MotionTrack: Learning Robust Short-term and Long-term Motions for Multi-Object Tracking - Supplementary Materials -

Zheng Qin[1†] Sanping Zhou[1†] Le Wang[1*] Jinghai Duan[2] Gang Hua[3] Wei Tang[4]
[1]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center for Visual Information and Applications,
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2]School of Software Engineering, Xi'an Jiaotong University
[3]Wormpex AI Research [4]University of Illinois at Chicago

Appendix

In this section, we provide more experimental results in this supplementary material to further evaluate the effectiveness of our proposed method. Besides, we provide architecture and training details, video demos and pseudo code to exhibit our method in more detail.

A. Architecture and Training Details.

Interaction Module. The embedding dimensions of self-attention and graph convolution are both 64. In interaction extraction, we use multi-head self-attention with 1 layer, and the number of heads is 4. We use 3 asymmetric convolution layers with kernel size 3. The threshold used by the signal function is set to 0.6. In motion prediction, the GCN includes 1 layer. Our proposed Interaction Module is trained using the Adam optimizer [1] for 100 epochs with data batches of size 256. The initial learning rate is set to 0.001, which is decayed to 0.0001 after 50 epochs. We take 3 consecutive frames as a training sample. The offset between the first 2 frames is taken as input, and the last frame is used as supervision. We generate input and ground truth after compensating for these 3 images when considering camera motion compensation.

Refind Module. We use 4 asymmetric convolution layers with kernel size 5. The embedding dimensions of tracklet and detection features are both 256. Our proposed Refind Module is trained using the Adam optimizer for 10 epochs with data batches of size 64. The initial learning rate is set to 0.001. When sampling, we limit the tracklets to be longer than 20 frames and the interval between tracklet and detection in a positive sample is from 10 to 120 frames.

†Co-first authors. *Corresponding author.

B. Video Demos.

We provide video demos, where different colored boxes represent different identities and red bolded boxes represent the location during occlusion after the long-term refind. For the cases in our demo video (red bolded), almost all other methods fail to track them (tracklets before and after crowds or occlusion have different identities). https://youtube.com/shorts/CFmHsWB_Sus

C. Algorithm.

In Algorithm 1, we provide the pseudo code of our MotionTrack. Following our description in Section 3.2 of the main paper, given the detections $\mathcal{D}^t$ and current tracklets $\mathbb{T}$, this algorithm demonstrates how our approach updates tracklets at each time-step.

D. Generalization to other Settings.

TAO: It is a *multi-category* tracking dataset with 833 categories. Our MotionTrack achieves 19.6% mAP50 on the validation set, and recent methods SORT_TAO (ECCV'20), QDTrack (CVPR'21), and GTR (CVPR'22) achieve 13.2%, 16.1%, and 22.5% mAP50, respectively. While our method does not perform as well as GTR on TAO, we would note that GTR is built on a more powerful Transformer backbone and it was optimized on this dataset (for example, its performance on MOT17 is much lower than ours, i.e., -5.8 MOTA and -8.6 IDF1). We directly applied MotionTrack to this dataset with minimal modification, despite that it contains 833 categories rather than 1 person category and its videos have a much lower frame rate than MOT. Considering these significant differences, we believe that our method could be further improved in the field of multi-category tracking.

Setting	MOTA ↑	IDF1 ↑	HOTA ↑	AssA ↑	DetA ↑	IDs ↓
Linear interpolation	80.7	83.2	70.6	73.0	68.9	327
Error compensation	80.8	83.3	70.7	73.0	69.0	327
Improvement	**+0.1**	**+0.1**	**+0.1**	-	**+0.1**	-

Table 1. Comparison between linear interpolation and error compensation on the MOT17 validation set. Increases in metrics are marked in green.

E. Future work.

In the future, we will explore two directions to model the interaction between the Interaction Module and Refind Module. In the first direction, we will design a more integrated model of the two modules so that they could mutually benefit each other. On the one hand, the history trajectory information provides more reliance for interaction extraction and motion prediction. On the other hand, considering short-term motion also helps with long-term motion pattern learning. In the second direction, we will input the detections of the current frame into Interaction Module, concatenate them with the tracklet's features, and construct one more branch to predict the probability that a detection is unmatched. Then, the unmatched detection sampled for training Refind Module can be obtained more accurately and more reasonably than random sampling in the training phase. In the inference phase, we will set alive tracklets and lost tracklets to the same status when matching, which can break through the serial structure of the tracking framework more elegantly.

F. Speed.

As an appearance-free tracker, both our proposed modules (Interaction Module and Refind Module) take coordinate sequences as input instead of images and the architectures of our modules are very simple. On MOT17 and MOT20, our method achieves 15.7 FPS and 11.5 FPS, compared with 15.1 FPS and 10.8 FPS of the baseline, causing negligible additional computational burden.

G. Effect of Error Compensation.

After re-identifying the lost tracklet with its matched detection, we need to fill the trajectory. To make this process more accurate and reasonable, we correct the predicted trajectory instead of generating a new one. Compared with linear interpolation [2], which is subject to linearity, our error compensation method provides a post-processing trick to further improve performance without additional computational overhead. Table 1 shows the effectiveness of our error compensation and proves that Interaction Module and Refind Module still provide the major contributions combined with Table 3 in the main paper.

Algorithm 1: Data association of our MotionTrack.

Input: Tracklets at time t-1 $\mathbb{T}$; detections at time t $\mathcal{D}^t$; detection score thresholds τ^{high}, τ^{low}

Output: Updated tracklets at time t $\mathbb{T}$

/* Filter out high and low scoring detections. */

$\mathcal{D}^{\text{high}}, \mathcal{D}^{\text{low}} \leftarrow \text{Separation}(\mathcal{D}^t, \tau^{\text{high}}, \tau^{\text{low}})$

/* **Short-range association.** */

/* Apply camera motion compensation to tracklets, then extract the offset and absolute coordinates as input. */

$\mathbf{I}^t \leftarrow \text{Seq2offset}(\mathbb{T})$

$\mathbf{A}^{\text{adjc}} \leftarrow \text{InteractionExtraction}(\mathbf{I}^t)$

$\mathbf{P}^{\text{offs}} \leftarrow \text{MotionPrediction}(\mathbf{A}^{\text{adjc}}, \mathbf{O}^t)$

/* Assignment based on IoU distance. */

asgn1, $\mathcal{D}^{\text{u1}}$, $\mathbb{T}^{\text{u1}} \leftarrow \text{IoUAssociation}(\mathbb{T}, \mathcal{D}^{\text{high}}, \mathbf{P}^{\text{offs}})$

asgn2, $\mathcal{D}^{\text{u2}}$, $\mathbb{T}^{\text{u2}} \leftarrow \text{IoUAssociation}(\mathbb{T}^{\text{u1}}, \mathcal{D}^{\text{low}}, \mathbf{P}^{\text{offs}})$

/* Divide tracklets into alive part and lost part. */

$\mathbb{T}^{\text{alive}}, \mathbb{T}^{\text{lost}} \leftarrow \text{Division}(\mathbb{T})$

Update($\mathbb{T}^{\text{alive}}$, asgn1, asgn2)

Record($\mathbb{T}^{\text{lost}}$, $\mathbf{P}^{\text{offs}}$)

/* **Long-range association.** */

$\mathbf{D}^{\text{rest}}, \mathbf{T}^{\text{lost}} \leftarrow \mathcal{D}^{\text{u1}}, \mathbb{T}^{\text{lost}}$

$\mathbf{C}^{\text{corr}} \leftarrow \text{CorrelationCalculation}(\mathbf{D}^{\text{rest}}, \mathbf{T}^{\text{lost}})$

/* Assignment based on correlation matrix. */

asgn3, $\mathcal{D}^{\text{u3}}$, $\mathbb{T}^{\text{u3}} \leftarrow \text{Greedy}(\mathbb{T}^{\text{lost}}, \mathcal{D}^{\text{u1}}, \mathbf{C}^{\text{corr}})$

ErrorCompensation($\mathbb{T}^{\text{lost}}$, asgn3)

Initialization($\mathcal{D}^{\text{u3}}$)

H. Analysis of Correlation Calculation.

Following our description in Section 3.4 of the main paper, we claim that the long-range motions based on the history trajectory is more discriminative and can better match the tracklets and detections in case of extreme occlusion. Therefore, we follow the experimental settings in Section 3.5 of the main paper and intuitively verify that the correlation accuracy of 97% and 99% are achieved on MOT17 and MOT20 respectively, which proves the reliability of long-range motions.

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[2] 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, 2022. 2