Focus-MOT: Multi-target tracking detection algorithm with fine-grained feature extraction aggregation

Hongyu Jia[1], Wenwu Yang[1], and Lulu Zhang[1]

[1]Dalian Maritime University

June 26, 2023

Abstract

This work proposes a multi-target tracking and detection algorithm Focus-MOT based on feature refinement extraction fusion, t through the designed Field Enhancement Refinement Module and Information Aggregation Module, which effectively reduces the number of target ID switching.Jointly learns the Detector and Embedding model method becomes the mainstream of multi-target tracking and detection due to its fast detection speed, its Re-ID branch needs to use low-dimensional features and high-dimensional features to accommodate both large and small targets, however, its insufficient feature extraction leads to high ID_SW. Therefore this work aims to extract features of different levels for aggregation as a way to reduce the number of ID switching. The experimental results show a 2.7% improvement in MOTA and a 2300 times decrease in ID_SW relative to the results of the FairMOT algorithm on the MOT17 dataset.

Focus-MOT: Multi-target tracking detection algorithm with fine-grained feature extraction aggregation

Jia Hongyu[1], Yang Wenwu[2,], Zhang Lulu[3]

[1] *Dalian Maritime University, No.1 Linghai Road, Dalian, China*

[2] *Dalian Maritime University, No.1 Linghai Road, Dalian, China*

[3] *Dalian Maritime University, No.1 Linghai Road, Dalian, China*

Email: yangwenwu@dlmu.edu.cn.

Abstract This work proposes a multi-target tracking and detection algorithm Focus-MOT based on feature refinement extraction fusion, t through the designed Field Enhancement Refinement Module and Information Aggregation Module, which effectively reduces the number of target ID switching.Jointly learns the Detector and Embedding model method becomes the mainstream of multi-target tracking and detection due to its fast detection speed, its Re-ID branch needs to use low-dimensional features and high-dimensional features to accommodate both large and small targets, however, its insufficient feature extraction leads to high ID_SW. Therefore this work aims to extract features of different levels for aggregation as a way to reduce the number of ID switching. The experimental results show a 2.7% improvement in MOTA and a 2300 times decrease in ID_SW relative to the results of the FairMOT algorithm on the MOT17 dataset.

Introduction: Deep learning based multi-target tracking and detection methods can be generally classified into Tracking-By-Detection (referred to as TBD paradigm) and Jointly learns the Detector and Embedding model (referred to as JDE paradigm).The TBD paradigm is represented by Faster R-CNN as the detector of Sort, DeepSort algorithm, MOTDT algorithm, etc. [1-3]. Since the TBD paradigm treats feature vector acquisition and target detection as two separate models and features are not shared, both parts require separate computation time, and the total time is the sum of both, resulting in a lot of time wastage. In contrast, the JDE paradigm uses a single network to fuse target detection and embedding learning, extracts

Re-ID features while target detection, and reduces repeated computational inference by sharing features, thus improving the time efficiency of the model while maintaining the same accuracy as the TBD paradigm. For example, Fair-MOT and TADAM algorithms that improve the JDE paradigm [4-6].

The JDE paradigm relies on the feature extraction of the backbone network for recognition tracking, and the degree of its extraction seriously affects the detection tracking accuracy.

Focus-MOT improves the feature extraction and fusion strategy under the single network multitasking model, and adopts the JDE paradigm to design the Field Enhancement Refinement Module and Information Aggregation Module, aiming at extracting features of different levels for aggregation through the backbone network. in order to reduce the number of ID exchanges and pursue a balanced progress between detection speed and accuracy.

Focus-MOT's network structure

Figure 1 shows the proposed network structure of Focus-MOT. Focus-MOT uses Res2Net-50 [7] as a backbone network to increase the perceptual field of the network layers by constructing hierarchical residual class connections within a single residual block. The input image is normalized to 3*608*1088, and five-layer feature maps with sizes of 64*304*544, 256*152*272, 512*76*136, 1024*38*78, and 2048*19*39 are obtained through the backbone network, and the obtained five-layer feature maps are enhanced by the designed Field Enhancement Refinement Module to expand the perceptual field of the high-dimensional features, while completing the refinement of the features from the spatial dimension and the channel dimension, and then through the Information Aggregation Module, the bottom-up feature fusion from the high level to the low level, completing the information interaction between the high level semantic information and the low level detail information.

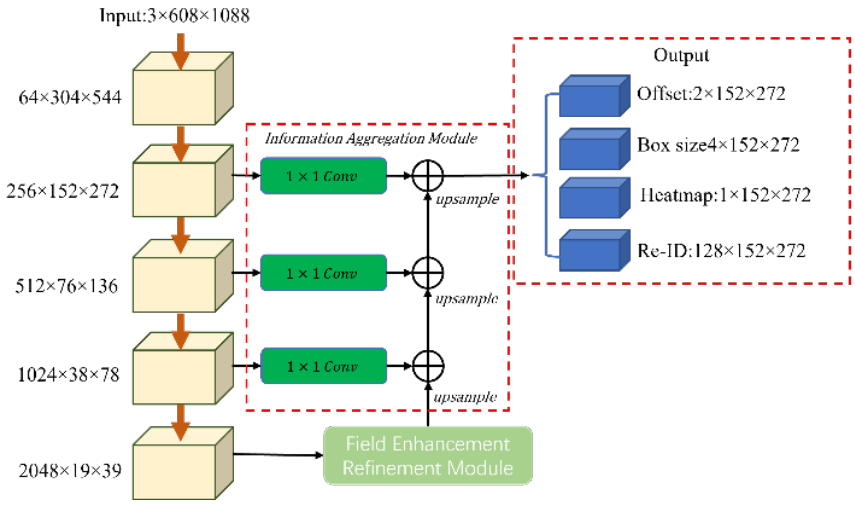

Fig 1 *The network structure of Focus-MOT.*

Figure 2 shows the network of Field Enhancement Refinement Module. The Field Enhancement Refinement Module first goes through five parallel modules: adaptive pooling, 3×3 convolution with hole rates of 6, 8, and 12, and 1×1 convolution, respectively, and then stitches them together so that multi-scale information can be captured while expanding the feature map sensory field. And after completing such an operation, we design two parallel modules to capture rich contextual relationships to better achieve compact feature representation within the class.

First is the branch above, A is the feature map of the input parallel module with size C × H × W. First, A is subjected to a convolution operation to obtain new feature maps B, C (B = C, size C × H × W), and

then BC are reshape to the size of C × N, where N = H × W. B is transposed and multiplied with C, and the obtained result is then subjected to a softmax operation to obtain the feature map S of size The sum of each row in S is 1. s_ji can be interpreted as the weight of pixel at position j to pixel at position i, i.e., the weight of all pixels j to a fixed pixel i is 1.

$$s_{\mathrm{ji}} = \frac{\exp\left(B_i \cdot C_j\right)}{\sum_{i=1}^{N} exp\left(B_i \cdot C_j\right)}$$

Meanwhile, A is subjected to another convolution operation to obtain the feature map D (of size C × H × W), with the same reshape of size C × N. Multiply it with the transpose of S to get the result map of size C × N, and then reshape it back to size C × H × W, multiplying it by a coefficient γ. Finally, add it to A to get the final feature map result E incorporating location information. where γ is a weight parameter to be learned, with an initial value of 0.

$$E_j = \gamma \sum_{i=1}^{N} \left(s_{\mathrm{ji}} D_i\right) + A_j$$

Such a branch is able to build rich contextual relationships on local features, encoding broader contextual information into local features and thus enhancing their representational power.Then comes the next branch, where we argue that the channel graph of each high-level feature can be regarded as a class-specific response, and by mining the interdependencies between channel graphs, the interdependent feature graphs can be highlighted and the semantics-specific feature representation can be improved. Therefore, this branch of the paper aims at building a channel attention module to explicitly model the dependencies between channels.Similar to the previous branch, except that instead of performing a convolution operation on the feature map A, the operation is performed directly on A. Similarly, A is reshaped to a size of C × N, denoted as B, and then B is multiplied with its own transpose and then subjected to a softmax operation to obtain a feature map X of size C × C.

$$x_{\mathrm{ji}} = \frac{\exp\left(A_i \cdot A_j\right)}{\sum_{i=1}^{C} exp\left(A_i \cdot A_j\right)}$$

The transpose of X is multiplied by B and then reshape back to the size of C×H×W, multiplied by a factor β, denoted as D. Adding A to D gives the final feature map E with fused channel information. β also has an initial value of 0.

$$E_j = \beta \sum_{i=1}^{C} \left(x_{\mathrm{ji}} A_i\right) + A_j$$

After the input features are processed by these two parallel branches, the two feature maps are added element by element to complete the fusion of the two feature maps, and the 1×1 convolution is used to reduce the dimensionality, so that the whole Spatial Fusion module can enhance the fusion of the low-level features.

After completing this series of operations, we up-sample the high-level features one by one by the designed Information Aggregation Module, and each up-sampling will be added element by element with the feature maps of the same resolution output by Res2net-50, and then the final four feature maps will be output.

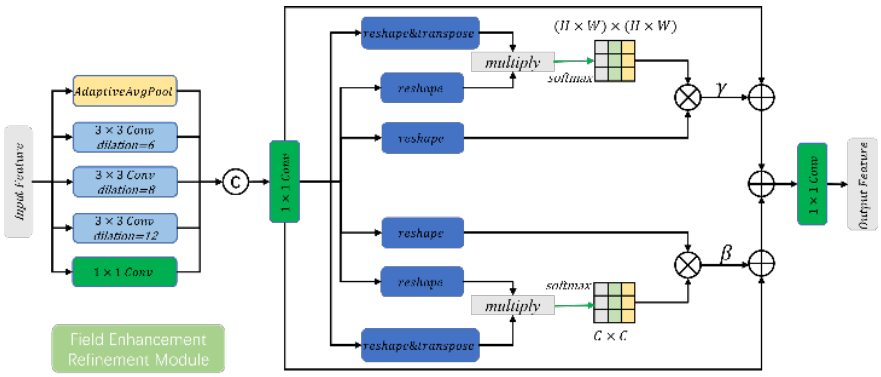

Fig 2 *Field Enhancement Refinement.*

Focus-MOT has four loss functions, which correspond to the loss hm_loss for heatmap, wh_loss for boxsize, off_loss for offset, and id_loss for Re-ID. for hm_loss, we use the MSE loss function to calculate; for wh_loss, we use the L1 loss function to calculate; for off_loss, we use the multivariate cross-entropy function to calculate. use the L1 loss loss function to calculate; for off_loss, we use the L1 loss function to calculate, and for Re-ID loss we use the multivariate cross entropy function to calculate.

Then the total loss is:

$$\text{loss} = \text{hm\_loss} + \text{wh\_loss} + \text{off\_loss} + 0.1 \times \text{id\_loss}$$

Focus-MOT uses the training set provided by the six datasets MOT17, Caltech, Citypersons, Cuhksysu, PRW, and ETH, and the test set of MOT15 and MOT17 is used for testing [8-12].

Experiment

Normalize all the input images to 608×1088, with an initial learning rate of 0.0001, and a batchszie of 4. Using the Adam optimizer, the learning rate decays to one percent of the initial learning rate after 100 epochs.

Figure 3 shows a visual display of the experimental results of Focus-MOT on MOT17 versus MOT15 datasets.

Fig 3 *MOT17 and MOT15 data set results show.*

We selected four types of evaluation metrics, MOTA, IDF, ID_SW, and FPS, to evaluate Focus-MOT and compare it with the methods in recent years, and the metric values are all from the published values of their papers, which have considerable objectivity. By comparing the results, the total number of ID_SW of Focus-MOT on MOT15 dataset is 356 times. On the MOT17 dataset, the total number of IDs is 568, both of which are the methods with the least ID_SW.

Table 1. Comparison of test results for MOT15 dataset.

	MOTA	IDF1	FPS	ID_SW
MDP_SubCNN	47.5	55.7	628	<1.7
AP_HWDPL	53.0	52.2	708	6.7
Rar15	56.5	61.3	428	<3.4
FairMOT*	**59.0**	62.2	582	**25.9**
SFP-JDE*	48.1	60.9	626	8.7
Focus-MOT	52.7	**67.4**	**356**	6.73

* The marked * is the algorithm that takes the JED paradigm.

Table 2. Comparison of test results for MOT17 dataset.

	MOTA	IDF1	FPS	ID_SW
SST	52.4	49.5	8431	<3.9
TubeTK*	63.0	58.6	4137	3.0
CenterTrack*	67.8	64.7	2583	17.5
FairMOT*	67.5	69.8	2868	**25.9**
TransCenter*	58.5	-	4659	-
Focus-MOT	**70.2**	**76.0**	**568**	7.75

* The marked * is the algorithm that takes the JED paradigm.

Ablation experiments

A detailed ablation experiment was conducted to verify the role and magnitude of each module of Focus-MOT. The hyperparameters and environment of the ablation experiments are consistent with those described previously, and the test analysis is performed on the MOT15 dataset with the design shown below:A: Parallel module without expanded feature map perceptual fields in Field Enhancement Refinement Module;B: Field Enhancement Refinement Module without the dual attention module in the Field Enhancement Refinement Module;C: Without Information Aggregation Module, the features in Field Enhancement Refinement Module are upsampled and summed with the fourth layer features and output.

Table 3. The results of the ablation experiments .

A	B	C	MOTA↑	IDF1	IDs	FPS
×	[?]	[?]	47.3	57.7	543	7.5?¿?
	×	[?]	46.2	54.8	572	**9.2**?¿?
	[?]	×	48.7	60.6	445	8.2?¿?
	[?]	[?]	**52.7**	**67.4**	**356**	6.73

From the results of the ablation experiments, it can be seen that increasing the perceptual field and refining the filtering of features in spatial and channel dimensions can bring different gains to the accuracy of multi-target tracking from different perspectives, while the bottom-up fusion in the information aggregation module can fully fuse the low-level information with the high-level information, which is also the place where the model gains the most.

Conclusion: Focus-MOT takes the extraction and fusion of features at different scales as the main direction, which retains more effective feature information and effectively reduces the number of ID switching during model tracking. Field Enhancement Refinement Module and Information Aggregation Module are proposed to improve the network's ability to extract key features of the target and enhance the model's effect of extracting features under different sensory fields. Moreover, it can effectively improve the tracking ability when the target scale is small and the targets overlap, and effectively improve the accuracy of model detection and tracking. The experimental results show that the method has a strong comprehensive performance by effectively reducing the number of ID switching at a higher MOTA.

References

1. Bewley A, Ge Z, Ott L, et al. Simple online and realtime tracking[C]. 2016 IEEE International Conference on Image Processing, Phoenix, Arizona, USA, 2016: 3464-3468

2. Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2015:28.

3. Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]. 2017 IEEE International Conference on Image Processing, Beijing, China, IEEE, 2017: 3645-3649

4. Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada, USA, 2016: 779-788

5. ZHANG Y F,WANG C Y,WANG X G,et al. FairMOT:on the fairness of detection and re-identification in mul–tiple object tracking[J]. International Journal of Computer Vision,2021,129(11):3069-3087

6. GUO Song,WANG Jingya,WANG Xinchao,et al. Online multiple object tracking with cross-task synergy[C].2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Nashville,2021: 8132-8141

7. Shang-Hua Gao,Ming-Ming Cheng,Kai Zhao,Xin-Yu Zhang,Ming-Hsuan Yang,Philip Torr.Res2Net:A New Multi-scale Backbone Architecture.arXiv:1904.01169:1-7

8. Dollár P, Wojek C, Schiele B, et al. Pedestrian detection: A benchmark[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 304-311

9. Zhang S, Benenson R, Schiele B. Citypersons: A diverse dataset for pedestrian detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3213-3221

10. Zhang S, Benenson R, Schiele B. Citypersons: A diverse dataset for pedestrian detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3213-3221

11. Xiao T, Li S, Wang B, et al. Joint detection and identification feature learning for person search[C]//Proceedings of the .IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3415-3424

12. Zheng L, Zhang H, Sun S, et al. Person re-identification in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1367- 1376