SCIENCE CHINA
Information Sciences

• LETTER •

August 2024, Vol. 67, Iss. 8, 189102:1–189102:2
https://doi.org/10.1007/s11432-022-4097-y

Wavelet-domain feature decoupling for weakly supervised multi-object tracking

Yu-Lei LI, Yan YAN, Yang LU* & Hanzi WANG

Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, China

Received 16 November 2022/Revised 10 June 2023/Accepted 17 April 2024/Published online 24 July 2024

Multi-object tracking (MOT) aims to discover interesting foreground targets and meanwhile generate robust trajectories of these targets with different data association algorithms based on identity embedding features (Re-ID features) [1] or motion cues [2]. During the past few decades, some advanced MOT methods have achieved considerable progress due to their wide range of applications in modern autonomous driving and intelligent video analysis.

Despite the continuous progression of supervised tracking, the MOT task is still challenging due to the adverse influence of crowded scenes including many heavily or completely occluded targets (for example, occluded persons) and the limitation of manual identity labels for supervised training. The crowded scenes most likely bring unreliable detection predictions and noisy embedding features or motion predictions in the MOT task, and they adversely affect the effective capability of the competing methods to generate robust trajectories. To address the above problems, some tracking-by-detection methods adopt expensive detectors to obtain accurate detection predictions and suppress the adverse influence of heavily occluded targets. Moreover, some joint-detection-and-tracking methods aim to alleviate the limitation of noisy embedding features (or motion predictions). They simultaneously perform detection predicting and identity embedding (or motion predicting) processes to effectively discover and track heavily occluded targets and suppress identity switches (IDS) in crowded scenes. Some Transformer-based methods focus on extracting discriminative tracking queries and performing both detection and data association sub-tasks in their unified detection and tracking networks.

These supervised MOT methods heavily depend on manual identity labels to extract discriminative intermediate features for identifying multiple targets and associating them into trajectories. In contrast, some weakly supervised methods [3,4] perform the MOT task by generating pseudo identity labels with detection predictions. However, the pseudo identity labels contain many false or missing labels of heavily occluded targets and adversely affect their network optimizations, resulting in noisy intermediate features in their networks. To alleviate the above challenging problem, we propose a wavelet-domain feature-decoupling Transformer-based tracking network (FDMOT) for the weakly supervised MOT task. The proposed FDMOT effectively extracts noise-decoupled intermediate features based on the different wavelet coefficients of noise and target features in the wavelet domain. Moreover, it further improves the embedding features obtained from the identity embedding branch to well-refined embedding features in the data association branch.

Proposed method. We propose a feature-decoupling Transformer-based network to decouple noisy features and extract noise-decoupled intermediate features for detection, identity embedding and data association. The proposed feature-decoupling Transformer module (FDT) mainly consists of a feature decoupler and two multi-layer perceptrons (MLP). In particular, the feature decoupler component has some variants in terms of convolution layers, i.e., none, spatial-domain, Fourier-domain, and wavelet-domain convolution layers. Moreover, the MLP component includes the following variants: FFN, RepMLP, ResMLP, and MLP-Mixer. See the appendix for more details. Our FDMOT takes the noisy intermediate features $\boldsymbol{x}_t^n$ as input, and outputs the noise-decoupled intermediate features $\boldsymbol{x}_t^d$, which can be written as follows:

$$\boldsymbol{x}_t^w = \mathrm{Conv}[\mathrm{WT}(\boldsymbol{x}_t^n)], \tag{1}$$

$$\boldsymbol{x}_t^m = \mathrm{softmax}\left[\frac{\boldsymbol{x}_t^w(\boldsymbol{x}_t^w)^{\mathrm{T}}}{\sqrt{d_{\boldsymbol{x}_t^w}}}\right]\boldsymbol{x}_t^w, \tag{2}$$

$$\boldsymbol{x}_t^i = \mathrm{IWT}[\mathrm{Conv}(\boldsymbol{x}_t^m)] + \mathrm{MLP}_M(\boldsymbol{x}_t^n), \tag{3}$$

$$\boldsymbol{x}_t^d = \mathrm{MLP}_F[\boldsymbol{x}_t^i] + \boldsymbol{x}_t^i, \tag{4}$$

where $\boldsymbol{x}_t^w$, $\boldsymbol{x}_t^m$ and $\boldsymbol{x}_t^i$ are the intermediate features to calculate $\boldsymbol{x}_t^d$. $\mathrm{WT}(\cdot)$, $\mathrm{IWT}(\cdot)$ and $\mathrm{Conv}(\cdot)$ respectively denote the wavelet transform, the inverse-wavelet transform and the feature extraction with convolution layers. $\mathrm{MLP}_M(\cdot)$ and $\mathrm{MLP}_F(\cdot)$ denote the MLPMixer and FFN components. $d_{\boldsymbol{x}_t^w}$ and $(\boldsymbol{x}_w)^{\mathrm{T}}$ are respectively the dimension and the matrix transpose of $\boldsymbol{x}_t^w$. $\mathrm{softmax}(\cdot)$ is the softmax function. Then, FDMOT extracts the reliable detection predictions $\boldsymbol{d}_t$ and noise-decoupled embedding features $\boldsymbol{e}_t$ respectively by the detection and identity embedding heads of the two branches. We further perform feature decoupling on the noise-decoupled embedding features to extract the

* Corresponding author (email: luyang@xmu.edu.cn)

Table 1 Comparisons of FDMOT with some state-of-the-art methods on the MOT20 test set[a)]

Method	MOTA↑	IDF1↑	HOTA↑	MT↑	ML↓	IDS↓	FPS↑
FairMOT [1]	61.8	67.3	54.6	68.8	7.6	5243	13.2
ByteTrack [2]	77.8	75.2	61.3	69.2	9.5	1223	17.5
MeMOT [5]	63.7	66.1	54.1	57.5	14.3	1938	–
UTrack† [4]	68.5	69.4	–	57.9	12.2	2147	12.4
UEANet† [3]	**73.0**	75.6	58.6	55.0	13.9	**1423**	**12.8**
FDMOT (ours)†	72.8	**76.0**	**59.5**	**59.3**	**12.0**	1776	12.1

a) The symbol '†' denotes that the method is based on weakly supervised training. For each metric, we mark the best result in bold.

well-refined embedding features $\boldsymbol{e}_t^{\mathrm{wr}}$ with the above feature-decoupling Transformer module, which can be written as

$$\boldsymbol{e}_t^{\mathrm{wr}} = \mathrm{FDT}(\boldsymbol{e}_t), \tag{5}$$

where $\mathrm{FDT}(\cdot)$ denotes the feature-decoupling Transformer module with the embedding features $\boldsymbol{e}_t$ as input in the data association branch.

Finally, the proposed method generates robust trajectories with the well-refined embedding features $\boldsymbol{e}_t^{\mathrm{wr}}$ and $\boldsymbol{e}_{t-1}^{\mathrm{wr}}$ in the data association process by

$$\boldsymbol{a}_{t,t-1} = \mathrm{Branch}_A\left(\boldsymbol{e}_t^{\mathrm{wr}}, \boldsymbol{e}_{t-1}^{\mathrm{wr}}\right), \tag{6}$$

where $\boldsymbol{a}_{t,t-1}$ denotes the association matrix that is obtained by the embedding-based association algorithm [1] in the data association head $\mathrm{Branch}_A(\cdot,\cdot)$. Based on the proposed feature-decoupling Transformer module, our method can suppress noise features and enhance target features of occluded targets according to their different feature coefficients in the wavelet domain. Moreover, we improve noise-decoupled embedding features to well-refined embedding features to accurately identify heavily occluded targets in the data association branch.

For the weakly supervised training of FDMOT, we propose a two-step training procedure and a joint loss. We first adopt the Kalman filter tracker to generate the pseudo identity labels, and then use these pseudo labels to optimize the tracking network with the joint loss L_t by

$$L_t = L_{\mathrm{det}}(\hat{\boldsymbol{d}}_t, \boldsymbol{d}_t) + L_{\mathrm{ie}}(\hat{\boldsymbol{p}}_t, \boldsymbol{p}_t) + L_{\mathrm{ea}}(\hat{\boldsymbol{p}}_t, \boldsymbol{p}_t^{\mathrm{wr}}), \tag{7}$$

where $L_{\mathrm{det}}(\cdot,\cdot)$ denotes the detection loss [1]. $L_{\mathrm{ie}}(\cdot,\cdot)$ and $L_{\mathrm{ea}}(\cdot,\cdot)$ are the k-class cross-entropy loss. k is the identity number of trajectories. $\hat{\boldsymbol{d}}_t$ and $\boldsymbol{d}_t$ denote the ground-truth detection labels and the detection predictions. $\boldsymbol{p}_t$ and $\boldsymbol{p}_t^{\mathrm{wr}}$ are the predicted identities of the noise-decoupled and well-refined embedding features.

Experiments. When dealing with the crowded scenes of the MOT20 dataset, our FDMOT obtains comparable results on the evaluation metrics of MOTA, HOTA and ML against the previous supervised MOT methods in Table 1. In particular, FDMOT achieves a higher IDF1 (i.e., 76.0) than the tracking-by-detection and joint-detection-and-tracking methods [1, 2]. And it also outperforms the Transformer-based method [5] by +9.1 MOTA, +9.9 IDF1 and +5.4 HOTA. Moreover, it surpasses the other weakly supervised MOT methods [3, 4] in terms of IDF1, HOTA and MT. The comparison results show that FDMOT can effectively handle heavily occluded targets in crowded scenes by improving the noise-decoupled embedding features into the well-refined ones for accurately identifying and tracking targets. With the cooperation of the three feature-decoupling Transformer-based branches, FDMOT effectively decouples noisy features to generate the reliable detection predictions and the well-refined embedding features for robust weakly supervised tracking. Please refer to the Supporting information for more detailed results.

Conclusion. We present a wavelet-domain feature-decoupling Transformer-based tracking network for the weakly supervised MOT task (FDMOT). Our FDMOT has two improvements over the previous weakly supervised methods. First, FDMOT decouples noisy intermediate features caused by noisy pseudo identity labels in the wavelet domain, extracting discriminative features for accurately detecting and identifying multiple targets. Second, FDMOT further improves the noise-decoupled embedding features into the well-refined ones with the cooperation of the three feature-decoupling Transformer-based branches, which can accurately identify and track heavily occluded targets in crowded scenes. Experimental results show the superiority of FDMOT compared with several state-of-the-art supervised and weakly supervised MOT methods.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. U21A20514, 62002302) and FuXiaQuan National Independent Innovation Demonstration Zone Collaborative Innovation Platform Project (Grant No. 3502ZCQXT2022008).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Zhang Y, Wang C, Wang X, et al. FairMOT: on the fairness of detection and re-identification in multiple object tracking. Int J Comput Vis, 2021, 129: 3069–3087
2 Zhang Y, Sun P, Jiang Y, et al. Bytetrack: multi-object tracking by associating every detection box. In: Proceedings of the European Conference on Computer Vision, Tel Aviv, 2022. 1–18
3 Li Y L. Unsupervised embedding and association network for multi-object tracking. In: Proceedings of the International Joint Conference on Artificial Intelligence, Vienna, 2022. 1123–1129
4 Liu Q, Chen D, Chu Q, et al. Online multi-object tracking with unsupervised re-identification learning and occlusion estimation. Neurocomputing, 2022, 483: 333–347
5 Cai J, Xu M, Li W, et al. MeMOT: multi-object tracking with memory. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, 2022. 8090–8100