SCIENCE CHINA
Information Sciences

• **Supplementary File** •

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

Yu-Lei LI[1], Yan YAN[1], Yang LU[1] & Hanzi WANG[1*]

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

Appendix A Related Work

Appendix A.1 Supervised MOT Methods

Tracking-by-detection methods. Most tracking-by-detection methods [10, 24, 38] adopt reliable detection predictions (obtained by high-cost detectors) and the Kalman filter or the bounding box-IoU-based data association algorithm to perform the MOT task. Moreover, some MOT methods [5,20,25,33,36] adopt Transformers to generate discriminative tracking queries for performing detection and data association in unified networks. These tracking-by-detection methods focus on discovering heavily occluded targets, which highly influence the upper bound of supervised tracking. However, their detection branches require high computational costs and training data with manual annotations.

Joint-detection-and-tracking methods. Recently, the joint-detection-and-tracking methods [13, 17, 26, 30, 32, 34, 39] simultaneously extract discriminative embedding features and detection predictions in unified networks. For example, FairMOT [39] addresses the challenging issues of the anchors of bounding boxes and the feature sharing by treating detection and re-ID (identity embedding) branches equally. DeepMOT [34] adopts a deep network to calculate the association matrix in an end-to-end way, which discards the dependence on hand-crafted data association algorithms and further improves the performance. The above joint-detection-and-tracking methods can obtain competitive trajectories with a low computational cost compared with those tracking-by-detection methods. However, the training data with manual detection and identity annotations are necessary for the optimizations of the detection, identity embedding and data association branches.

Transformer-based methods. Recently, Transformers are adopted in some MOT methods [5, 20, 25, 33, 36], which focus on generating discriminative detection and tracking queries and then performing the detection and data association processes in unified networks. For example, Trackformer [20], Transtrack [25] and MOTR [36] adopt target features from the previous frame to create tracking queries for effectively detecting and tracking targets in the current frame. These Transformer-based methods perform MOT as a variant of detection by discovering the current-frame targets with the past-frame tracking queries, which are severely affected by the high-cost detectors.

As a result, those supervised MOT methods effectively exploit the manual labels, and they generate reliable detection predictions and robust trajectories based on discriminative embedding features for occluded targets. However, the detection and identity labels bring a high economic cost in the manual annotations of trajectories.

Appendix A.2 Weakly Supervised MOT Methods

Recently, some MOT methods [1, 14–16, 18] generate identity embedding features based on pseudo identity labels or self-supervised learning to avoid the high-cost manual identity labels. For example, MCM [14] generates the pseudo identity labels with detection predictions and the bounding box-IoU algorithm and adopts these pseudo labels to train the identity embedding network for extracting embedding features and performing data association in a weakly supervised way. Moreover, UTracK [18] optimizes the identity embedding model without any manual identity labels and performs the MOT task based on its self-supervised matching loss. Despite the manual detection labels for the detection branches, the above weakly supervised methods adopt pseudo identity labels to extract identity embedding features for identifying and tracking multiple targets. Due to the adverse optimization with noisy pseudo labels, previous weakly supervised MOT methods [1,14,15,18] suffer from noisy intermediate features in their networks, resulting in noisy embedding features and unsatisfactory trajectories for the weakly supervised MOT task.

Unlike those supervised and weakly supervised MOT methods, we propose an effective feature-decoupling Transformer-based network for the weakly supervised MOT task (named FDMOT). Our FDMOT decouples noise and signal intermediate features according to their different wavelet coefficients and extracts noise-decoupled features with a feature-decoupling Transformer module in the wavelet domain. Moreover, it further improves identity embedding features to well-refined embedding features for accurately identifying and tracking heavily occluded targets in the feature-decoupling Transformer-based data association branch.

Appendix B Proposed Method

Appendix B.1 Motivation of FDMOT

Since the pseudo identity labels are generated with the detection predictions, the pseudo labels may contain many false or missing trajectories of heavily or completely occluded targets. The previous weakly supervised MOT methods [14,15,18] optimized by these noisy pseudo identity labels extract noisy intermediate features and identity embedding features for occluded targets, which are

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

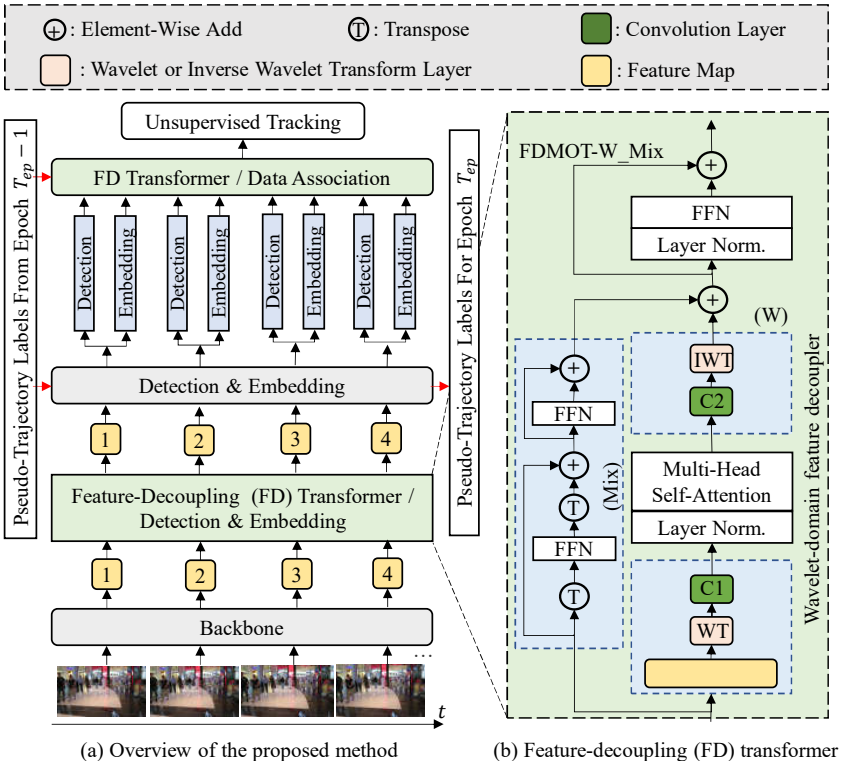

Figure A1 Overview of the proposed FDMOT. (a) FDMOT consisting of the backbone and the detection, identity embedding and data association branches integrated with the feature-decoupling Transformer module. (b) The architecture of the feature-decoupling Transformer module. The red line denotes that FDMOT generates the pseudo identity labels with the detection predictions and the Kalman filter tracker, and then adopts them to optimize the identity embedding branch and the data association branch.

not sufficiently robust to detect and track heavily occluded targets for the weakly supervised MOT task. In contrast, 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. Based on the proposed feature-decoupling Transformer module, our method can suppress noise features and enhances signal features of occluded targets according to their different feature coefficients in the wavelet domain. It extracts reliable detection predictions and noise-decoupled embedding features in the detection and identity embedding branches. Moreover, we improve noise-decoupled embedding features to well-refined embedding features to accurately identify heavily occluded targets in the data association branch. The number and function of the Transformer module used in the previous Transformer-based methods and our FDMOT are different.

Appendix B.2 Weakly Supervised Tracking via Feature Decoupling

Feature-decoupling detection and embedding. As shown in Figure A1(b), the proposed feature-decoupling Transformer module (i.e., FDT) mainly consists of a feature decoupler and two multi-layer perceptrons (MLP) components. In particular, the feature decoupler component has some variants in terms of convolution layers, i.e., none (N), spatial-domain (S), Fourier-domain (F), and wavelet-domain (W) convolution layers. Moreover, the MLP component includes the following variants: FFN [29], RepMLP (Rep) [8], ResMLP (Res) [28], and MLPMixer (Mix) [27]. Specifically, FDMOT (using FDT-W Mix+FDT-W Mix) takes the noisy intermediate features $\mathbf{x}_t^n$ as input, and outputs the noise-decoupled intermediate features $\mathbf{x}_t^d$, which can be written as follows:

$$\mathbf{x}_t^w = Conv[WT(\mathbf{x}_t^n)], \tag{B1}$$

$$\mathbf{x}_t^m = softmax[\frac{\mathbf{x}_t^w(\mathbf{x}_t^w)^T}{\sqrt{d_{\mathbf{x}_t^w}}}]\mathbf{x}_t^w, \tag{B2}$$

$$\mathbf{x}_t^i = IWT[Conv(\mathbf{x}_t^m)] + MLP_M(\mathbf{x}_t^n), \tag{B3}$$

$$\mathbf{x}_t^d = MLP_F[\mathbf{x}_t^i] + \mathbf{x}_t^i, \tag{B4}$$

where $\mathbf{x}_t^w$, $\mathbf{x}_t^m$ and $\mathbf{x}_t^i$ are the intermediate features to calculate $\mathbf{x}_t^d$. WT, IWT and $Conv$ respectively denote the wavelet transform, the inverse-wavelet transform and the feature extraction with convolution layers. MLP_M and MLP_F denote the MLPMixer and FFN components. $d_{\mathbf{x}_t^w}$ and $(\mathbf{x}_w)^T$ are respectively the dimension and the matrix transpose of $\mathbf{x}_t^w$. $softmax(\cdot)$ is the softmax function. Then, FDMOT extracts the reliable detection predictions $\mathbf{d}_t$ and noise-decoupled embedding features $\mathbf{e}_t$ respectively by

$$\mathbf{d}_t = Branch_D(\mathbf{x}_t^d), \tag{B5}$$

$$\mathbf{e}_t = Branch_E(\mathbf{x}_t^d), \tag{B6}$$

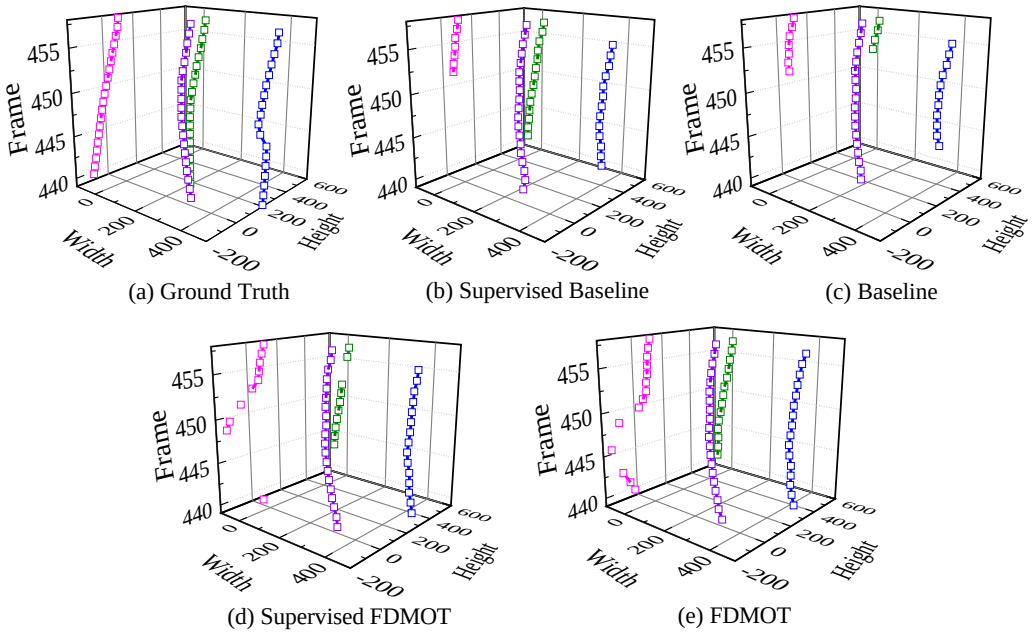

Figure A2 Qualitative comparisons of trajectories. The predicted trajectories are obtained by the ground truth (a), the supervised baseline [39] (b), the weakly supervised baseline (c), the supervised FDMOT (d) and the weakly supervised FDMOT (e), when using frame 439 to frame 458 in the No. 0005 video of MOT17. The height, width and frame axes respectively represent two-dimensional space coordinates and one-dimensional time coordinates of objects in a trajectory.

where $Branch_D$ and $Branch_E$ represent the detection and identity embedding heads of the two branches.

As a result, we first transform the noisy intermediate features from the spatial domain to the wavelet domain, which separates the noisy features into noise features and signal features based on their different wavelet coefficients. Then, the feature-decoupling Transformer module learns to suppress noise features and enhance signal features for the feature-decoupling detection and identity embedding processes.

Data association with well-refined embedding features. We further perform feature decoupling on the noise-decoupled embedding features to extract the well-refined embedding features $\mathbf{e}_t^{wr}$ with the feature-decoupling Transformer module, which can be written as:

$$\mathbf{e}_t^{wr} = MLP_F[\mathbf{e}_t^i] + \mathbf{e}_t^i, \tag{B7}$$

where the intermediate features $\mathbf{e}_t^i$ are defined as follows:

$$\mathbf{e}_t^w = Conv[WT(\mathbf{e}_t)], \tag{B8}$$

$$\mathbf{e}_t^m = softmax(\frac{\mathbf{e}_t^w(\mathbf{e}_t^w)^T}{\sqrt{d_{\mathbf{e}_t^w}}})\mathbf{e}_t^w, \tag{B9}$$

$$\mathbf{e}_t^i = IWT[Conv(\mathbf{e}_t^m)] + MLP_M(\mathbf{e}_t), \tag{B10}$$

and $\mathbf{e}_t^w$ and $\mathbf{e}_t^m$ are obtained by the feature-decoupling Transformer module for calculating $\mathbf{e}_t^i$. $d_{\mathbf{e}_t^w}$ and $(\mathbf{e}_t^w)^T$ respectively denote the dimension and the matrix transpose of $\mathbf{e}_t^w$.

Finally, the proposed method generates robust trajectories with the well-refined embedding features $\mathbf{e}_t^{wr}$ and $\mathbf{e}_{t-1}^{wr}$ in the data association process, which can be mathematically expressed as

$$\mathbf{a}_{t,t-1} = Branch_A\left(\mathbf{e}_t^{wr}, \mathbf{e}_{t-1}^{wr}\right), \tag{B11}$$

where $\mathbf{a}_{t,t-1}$ denotes the association matrix that is obtained by the embedding-based association algorithm [30, 39] in the data association head $Branch_A$.

Differences between FDMOT and other weakly supervised methods. (1) Wavelet domain: Other weakly supervised methods [14–16, 18] learn the intermediate features in the spatial domain. FDMOT performs feature decoupling on the intermediate features in the wavelet domain, where noise and signal intermediate features can be effectively separated according to their different wavelet coefficients in a common feature space. (2) Noise-decoupled features: Other weakly supervised methods use the noisy intermediate features for identity embedding, and output noisy embedding features. FDMOT adopts the feature-decoupling Transformer module to extract noise-decoupled intermediate features, and it generates noise-decoupled embedding features for occluded targets. (3) Well-refined embedding features: Compared with the noisy embedding features in [14, 15, 18] and the enhanced embedding features in [16], the proposed FDMOT further refines the embedding features to achieve the well-refined embedding features for accurately identifying and tracking heavily occluded targets.

Appendix B.3 Training Procedure

To perform weakly supervised training of FDMOT, we propose a two-step training procedure as follows: (a) Before the epoch T, we firstly generate detection predictions with the detection branch equipped with the network parameters of the past epoch $T-1$; We

Table B1 Ablation study for the influence of the pseudo identity label. * denotes the method is trained with manual identity labels.

FDMOT-Det.&Emb.	FDMOT-Asso.	MOTA↑	IDF1↑	IDS↓
None*	None*	69.1	72.9	299
None	None	68.4	69.9	751
FDT-W_Mix*	FDT-W_Mix*	**71.0**	75.0	**271**
FDT-W_Mix	FDT-W_Mix	70.9	**75.2**	282

Table B2 Ablation study for the influence of the feature-decoupling Transformer-based detection and identity embedding branches.

FDMOT-Det.&Emb.	FDMOT-Asso.	MOTA↑	IDF1↑	IDS↓
FDT-N_Mix	None	69.6	72.5	271
FDT-S_Mix	None	69.7	73.1	304
FDT-F_Mix	None	70.1	73.2	335
FDT-W_FFN	None	**70.2**	73.6	334
FDT-W_Rep	None	69.7	73.5	315
FDT-W_Res	None	70.1	73.4	272
FDT-W_Mix	None	**70.2**	**74.2**	**270**

adopt the Kalman filter tracker to generate the pseudo identity labels (i.e., $\hat{\mathbf{p}}_{T_{ep}}$) with the detection predictions from the previous epoch $T_{ep}-1$; (b) We train the backbone (i.e., DLA34 [40]) and the identity embedding and data association branches of FDMOT with the pseudo identity labels, and train the detection branch with the manual detection labels.

At the epoch T, we adopt a fully connected layer and a softmax function to map the noise-decoupled and well-refined embedding features $\mathbf{e}_t$ and $\mathbf{e}_t^{wr}$ to their corresponding identities $\mathbf{p}_t$ and $\mathbf{p}_t^{wr}$, respectively. During training, we calculate the joint loss L_t by the detection loss L_{det} (used in CenterNet [40]) and the identity embedding losses L_{ie} and L_{ea}. Mathematically, the above weakly supervised training process can be defined as

$$L_t = L_{det}(\hat{\mathbf{d}}_t, \mathbf{d}_t) + L_{ie}(\hat{\mathbf{p}}_t, \mathbf{p}_t) + L_{ea}(\hat{\mathbf{p}}_t, \mathbf{p}_t^{wr}), \tag{B12}$$

where L_{ie} and L_{ea} represent the common k-class cross-entropy losses. k is the identity number of $\hat{\mathbf{p}}_{T_{ep}}$. $\hat{\mathbf{p}}_t$ denote the pseudo identity labels for the current frame t. $\hat{\mathbf{d}}_t$ and $\mathbf{d}_t$ denote the ground-truth detection labels and the detection predictions. $\mathbf{p}_t$ and $\mathbf{p}^{wr}_t$ are the predicted identities of the noise-decoupled and well-refined embedding features.

With the joint training of the three branches of FDMOT, the proposed method effectively generates the reliable detection predictions and the well-refined embedding features by decoupling noisy features with the feature-decoupling Transformer.

Appendix C Experiments

Appendix C.1 Experiment Settings

Datasets. To show the effectiveness of FDMOT, we evaluate its performance on the two challenging datasets, MOT17 [21] and MOT20 [6]. For ablation studies in Table B1, we follow the previous methods [30, 39] and respectively adopt the first half and the last half of the MOT17 training set for training and validation. Similar to the previous methods [15, 24, 30, 38, 39], we use the extra datasets (i.e., CrowdHuman [23], Cityperson [37], ETHZ [11]) and the MOT17/MOT20 training set for training, while employing the MOT17/MOT20 test set for evaluation.

Evaluation metrics. To perform comprehensive comparisons of FDMOT with several state-of-the-art methods, we employ the common MOT evaluation metrics, such as the CLEAR metrics [3] (including Multi-Object Tracking Accuracy (MOTA), Mostly Tracked Targets (MT), Mostly Lost Targets (ML), IDs, etc.), ID F1 Score (IDF1) [22] and Higher Order Tracking Accuracy (HOTA) [19]. In particular, MOTA and HOTA evaluate the comprehensive performance of both detectors and data association algorithms, but MOTA focuses more on the discovery capabilities of detectors. IDF1 evaluates the identity preservation ability of the competing methods.

Implementation details. During training, we optimize the feature-decoupling Transformer-based network (using DLA34 [40] as the backbone) for 30 epochs with a learning rate of 1×10^{-4} and a mini-batch size of 12 on three RTX2080Ti GPUs when using an RTX2080Ti GPU for testing. We adopt 2 layers of feature-decoupling Transformers in the detection and identity embedding branches and 1 layer in the data association branch. Moreover, we set the detection threshold α to 0.4.

Appendix C.2 Ablation Studies

To verify the effectiveness of the proposed FDMOT, we perform several ablation studies for the influences of the feature-decoupling Transformer-based branches on the MOT17 validation set. From Table B1 to Table B4, we train the identity embedding and data as-

Table B3 Ablation study for the influence of the feature-decoupling Transformer-based data association branch.

FDMOT-Det.&Emb.	FDMOT-Asso.	MOTA↑	IDF1↑	IDS↓
None	FDT-N_Mix	68.8	73.7	336
None	FDT-S_Mix	69.0	73.3	328
None	FDT-F_Mix	69.3	73.6	340
None	FDT-W_FFN	69.1	73.8	337
None	FDT-W_Rep	69.5	74.3	**281**
None	FDT-W_Res	**69.6**	74.1	299
None	FDT-W_Mix	**69.6**	**74.3**	307

Table B4 Ablation study for the influence of the cooperation among the three feature-decoupling Transformer-based branches.

FDMOT-Det.&Emb.	FDMOT-Asso.	MOTA↑	IDF1↑	IDS↓
FDT-W_Mix	FDT-W_FFN	69.9	73.9	299
FDT-W_Mix	FDT-W_Rep	70.2	74.2	312
FDT-W_Mix	FDT-W_Res	70.7	74.0	289
FDT-W_FFN	FDT-W_FFN	70.8	74.4	**260**
FDT-W_Rep	FDT-W_Rep	70.6	74.3	281
FDT-W_Res	FDT-W_Res	70.7	74.4	303
FDT-W_Mix	FDT-W_Mix	**70.9**	**75.2**	282

sociation branches of FDMOT variants with the pseudo identity labels unless otherwise stated. In these tables, FDMOT-Det.&Emb. denotes the variants of feature-decoupling Transformer used in the detection and identity embedding branches. FDMOT-Asso. represents the variants of feature-decoupling Transformer used in the data association branch of FDMOT. ↑ (↓) means that the higher (lower) performance is the better. We mark the best result in bold.

Supervised vs. weakly supervised tracking. We evaluate the influence of the pseudo identity labels on the tracking performance in Table B1. The results show that the weakly supervised baseline (None+None) obtains poor performance on MOTA, IDF1 and IDS compared with the supervised baseline (None*+None*) [39]. On the contrary, the FDMOT (FDT-W_Mix+FDT-W_Mix) variant achieves comparable performance against the supervised FDMOT (FDT-W_Mix*+FDT-W_Mix*) variant on the three metrics. FDMOT effectively reduces the adverse influence of the branch competition caused by the heavily or completely occluded targets and extracts well-refined embedding features by the feature-decoupling Transformer module. The comparison results show that FDMOT suppresses the adverse influence of the pseudo identity labels and generates robust trajectories by feature decoupling.

Influence of the feature-decoupling Transformer-based detection and identity embedding branches. We adopt the feature-decoupling Transformer module in the detection and identity embedding branches, and evaluate its influence on detection and identity embedding based on the metrics MOTA and IDS. The FDMOT variants achieve higher and different tracking results in terms of different feature-decoupling Transformer modules against the weakly supervised baseline [39]. In particular, the FDT-W_Mix+None variant obtains remarkable performance with +2.5 MOTA and -64% IDS, which validates that performing wavelet-domain feature decoupling for detection and identity embedding is necessary, while extracting reliable detection predictions and embedding features for weakly supervised trajectory generation.

Influence of the feature-decoupling Transformer-based data association branch. In Table B3, we adopt different feature-decoupling Transformer variants in the data association branch, and achieve different tracking results on IDF1. The None+FDT-W_Mix variant obtains the best improvement of +5.2 IDF1 against the weakly supervised baseline, which owes its success to the embedding features refined by the feature-decoupling Transformer module. Moreover, it only has an improvement of +0.1 IDF1 against the FDT-W_Mix+None variant, and drops the performance from 75.2 to 74.3 on IDF1, compared with the FDT-W_Mix+FDT-W_Mix variant. The comparison results show the important role of the feature-decoupling Transformer-based detection and identity embedding branches in extracting the well-refined embedding features for data association.

Influence of the cooperation among the three feature-decoupling Transformer-based branches. We evaluate the cooperation relationship among the three feature-decoupling Transformer-based branches in Table B4. The FDMOT variants achieve better performance on MOTA and IDF1 than the variants in Table B2 and B3, and they obtain different performances on MOTA and IDF1 in terms of different feature-decoupling Transformer variants. It is worth noting that there is a performance drop of -1.0 MOTA and -1.3 IDF1 from the FDT-W_Mix+FDT-W_Mix variant to the FDT-W_Mix+FDT-W_FFN variant. The above observations verify the cooperation relationship among the three feature-decoupling Transformer-based branches while extracting detection predictions and embedding features for weakly supervised tracking. Moreover, adopting the same rather than the different feature-decoupling Transformer modules in the three branches of FDMOT is more beneficial for robust tracking. Therefore, we select the FDT-W_Mix+FDT-W_Mix variant for experiments on the MOT17 and MOT20 test sets.

Qualitative comparisons with the predicted trajectories. We report the predicted trajectories obtained by the baseline [39]

Table C1 Comparisons of FDMOT with state-of-the-art methods on the MOT17 test set. * denotes the method is based on a high-cost detector. † denotes the method is based on weakly supervised training. For each metric, we mark the best result in bold.

Method	MOTA↑	IDF1↑	HOTA↑	MT↑	ML↓	IDS↓	FPS↑
CenterTrack [41]	67.8	64.7	52.2	34.6	24.5	3039	3.8
FairMOT [39]	73.7	72.3	59.3	43.2	17.3	3303	25.9
TransCenter [33]	73.2	62.2	54.5	41.1	19.0	2964	1.0
CorrTracker [30]	76.5	73.6	60.7	47.6	12.7	3369	15.6
GSDT_V2 [31]	73.2	66.5	55.2	41.7	19.0	3891	4.9
TraDeS [32]	69.1	63.9	52.7	36.4	21.5	3555	22.3
MAATrack* [24]	79.4	75.9	62.0	57.6	12.0	1452	-
ByteTrack* [38]	80.3	77.3	63.1	53.2	14.5	2196	29.6
Unicorn* [35]	77.2	75.5	61.7	58.7	11.2	5379	-
MeMOT [5]	72.5	69.0	56.9	43.8	18.0	2724	-
MCM† [14]	48.1	-	-	17.7	39.8	2328	-
SUMOT† [15]	61.7	58.1	-	-	-	1864	-
UTrack† [18]	73.5	70.2	-	43.3	**15.2**	4110	**25.4**
V-Spatial† [1]	56.8	58.3	-	27.9	28.3	-	-
UEANet† [16]	77.2	77.0	62.7	41.7	19.0	**1533**	25.1
FDMOT(Ours)†	**78.9**	**77.5**	**63.6**	**48.4**	15.4	1812	23.8

and FDMOT in Figure A2 to qualitatively compare their performance. The weakly supervised baseline obtains worse detection predictions (colored squares) and trajectories than the supervised baseline, as the pseudo identity labels are noisy when generated with the detection predictions of occluded targets. The proposed FDMOT brings more reliable detection predictions (colored squares) for occluded targets than the baseline and associates them into robust trajectories, since the pseudo identity labels may lose some identity labels of the heavily or completely occluded targets, which reduces the branch competition and improves the robustness of detection predictions. As a result, FDMOT surpasses the baseline and achieves comparable tracking results against the supervised FDMOT (based on the supervised training) according to the completeness of trajectories and the ground truth.

Appendix C.3 Experiments on the MOT17 and MOT20 Test Sets

Results on MOT17. Table C1 shows that FDMOT achieves comparable or even better performance than previous supervised methods on MOT17. For example, FDMOT achieves competing tracking results compared with the high-cost detector-based methods (such as MAATrack [24], Unicorn [35] and ByteTrack [38]) on MOT17. Moreover, it surpasses the previous weakly supervised methods [1, 14–16, 18] by at least +1.7 MOTA, +0.5 IDF1 and +0.9 HOTA. As a result, our FDMOT alleviates the adverse influence of noisy pseudo identity labels by extracting reliable detection predictions and embedding features to accurately detect and track multiple targets with the feature-decoupling Transformer module.

Results on MOT20. While dealing with the more crowded scenes of MOT20 (compared with MOT17), our FDMOT obtains comparable results on the metrics, MOTA, HOTA and ML [3] against previous supervised MOT methods in Table C2. In particular, FDMOT achieves a higher IDF1 (i.e., 76.0) than those high-cost detector-based supervised methods [5,24,38]. Moreover, it surpasses other weakly supervised methods in terms of HOTA and outperforms the high-cost Transformer-based methods [5,25,33] by at least +9.1 MOTA, +9.9 IDF1 and +5.4 HOTA. The previous weakly supervised MOT methods and our method achieve worse tracking results than the ones in Table C1 on MOTA, IDF1 and HOTA due to the much heavier occlusion. Fortunately, our FDMOT still obtains the first performance on IDF1 and HOTA, which shows the better capability of FDMOT to alleviate the influence of occluded targets on weakly supervised tracking. The comparison results confirm 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.

Visualization of tracking results We also report the tracking results obtained by FDMOT on the MOT17 and MOT20 test sets to show its weakly supervised tracking capability in Figure C1. The proposed FDMOT effectively generates reliable detection predictions and final trajectories for multiple targets especially occluded targets in crowded scenes. FDMOT obtains remarkable tracking performance on the MOT17 and MOT20 datasets, since it effectively discovers and tracks heavily occluded targets in crowded scenes. There are many real-world applications of MOT technology for computer vision systems. As shown in Figure C2, we employ FDMOT to track multiple targets with a vehicle camera in a business district, which further verifies the effectiveness of our method in real-world traffic scenes.

Performance changes while the occlusion of targets increases. The MOT17 dataset has the same videos as MOT16. But the MOT17 dataset provides more detection labels and identity labels of heavily occluded targets. So the occlusion of targets increases from MOT16 to MOT17. We give the comparison results of the previous weakly supervised MOT methods and our FDMOT on the MOT16 and MOT17 test sets. As shown in Table C3, when the occlusion increases from MOT16 to MOT17,

Table C2 Comparisons of FDMOT with some state-of-the-art methods on the MOT20 test set. * denotes the method is based on a high-cost detector. † denotes the method is based on weakly supervised training. For each metric, we mark the best result in bold.

Method	MOTA↑	IDF1↑	HOTA↑	MT↑	ML↓	IDS↓	FPS↑
TransCenter [33]	61.9	50.4	44.3	49.4	15.5	4653	1.0
TransTrack [25]	64.5	59.2	48.5	49.1	13.6	3565	14.9
FairMOT [39]	61.8	67.3	54.6	68.8	7.6	5243	13.2
CorrTracker [30]	65.2	69.1	-	66.4	8.9	5183	8.5
GSDT [31]	67.1	67.5	53.6	53.1	13.2	3230	1.5
MAATrack* [24]	73.9	71.2	57.3	59.7	12.3	1331	-
ByteTrack* [38]	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† [18]	68.5	69.4	-	57.9	12.2	2147	12.4
UEANet† [16]	**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

the previous weakly supervised MOT methods obtain large performance drops, but the proposed FDMOT achieves a minimal performance drop on the evaluation metrics, MOTA, IDF1 and IDS. The comparison results show that our method alleviates the influence of heavily occluded targets on tracking performance. We update the descriptions of the performance changes following your suggestion in the revised manuscript and the appendix.

References

1 Bastani, F, He S, Madden S. Self-supervised multi-object tracking with cross-input consistency, in: Advances in Neural Information Processing Systems, Virtual, 2021. 13695–13706
2 Wang M, Zhang Y, Dong h, et al. Trajectory tracking control of a bionic robotic fish based on iterative. Science China Information Sciences, 2020, 63(7): 1-9
3 Bernardin K, Stiefelhagen R. Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing, 2008: 1–10
4 Xue Z, Wu W. Anomaly detection by exploiting the tracking trajectory in surveillance videos. Science China Information Sciences, 2020, 63(5): 1-3
5 Cai J, Xu M, Li W, et al. Memot: Multi-object tracking with memory, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, 2022. 8090–8100
6 Dendorfer P, Rezatofighi H, Milan A, et al. Mot20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003, 2020
7 Zhang D, Li T, Chen C L, et al. Target tracking algorithm based on a broad learning system. Science China Information Sciences, 2022, 65(5): 1-3
8 Ding X, Xia C, Zhang X, et al. Repmlp: re-parameterizing convolutions into fully-connected layers for image recognition. arXiv preprint arXiv:2105.01883, 2021
9 Tu Z, Pan W, Duan Y, et al. RGBT tracking via reliable feature configuration. Science China Information Sciences, 2022, 65(4): 1-13
10 Du Y, Song Y, Yang B, et al. Strongsort: Make deepsort great again. arXiv preprint arXiv:2202.13514, 2022
11 Ess A, Leibe B, Schindler K, et al. A mobile vision system for robust multi-person tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Anchorage, 2008. 1–8
12 Yong K, Chen M, Wu Q. Noncertainty-equivalent observer-based noncooperative target tracking control for unmanned aerial vehicles. Science China Information Sciences, 2022, 65(5): 1-15
13 Guo S, Wang J, Wang X, et al. Online multiple object tracking with cross-task synergy, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, virtual, 2021. 8136–8145
14 Ho K, Kardoost A, Pfreundt F J, et al. A two-stage minimum cost multicut approach to self-supervised multiple person tracking, in: Proceedings of the Asian Conference on Computer Vision, Singapore, 2021
15 Karthik S, Prabhu A, Gandhi V. Simple unsupervised multi-object tracking. arXiv preprint arXiv:2006.02609, 2020
16 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
17 Liang C, Zhang Z, Zhou X, et al. Rethinking the competition between detection and reid in multi-object tracking. IEEE Transactions on Image Processing, 2022, 31: 3182–3196
18 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
19 Luiten J, Osep A, Dendorfer P, et al. Hota: A higher order metric for evaluating multi-object tracking. International journal of computer vision, 2021, 129: 548–578
20 Meinhardt T, Kirillov A, Leal-Taixé L, et al. Trackformer: Multi-object tracking with Transformers. arXiv preprint arXiv:2101.02702,2021
21 Milan A, Leal-Taixé, L, Reid I, et al. Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831, 2016
22 Ristani E, Solera F, Zou R, et al. Performance measures and a data set for multi-target, multi-camera tracking, in: Proceedings

Figure C1 Example images and tracking results obtained by FDMOT. Each row shows the sampled frames with a 20-frame interval form (a) MOT17 or (b) MOT20.

Table C3 Comparisons of the performance changes while the occlusion of targets increases. For each metric, we mark the best result in bold.

Method	MOTA↑	IDF1↑	HOTA↑	MT↑	ML↓	IDS↓	FPS↑
MOT16							
SUMOT† [15]	62.4	58.5	-	-	-	588	-
UTrack† [18]	74.2	71.1	-	44.8	**14.0**	1324	24.8
UEANet† [16]	77.9	77.6	63.2	43.5	17.3	**491**	**25.1**
FDMOT(Ours)†	**79.0**	**77.8**	**63.9**	**49.9**	14.4	1721	23.8
MOT17							
SUMOT† [15]	61.7	58.1	-	-	-	1864	-
UTrack† [18]	73.5	70.2	-	43.3	**15.2**	4110	**25.4**
UEANet† [16]	77.2	77.0	62.7	41.7	19.0	**1533**	25.1
FDMOT(Ours)†	**78.9**	**77.5**	**63.6**	**48.4**	15.4	1812	23.8
From MOT16 to MOT17, Performance Changes							
SUMOT† [15]	-0.7	-0.4	-	-	-	+1298	-
UTrack† [18]	-0.7	-0.9	-	**-1.5**	+1.2	+2786	**+0.6**
UEANet† [16]	-0.7	-0.6	-0.5	-1.8	+1.7	+1042	+0.0
FDMOT(Ours)†	**-0.1**	**-0.3**	**-0.3**	**-1.5**	**+1.0**	**+91**	+0.0

of the European Conference on Computer Vision, Amsterdam, 2016. 17–35
23 Shao S, Zhao Z, Li B, et al. Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123, 2018
24 Stadler D, Beyerer J. Modelling ambiguous assignments for multi-person tracking in crowds, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, Waikoloa, 2022. 133–142
25 Sun P, Cao J, Jiang Y, et al. Transtrack: Multiple object tracking with Transformer. arXiv preprint arXiv:2012.15460, 2020
26 Sun S, Akhtar N, Song H, et al. Deep affinity network for multiple object tracking. IEEE transactions on pattern analysis and machine intelligence, 2021, 43: 104–119
27 Tolstikhin I O, Houlsby N, Kolesnikov A, et al. Mlp-mixer: An all-mlp architecture for vision, in: Advances in Neural Information Processing Systems, Virtual, 2021. 24261–24272
28 Touvron H, Bojanowski P, Caron M, et al. Resmlp: Feedforward networks for image classification with data-efficient training. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022
29 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need, in: Advances in Neural Information Processing Systems, Virtual, Long Beach, 2017. 5998–6008
30 Wang Q, Zheng Y, Pan P, et al. Multiple object tracking with correlation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2021. 3876–3886
31 Wang Y, Kitani K, Weng X. Joint object detection and multi-object tracking with graph neural networks, in: Proceedings of the IEEE International Conference on Robotics and Automation, Xi'an, 2021. 13708–13715
32 Wu J, Cao J, Song L, et al. Track to detect and segment: An online multi-object tracker, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2021. 12352–12361
33 Xu Y, Ban Y, Delorme G, et al. Transcenter: Transformers with dense queries for multiple-object tracking. arXiv preprint arXiv:2103.15145, 2021
34 Xu Y, Osep A, Ban Y, et al. How to train your deep multi-object tracker, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 2020. 6786–6795
35 Yan B, Jiang Y, Sun P, et al. Towards grand unification of object tracking, in: Proceedings of the European Conference on Computer Vision, Tel Aviv, 2022. 733–751
36 Zeng F, Dong B, Zhang Y, et al. Motr: End-to-end multiple-object tracking with Transformer, in: Proceedings of the European Conference on Computer Vision, Tel Aviv, 2022. 659–675
37 Zhang S, Benenson R, Schiele B. Citypersons: A diverse dataset for pedestrian detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 3213–3221
38 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
39 Zhang Y, Wang C, Wang X, et al. Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision, 2021, 129: 3069–3087
40 Zhou X, Wang D, Krähenbühl P. Objects as points. arXiv preprint arXiv:1904.07850, 2019
41 Zhou X, Koltun V, Krähenbühl P. Tracking objects as points, in: Proceedings of the European Conference on Computer Vision, Virtual, 2020. 474–490
42 He K M, Gkioxari G, Dollár P, Girshick R B. Tracking objects as points, in: Proceedings of the European Conference on Computer Vision, Venice, 2017. 2980-2988

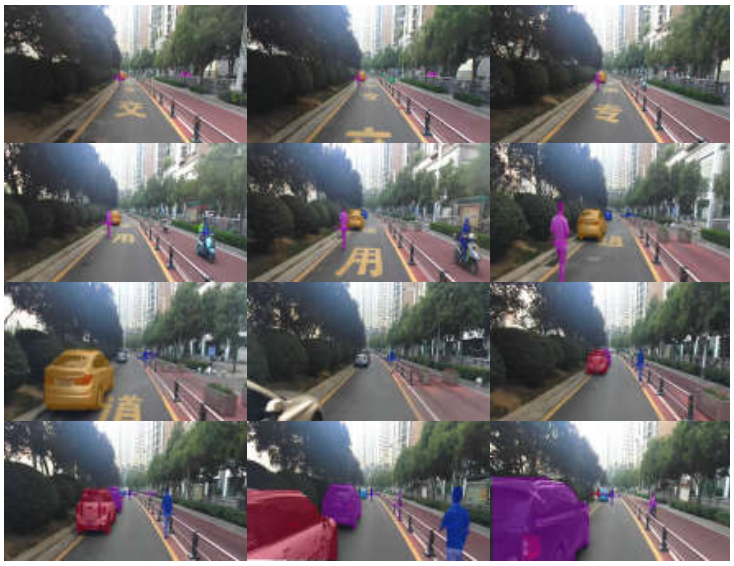

Figure C2 An example application of FDMOT on detecting and tracking multiple targets in a real-world traffic scene. In addition, we adopt pixel-level instances (generated by Mask-RCNN [42]) to locate cars and persons, and then associate them into robust trajectories with the identity embedding and data association branches of FDMOT.