DiffMOT: A Real-time Diffusion-based Multiple Object Tracker with Non-linear Prediction

Supplementary Material

A. Preliminary

The Diffusion Probabilistic Model (DPM) [5] has shown great potential in modeling non-linear mapping, yet it suffers from prolonged inference time caused by thousands of sampling steps. DDM [6] attempts to speed up the inference process by applying a decoupled diffusion process. Specifically, the forward process of decoupled diffusion is split into the analytic image attenuation process and the increasing process of normal noise:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0 + \int_0^t \mathbf{f}_t \mathrm{d}t, t\mathbf{I}), \tag{1}$$

where $\mathbf{x}_0$ and $\mathbf{x}_t$ are the clean and noisy signals respectively, $\mathbf{f}_t$ denotes the analytic function representing the attenuation velocity of $\mathbf{x}_0$ over time t ($t \in [0, 1]$), and $\mathbf{I}$ is the identity matrix. In practice, the proposed D^{2}MP uses the specific form—constant function: $\mathbf{f}_t = \mathbf{c}$. [6] has proved that the corresponding reversed process supports sampling with arbitrary time interval Δt and is expressed by:

$$\begin{aligned} q(\mathbf{x}_{t-\Delta t}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t + \int_t^{t-\Delta t} \mathbf{f}_t \mathrm{d}t \\ - \frac{\Delta t}{\sqrt{t}}\boldsymbol{z}, \frac{\Delta t(t-\Delta t)}{t}\mathbf{I}), \end{aligned} \tag{2}$$

where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, is the noise added on $\mathbf{x}_0$. Actually, $\boldsymbol{z}$ and $\mathbf{f}_t$ are unknown in the reversed process, therefore, we need to parameterize $\mathbf{f}_t$ and $\boldsymbol{z}$ using a neural network $\boldsymbol{\Theta}$. In the training stage, the decoupled diffusion model uses $\boldsymbol{z}$ and $\mathbf{f}_t$ to supervise the parameterized $\boldsymbol{z}_{\boldsymbol{\Theta}}$ and $\mathbf{f}_{\boldsymbol{\Theta}}$ simultaneously:

$$\min_{\boldsymbol{\Theta}} \mathbb{E}_{q(\mathbf{x}_0)} \mathbb{E}_{q(\boldsymbol{z})} [\|\mathbf{f}_{\boldsymbol{\Theta}} - \mathbf{f}\|^2 + \|\boldsymbol{z}_{\boldsymbol{\Theta}} - \boldsymbol{z}\|^2]. \tag{3}$$

The reversed process aims to generate $\mathbf{x}_0$ from $\mathbf{x}_1$ via Eq. 2 iteratively. Due to the analyticity of the image attenuation process, we can conduct one-step sampling when $\Delta t = t = 1$, removing the low speed of iterative generation.

B. Pseudo-code of DiffMOT

The inference of DiffMOT consists of three parts: detection, motion prediction, and association, and the pseudo-code is shown in Alg. 1. For the f-th frame of the given video sequence, we use a detector to obtain the bounding boxes of objects. The detections are divided into two groups according to their confidence scores ($det.conf$). Specifically, we set two different thresholds τ_{high} and τ_{low} and group the detections by:

$$\begin{cases} \mathcal{D}_{first} = \mathcal{D}_{first} \cup \{det\} & det.conf > \tau_{high} \\ \mathcal{D}_{second} = \mathcal{D}_{second} \cup \{det\} & \tau_{low} < det.conf < \tau_{high}. \end{cases} \tag{4}$$

On the other hand, we use the proposed D^{2}MP to obtain the predicted boxes of objects in the previous trajectories. During the association stage, we match the detected and predicted bounding boxes twice since the detections are divided into two groups. We first match D_{first} with the predicted bounding boxes via the similarity of reid features and IoU of bounding boxes. Afterwards, D_{second} is matched with the predicted bounding boxes via IoU. The matched detections will update the trajectories. The unmatched tracks will be deleted. The unmatched detections will be initialized as new tracks.

Method	HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑
DiffusionTrack[8]	55.3	66.3	51.3	72.8	59.9
MotionTrack[10]	59.7	71.3	56.8	76.4	-
ByteTrack[11]	61.3	75.2	59.6	77.8	63.4
OC-SORT[1]	62.4	76.3	62.5	75.7	62.4
StrongSORT[4]	62.6	77.0	64.0	73.8	61.3
SparseTrack[7]	63.5	77.6	63.1	**78.1**	**64.1**
Deep OC-SORT[9]	**63.9**	**79.2**	**65.7**	75.6	62.4
DiffMOT	61.7	74.9	60.5	76.7	63.2

Table 1. Comparison with SOTA MOT trackers on the MOT20 test sets under the "private detector" protocol. All methods use the same YOLOX detector. ↑ means the higher the better and ↓ means the lower the better. **Bold** numbers indicate the best result.

C. Benchmark Evaluation on MOT20

MOT20 [3] is also one of the commonly used pedestrian-dominant datasets in MOT, characterized by higher density, and the motion is more closely approximated as linear. We conduct the experiment on the MOT20 test sets under the "private detector" protocol to further demonstrate the performance of DiffMOT on pedestrian-dominant scenarios. As shown in Tab. 1, DiffMOT achieves 61.7% HOTA, 74.9% IDF1, 60.5% AssA, 76.7% MOTA, and 63.2% DetA. The results indicate that even in pedestrian-dominant scenarios, DiffMOT, designed specifically for non-linear motion scenarios, can achieve comparable performance.

Algorithm 1: Pseudo-code of DiffMOT.

Input : A video sequence V; the detector $\mathbf{D}$; HMINet model $\mathbf{M}$; detection score threshold τ_{high}, τ_{low}; tracking score threshold ϵ

Parameter: Detections $\mathcal{D}_f$; predicted boxes P_f; pure noise z; objects motion M_f; conditions C_f

Output : Tracks $\mathcal{T}$ of the video

Initialization: $\mathcal{T} \leftarrow \emptyset$
for *frame* f *in* V **do**
  $\mathcal{D}_{first} \leftarrow \emptyset$; $\mathcal{D}_{second} \leftarrow \emptyset$
  /* Detection */
  $\mathcal{D}_f \leftarrow \mathbf{D}(f)$
  for det *in* $\mathcal{D}_f$ **do**
    if $det.conf > \tau_{high}$ **then**
      $\mathcal{D}_{first} \leftarrow \mathcal{D}_{first} \cup \{det\}$
    end
    else if $\tau_{low} < det.conf < \tau_{high}$ **then**
      $\mathcal{D}_{second} \leftarrow \mathcal{D}_{second} \cup \{det\}$
    end
  end
  /* Motion Prediction */
  for trk *in* $\mathcal{T}$ **do**
    $trk.C_f \leftarrow$ motion information from trk
    $trk.M_f \leftarrow \mathbf{M}(z, trk.C_f)$
    $trk.P_f \leftarrow trk.M_f + trk.last\_location$
  end
  /* Association */
  Match P_f and D_{first} using reid feature and IoU
  $P_f^{remain} \leftarrow$ remaining predicted boxes from P_f
  $D_f^{remain} \leftarrow$ remaining detected boxes from D_{first}
  Match P_f^{remain} and D_{second} using IoU
  $P_f^{re-remain} \leftarrow$ remaining predicted boxes from P_f^{remain}
  /* Delete unmatched tracks */
  $\mathcal{T}^{unmatched} \leftarrow$ remaining unmatched tracks from $P_f^{re-remain}$
  $\mathcal{T} \leftarrow \mathcal{T} \backslash \mathcal{T}^{unmatched}$
  /* Initialize new tracks */
  for det *in* D_f^{remain} **do**
    if $det.conf > \epsilon$ **then**
      $\mathcal{T} \leftarrow \mathcal{T} \cup \{det\}$
    end
  end
end
Return: $\mathcal{T}$

Train Dataset	Test Dataset	HOTA↑	IDF1↑	MOTA↑
SportsMOT	SportsMOT	76.2	76.1	97.1
DanceTrack	SportsMOT	75.6	75.1	97.1
MOT17	MOT17	64.5	79.3	79.8
DanceTrack	MOT17	61.8	74.4	79.1
MOT20	MOT20	61.7	74.9	76.7
DanceTrack	MOT20	61.2	73.8	76.2

Table 2. Generalization experiments for D^{2}MP.

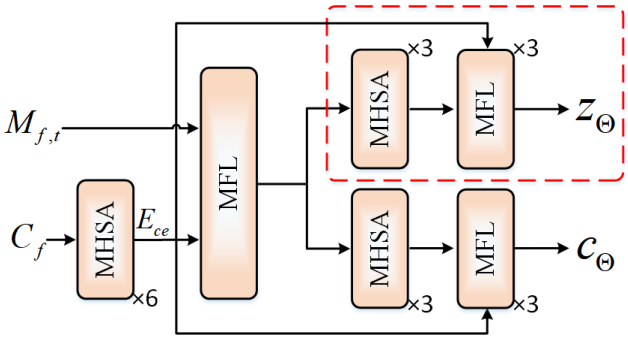
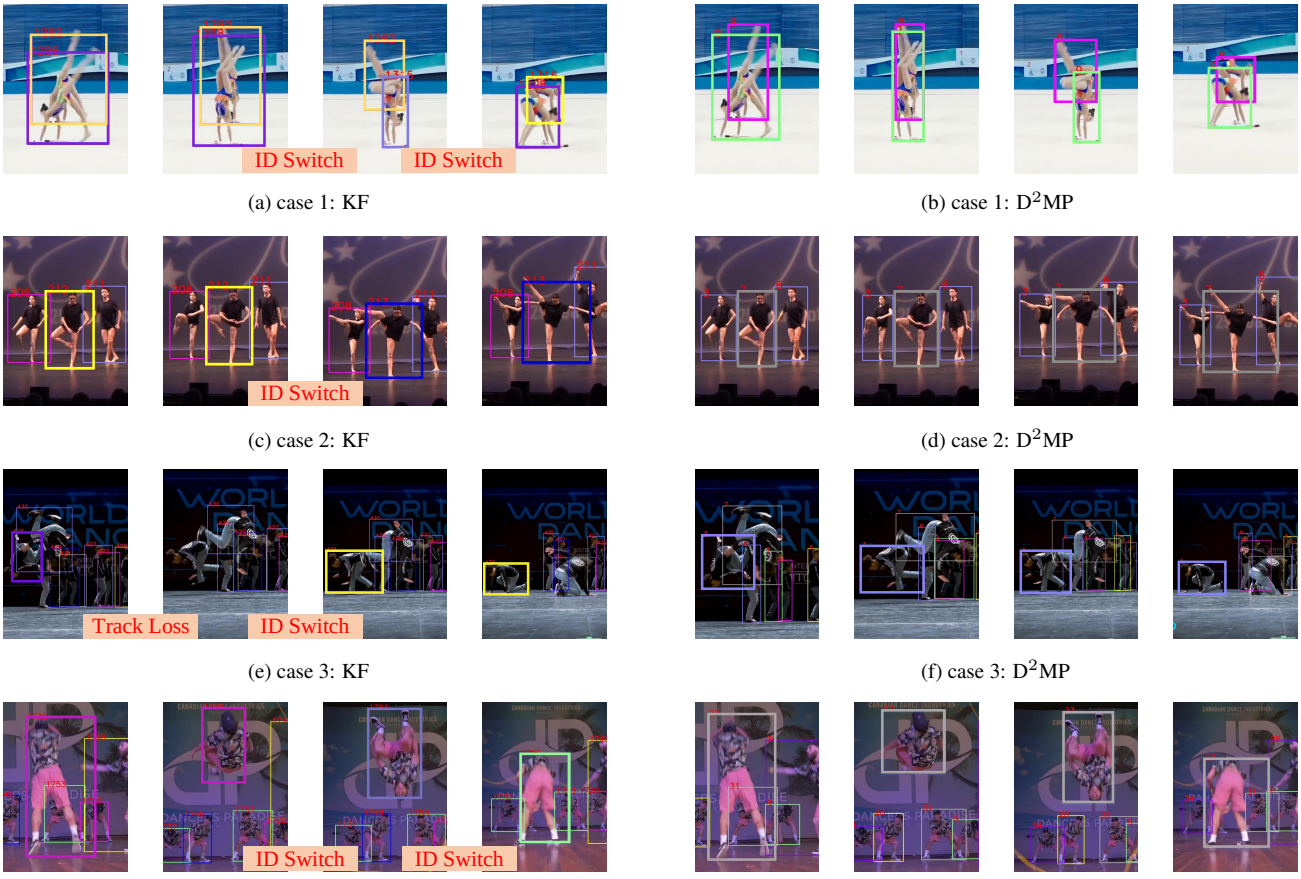


Figure 1. The architecture of D^{2}MP-TB. The distinction from D^{2}MP-OB is enclosed within the red dashed box.

D. Generalization of D^{2}MP

To demonstrate the generalization of our D^{2}MP, we directly use the model trained on the DanceTrack dataset to test the performances on other datasets. The results are shown in Tab. 2. As we can see in the table, on SportsMOT, the model trained on DanceTrack is only 0.6% and 1.0% lower than the model trained on SportsMOT in HOTA and IDF1. On MOT17 and MOT20, the model trained with DanceTrack is 2.7% / 0.5% and 4.9% / 1.1% lower than the model trained on MOT17 / 20 in HOTA and IDF1. Compared with Tab. **??**, Tab. **??**, and Tab. 1, the model trained solely on DanceTrack has already achieved performance comparable to the state-of-the-art methods. Especially on SportsMOT, the model trained solely on DanceTrack has achieved SOTA performance, which outperforms previous SOTA MixSort-OC [2] by 1.4% in HOTA, 0.7% in IDF1, and 0.6% in MOTA. The above observation indicates that D^{2}MP possesses strong generalization capabilities, as it directly learns the distribution of all objects' motion using the diffusion model rather than learning individual object trajectories. Our model can be applied to new scenarios without retraining, demonstrating the advantage of using the diffusion model for motion prediction.



(a) case 1: KF (b) case 1: D^{2}MP

(c) case 2: KF (d) case 2: D^{2}MP

(e) case 3: KF (f) case 3: D^{2}MP

(g) case 4: KF (h) case 4: D^{2}MP

Figure 2. Qualitative comparison between using KF or D^{2}MP as the motion model on the DanceTrack test set. (a), (c), (e), and (g) represent the results using KF as the motion model. (b), (d), (f), and (h) represent the results using D^{2}MP as the motion model. Each pair of rows shows the comparison of the results for one sequence. Boxes of the same color represent the same ID. Best viewed in color and zoom-in.

E. Architecture of D^{2}MP-TB

We have conducted experiments on different architectures of D^{2}MP in the ablation study. The architecture of D^{2}MP-OB can be observed in Figure 3 in the manuscript. As shown in in Fig. 1, D^{2}MP-TB is a two-branch structure and predicts $\boldsymbol{z}_{\Theta}$ and $\mathbf{c}_{\Theta}$ respectively. The distinction from D^{2}MP-OB is enclosed within the red dashed box in the figure.

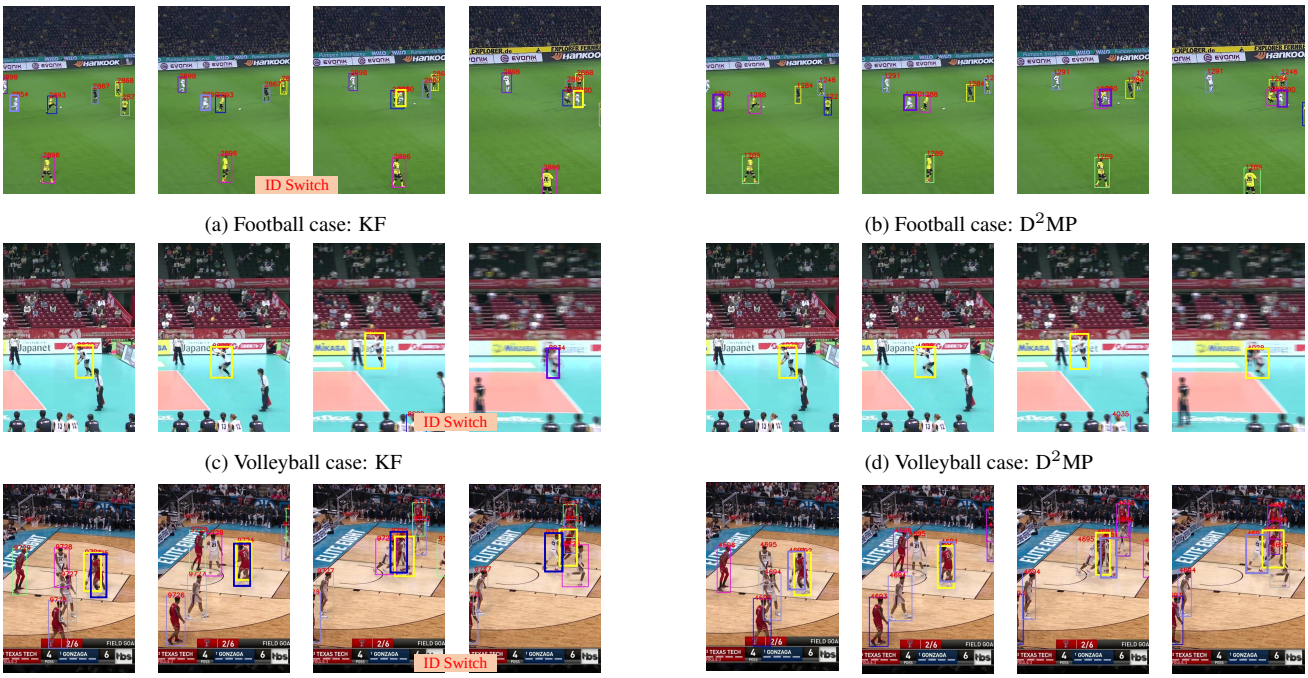
F. More Visualization

F.1. Qualitative Comparison on DanceTrack

Fig. 2 shows some samples on the test set of DanceTrack where trackers with KF suffer from discontinuous trajectories and high ID switches while DiffMOT with D^{2}MP has strong robustness in non-linear motion scenes. For example, the issue happens on the tracking results by trackers with KF at: (a) ID1298 $\rightarrow$ ID1315, and ID1282 $\rightarrow$ ID1316; (c) ID212 $\rightarrow$ ID213; (e) ID426 being lost and then switch to ID436; (g) ID1736 $\rightarrow$ ID1766 $\rightarrow$ ID1769. The visual comparison indicates that when the objects exhibit non-linear motions in dance scenarios, trackers with KF are unable to predict the accurate trajectories' position, resulting in a large ID switch. In contrast, D^{2}MP exhibits greater robustness in handling these non-linear motions.

F.2. Qualitative Comparison on SportsMOT

Fig. 3 depicts more qualitative comparisons between employing KF and D^{2}MP as the motion model on the test set of SportsMOT. We select samples from diverse scenes, including football, volleyball, and basketball scenes. The issue happens on the tracking results by trackers with KF at: (a) ID2854 $\rightarrow$ ID2900; (c) ID8820 $\rightarrow$ ID8834; (e) ID switch between ID9724 and ID9725. The visual comparison in the figure highlights that when the objects exhibit non-linear motions such as acceleration or deceleration in sports scenarios, KF often hard to provide accurate predictions, while DiffMOT demonstrates the ability to predict the objects' position accurately in such scenarios.



(a) Football case: KF (b) Football case: D^{2}MP

(c) Volleyball case: KF (d) Volleyball case: D^{2}MP

(e) Basketball case: KF (f) Basketball case: D^{2}MP

Figure 3. Qualitative comparison between using KF or D^{2}MP as the motion model on the SportsMOT test set. (a), (c), and (e) represent the results using KF as the motion model. (b), (d), and (f) represent the results using D^{2}MP as the motion model. Each pair of rows shows the comparison of the results for one sequence. All of the cases are in scenarios with a moving camera. Boxes of the same color represent the same ID. Best viewed in color and zoom-in.

F.3. Visual Results on MOT17/20

Fig. 4 and Fig. 5 show several tracking results of our DiffMOT on the test set of MOT17 and MOT20, respectively. It can be observed that Although the proposed DiffMOT is designed specifically for non-linear motion scenes, it can still achieve appealing results.

G. Illustration of Failure Cases

We visualize two failure cases in Fig. 6. For the first case, when different objects are passing through each other, objects with ID "1" and "2", as well as objects with ID "4" and "5", underwent an exchange of identities. This is due to the absence of velocity direction constraints in the motion model. We believe that incorporating velocity direction constraints to restrict the generation of predicted boxes could help address this issue.

In the second case, the object with ID "3" disappears in "Frame 2" and reappears in "Frame 3" as ID "7". Simultaneously, the object with ID "4" disappears in "Frame 2" and reappears in "Frame 4" as ID "1", while the object with original ID "1" becomes a new ID "8". This phenomenon occurs due to the difficulty in recovering long-term lost objects of our motion model. When an object is lost for an extended period, it becomes challenging to re-associate the object accurately, leading to the generation of new IDs or ID switches. In our future work, we intend to explore the generation of multi-frame trajectories to improve the motion model's capacity for long-term matching.

References

[1] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *CVPR*, pages 9686–9696, 2023. 11

[2] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *ICCV*, pages 9921–9931, 2023. 12

[3] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 11

[4] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 2023. 11

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 11

[6] Yuhang Huang, Zheng Qin, Xinwang Liu, and Kai Xu. De-

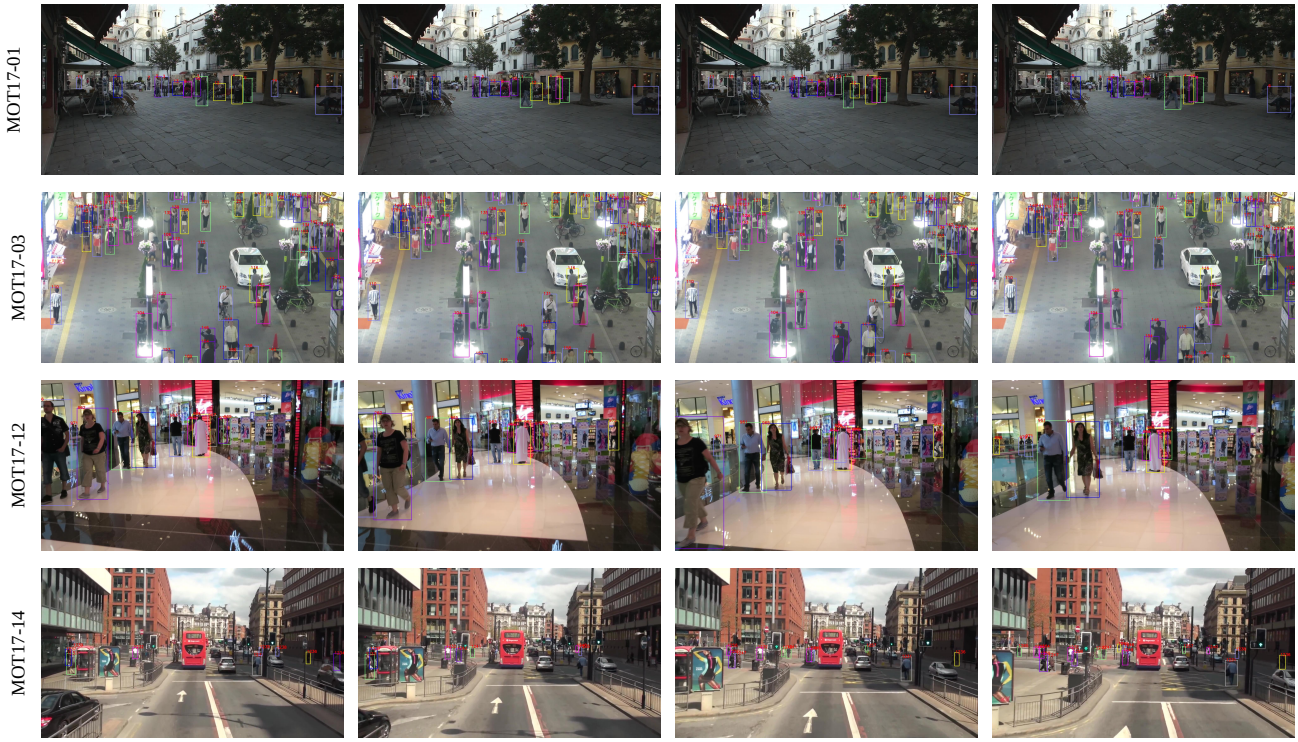


Figure 4. The visualization of DiffMOT tracking results on the test set of MOT17. Boxes of the same color represent the same ID. Best viewed in color and zoom-in.

Figure 5. The visualization of DiffMOT tracking results on the test set of MOT20. Boxes of the same color represent the same ID. Best viewed in color and zoom-in.

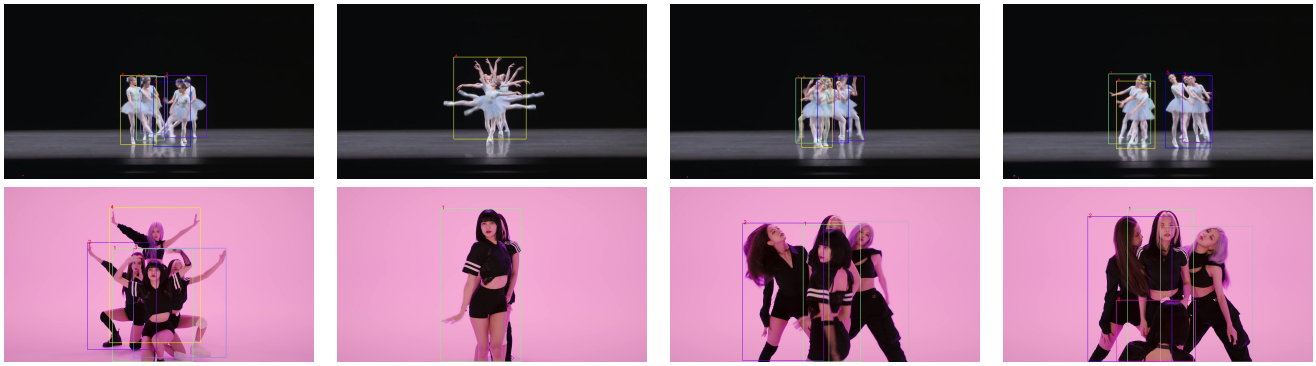


Figure 6. Illustration of two failure cases. We show the two most common failure cases of our approach. In the first row, due to the absence of velocity direction constraints, ID switches have occurred. In the second row, due to the difficulty in recovering long-term lost objects, the new ID is generated.

coupled diffusion models with explicit transition probability. *arXiv preprint arXiv:2306.13720*, 2023. 11

[7] Zelin Liu, Xinggang Wang, Cheng Wang, Wenyu Liu, and Xiang Bai. Sparsetrack: Multi-object tracking by performing scene decomposition based on pseudo-depth. *arXiv preprint arXiv:2306.05238*, 2023. 11

[8] Run Luo, Zikai Song, Lintao Ma, Jinlin Wei, Wei Yang, and Min Yang. Diffusiontrack: Diffusion model for multi-object tracking. *arXiv preprint arXiv:2308.09905*, 2023. 11

[9] Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification. *arXiv preprint arXiv:2302.11813*, 2023. 11

[10] Changcheng Xiao, Qiong Cao, Yujie Zhong, Long Lan, Xiang Zhang, Huayue Cai, Zhigang Luo, and Dacheng Tao. Motiontrack: Learning motion predictor for multiple object tracking. *arXiv preprint arXiv:2306.02585*, 2023. 11

[11] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, pages 1–21. Springer, 2022. 11