Appendices for Integrating Boxes and Masks: A Multi-Object Framework for Unified Visual Tracking and Segmentation

Yuanyou Xu[1,2]†, Zongxin Yang[1], Yi Yang[1]‡
[1]ReLER, CCAI, Zhejiang University, China [2]Baidu Research, China
{yoxu, zongxinyang, yangyics}@zju.edu.cn

A. Spatial-Temporal Propagation

A.1. Paradigm Comparison

Recent box-based tracking methods [2, 17] with leading performance use the one-stage paradigm with large-scale transformers [12, 4] pre-trained on ImageNet [3]. As a unified framework, we use the two-stage paradigm, but we further perform spatial-temporal propagation across multiple memory frames, which has been proven to be effective in video object segmentation [8, 1, 15]. Compared with common tracking methods (Figure 1), the memory propagation on whole frames has larger coverage on both temporal and spatial dimensions.

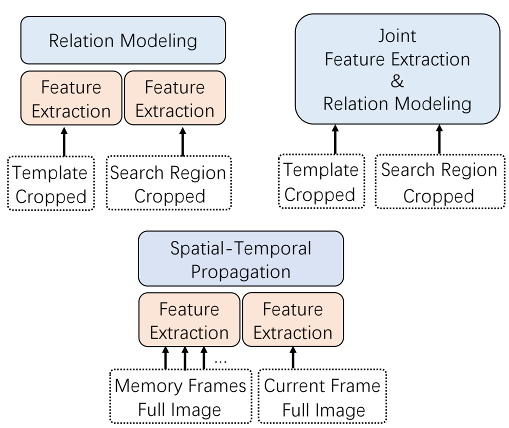

Figure 1. Comparison between two-stage (up left), one-stage (up right) tracking paradigms [13, 17] and memory based propagation paradigm [8, 15] (down middle) in our framework.

A.2. Memory Strategy

We use previous frames together with their predicted masks to update the memory storage by extending the stored keys and values for attention operations in the propagation module. For VOT task, we still predict masks for each frame for memory updating, as for VOS task. The memory storage is used in two types of attention in the propagation module, the long-term attention and the short-term attention [15, 16]. The long-term attention is between the current frame and all memory frames and is performed in a global manner. The short-term attention is performed between the current frame and a previous frame in a local window. In practice, we update the long-term memory every T_l frame, and set a max memory capacity of $T_{max} = 10$ frames to avoid memory explosion, which is very necessary in long-term tracking. For short-term memory, we always select the T_s-th frame before the current frame. Results of different long and short memory gaps are shown in Table 1. We find optimal memory gaps are relatively stable across similar benchmarks like LaSOT [5] and TrackingNet [7], as $T_l = 30, T_s = 10$. GOT-10k [6] is special because the

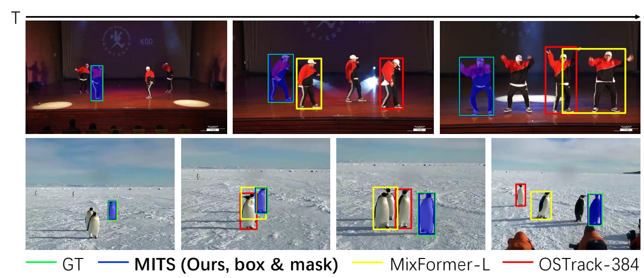

Figure 2. Qualitative results of MITS on VOT, compared with SOTA SOT methods MixFormer [2] and OSTrack [17].

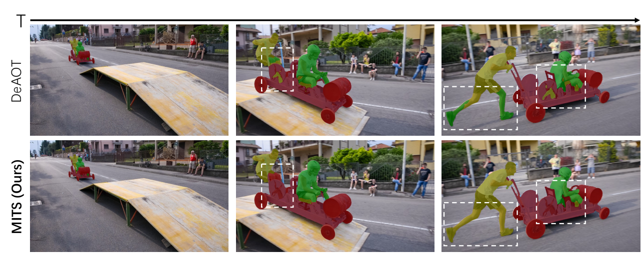

Figure 3. Qualitative results of MITS on VOS, compared with SOTA VOS method DeAOT [16]. For MITS, predictions from box predictor and mask decoder are both visualized.

†Yuanyou Xu worked on this at his Baidu Research internship.
‡Yi Yang is the corresponding author.

T_l	T_s	LaSOT [5] AUC	LaSOT [5] P_N	LaSOT [5] P	TrackingNet [7] AUC	TrackingNet [7] P_N	TrackingNet [7] P
40	13	71.8	79.7	77.8	83.2	88.6	84.1
30	10	**72.0**	**80.1**	**78.5**	**83.4**	**88.9**	**84.6**
20	6	70.5	77.9	76.5	83.2	88.8	84.3

Table 1. Results of different long T_l and short T_s memory gaps on SOT benchmarks.

videos in it are at 10 FPS, and we set $T_l = 10, T_s = 2$. For VOS benchmarks YouTube-VOS [14] and DAVIS [10], we set $T_l = 10, T_s = 3$.

B. Loss and Optimization

The loss functions we use for mask branch are Cross Entropy loss and Jaccard loss, and for box branch we use L1 loss and Generailized IoU loss [11]. For the ID decoder, we only use Cross Entropy loss to supervise the reconstruction of masks. Finally, we get the total loss:

$$L_{mask} = \frac{1}{N}\sum_{i=0}^{N}(-Y_i \log(\hat{P}_i) + \lambda_m(1 - IoU(\hat{Y}_i, Y_i)) \tag{1}$$

$$L_{box} = \frac{1}{N}\sum_{i=0}^{N}(|B_i - \hat{B}_i| + \lambda_b(1 - GIoU(\hat{B}_i, B_i)) \tag{2}$$

$$L_{ID} = \frac{1}{N}\sum_{i=0}^{N}(-Y_i \log(\tilde{P}_i)) \tag{3}$$

$$L_{total} = \alpha L_{mask} + \beta L_{box} + \gamma L_{ID} \tag{4}$$

where $\hat{P}$, $\tilde{P}$ and $\hat{B}$ are predictions from mask decoder, ID decoder and box predictor for N objects respectively, and Y and B are the ground truth one-hot masks and boxes for N objects. The losses are averaged among all objects. We set $\lambda_m = 1, \lambda_b = 0.5, \alpha = 0.5, \beta = 0.2, \gamma = 1$.

C. More Visualization Results

More visualization results are in Figure 2 and 3. In Figure 2, we select challenging scenes with multiple similar objects from LaSOT test set [5]. Advanced SOT methods like OSTrack [17] and MixFormer [2] fail to track the target object under the complex interaction among multiple similar objects with occlusion. Compared with them, MITS with unified multi-object identification has learned how to deal with multiple similar objects during training, so it is able to identify and track the target robustly from similar objects. In Figure 3, we compare MITS with SOTA VOS method DeAOT [16] in a challenging scene with multiple moving objects in DAVIS [9, 10]. More visualized video clips are available in the supplementary video.

References

[1] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 1

[2] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13608–13618, 2022. 1, 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[5] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 1, 2

[6] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019. 1

[7] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pages 300–317, 2018. 1, 2

[8] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019. 1

[9] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 2

[10] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 2

[11] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 2

[12] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing

convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021. 1

[13] Fei Xie, Chunyu Wang, Guangting Wang, Yue Cao, Wankou Yang, and Wenjun Zeng. Correlation-aware deep tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8751–8760, 2022. 1

[14] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 2

[15] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[16] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. *Advances in Neural Information Processing Systems*, 2022. 1, 2

[17] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 341–357. Springer, 2022. 1, 2