaluate the algorithm's performance in complex environments, particularly in handling occlusions and dense objects.

4.1.2. Metrics

We adopt HOTA [19], MOTA [20], and IDF1 [21] as evaluation metrics. The primary metric, high-level tracking accuracy (HOTA), combines several sub-metrics that evaluate the algorithm from different perspectives, balancing the effects of accurate detection, association, and localization into a unified measure, thereby providing a comprehensive assessment of algorithm performance. IDF1 evaluates the association performance of the tracker, while MOTA places more emphasis on detection performance.

4.1.3. Implementation details

To fairly demonstrate the advantages of our GradsSORT algorithm, we keep the detection and ReID components of the baseline HybridSORT-ReID unchanged, meaning we use YOLOX as the detection model. Similarly, for the ReID component, we adopt the model from BoT-SORT. The implementation is carried out using the PyTorch framework, and we employ the Adam optimizeon an NVIDIA GeForce RTX 3080 Ti for training and evaluation.

4.2. Evaluation of Different Benchmark

Our GradsSORT tracker underwent performance evaluation on the MOT17 and MOT20 test sets and was compared with other methods. The results are presented in Table 1 and Table 2. Experimental results demonstrate that GradsSORT performs exceptionally well on both the MOT17 and MOT20 datasets, with consistent improvements in HOTA, IDF1, and MOTA. Notably, GradsSORT exhibits even more significant tracking performance on MOT20 compared to MOT17.

4.2.1. MOT17

We present the performance of GradsSORT on MOT17 in Table 1. Specifically, GradsSORT outperforms the previous state-of-the-art tracker HybridSORT-ReID in all metrics, achieving 0.3 HOTA, 0.8 IDF1, and 0.33 MOTA. The additional computational overhead is negligible. It is worth noting that our method is primarily designed to address the challenges of object clustering and complex motion patterns. However, even when applied to the MOT17 dataset, which represents more general and simpler linear motion pattern scenarios, our method consistently demonstrates enhanced tracking performance.

Table 1. The comparison of GradsSORT with other methods on the MOT17-test set.

Tracker	HOTA↑	IDF1↑	MOTA↑
ByteTrack	63.1	77.3	80.3
StrongSOR	63.5	78.5	78.3
BoT-SORT-ReID	65.0	80.2	80.5
Deep OC-SORT	64.9	80.6	79.4
HybridSORT-ReID	63.2	78.1	78.4
GradsSORT (ours)	63.5	78.9	78.7

4.2.2. MOT20

Our algorithm demonstrates outstanding performance on the MOT20 test set, as shown in Table 2, with negligible additional computational overhead. Specifically, GradsSORT outperforms HybridSORT-ReID in all metrics, achieving 0.5 HOTA, 1.6 IDF1, and 0.6 MOTA, with an excellent HOTA score of 64.8. The results indicate that the method has certain effectiveness, robustness, and generalization in dealing with weak cues in scenarios resembling clustered dense objects and severe occlusions.

Table 2. The comparison of GradsSORT with other methods on the MOT20-test set.

Tracker	HOTA↑	IDF1↑	MOTA↑
ByteTrack	61.3	75.2	77.8
StrongSOR	61.5	75.9	72.2
BoT-SORT-ReID	63.3	77.5	77.8
Deep OC-SORT	63.9	79.2	75.6
HybridSORT-ReID	63.3	78.0	75.5
GradsSORT (ours)	64.8/0.5	79.6/1.6	76.1/0.6

4.3. Ablations

We conducted ablation experiments by integrating the Grads plugin for gradient-based hierarchical association into both HybridSORT and HybridSORT-ReID. The results, as shown in Table 3, indicate that adding the Grads plugin enhances tracking performance for both HybridSORT and HybridSORT-ReID on the MOT17 validation set. This demonstrates that the proposed scheme of gradient-based hierarchical association for target occlusion is effective for subsequent tracking and matching. Detailed experimental results can be found in Table 4 and Table 5.

Table 3. Ablation experiments with the Grads plugin on HybridSORT and HybridSORT-ReID.

Tracker	HOTA↑	IDF1↑	MOTA
HybridSORT	66.943	77.717	75.795
HybridSORT+Grads	67.286	78.442	76.454
HybridSORT-ReID	68.301	80.668	76.96
HybridSORT-ReID+Grads (ours)	68.578	81.246	76.725

Table 4. The results of HybridSORT with the Grads plugin on MOT17 val.

Sequence	HybridSORT	GradsSORT (ours)
MOT17-02	42.663	42.912 (+0.584%)
MOT17-04	79.446	79.906 (+0.579%)
MOT17-05	59.493	60.830 (+2.247%)
MOT17-09	66.097	66.556 (+0.694%)
MOT17-10	54.739	54.157 (-1.063%)
MOT17-11	64.292	64.411 (+0.185%)
MOT17-13	63.402	67.286 (+0.512%)
Overal	66.943	67.286 (+0.512%)

Table 5. The results of HybridSORT-ReID with the Grads plugin on MOT17 val.

Sequence	Base	GradsSORT (ours)
MOT17-02	47.974	48.452 (+0.996%)
MOT17-04	79.557	78.988 (−0.715%)
MOT17-05	59.177	59.103 (−0.125%)
MOT17-09	65.826	65.833 (+0.01%)
MOT17-10	56.493	58.24 (+3.092%)
MOT17-11	64.389	67.973 (+5.566%)
MOT17-13	67.551	68.022 (+0.697%)
Overal	68.301	68.578 (+0.406%)

5. Conclusion

In this paper, we demonstrate that pseudo-gradients enable effective sparse decomposition of target aggregation blocks, offering a novel solution to the long-standing challenge of severe occlusions and clustering. Extensive experiments validate the strong generalization capability of this algorithm across different trackers and scenarios.

Acknowledgment

We thank our fellow lab members for their assistance. This research is partially supported by the National Natural Science Foundation of China (62173078), the Key Research and Development Program of Liaoning Province (2024JH2/102400028), and the Scientific Research Project of Liaoning Provincial Department of Education (YTMS20231210).

References

[1] Vandenhende S, Georgoulis S, Van Gansbeke W, et al. Multi-task learning for dense prediction tasks: A survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 44(7): 3614-3633.

[2] Bewley A, Ge Z, Ott L, et al. Simple online and realtime tracking[C]//2016 IEEE international conference on image processing (ICIP). IEEE, 2016: 3464-3468.

[3] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.

[4] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]//2017 IEEE international conference on image processing (ICIP). IEEE, 2017: 3645-3649.

[5] Aharon N, Orfaig R, Bobrovsky B Z. BoT-SORT: Robust associations multi-pedestrian tracking[J]. arXiv preprint arXiv:2206.14651, 2022.

[6] Chen L, Ai H, Zhuang Z, et al. Real-time multiple people tracking with deeply learned candidate selection and person re-identification[C]//2018 IEEE international conference on multimedia and expo (ICME). IEEE, 2018: 1-6.

[7] Cao J, Pang J, Weng X, et al. Observation-centric sort: Rethinking sort for robust multi-object tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 9686-9696.

[8] Zhang Y, Sun P, Jiang Y, et al. Bytetrack: Multi-object tracking by associating every detection box[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 1-21.

[9] Yang M, Han G, Yan B, et al. Hybrid-sort: Weak cues matter for online multi-object tracking[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(7): 6504-6512.

[10] Du Y, Zhao Z, Song Y, et al. Strongsort: Make deepsort great again[J]. IEEE Transactions on Multimedia, 2023, 25: 8725-8737.

[11] Liu Z, Wang X, Wang C, et al. Sparsetrack: Multi-object tracking by performing scene decomposition based on pseudo-depth[J]. arXiv preprint arXiv:2306.05238, 2023.

[12] Qin Z, Zhou S, Wang L, et al. Motiontrack: Learning robust short-term and long-term motions for multi-object tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 17939-17948.

[13] Zeng F, Dong B, Zhang Y, et al. Motr: End-to-end multiple-object tracking with transformer[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 659-675.

[14] Quach K G, Nguyen P, Duong C N, et al. Depth perspective-aware multiple object tracking[M]//Engineering Applications of AI and Swarm Intelligence. Singapore: Springer Nature Singapore, 2024: 181-205.

[15] Liu Q, Chen D, Chu Q, et al. Online multi-object tracking with unsupervised re-identification learning and occlusion estimation[J]. Neurocomputing, 2022, 483: 333-347.

[16] Liu Y, Zhang X, Zhang B, et al. Multi-camera vehicle tracking based on occlusion-aware and inter-vehicle information[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 3257-3264.

[17] Milan A. MOT16: A benchmark for multi-object tracking[J]. arXiv preprint arXiv:1603.00831, 2016.

[18] Dendorfer P. Mot20: A benchmark for multi object tracking in crowded scenes[J]. arXiv preprint arXiv:2003.09003, 2020.

[19] Luiten J, Osep A, Dendorfer P, et al. Hota: A higher order metric for evaluating multi-object tracking[J]. International journal of computer vision, 2021, 129: 548-578.

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

[21] Ristani E, Solera F, Zou R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]//European conference on computer vision. Cham: Springer International Publishing, 2016: 17-35.