LVOS: A Benchmark for Long-term Video Object Segmentation
Supplementary Materials

1. Dataset Construction and Annotations

For the class selection, we summarize the classes of LaSOT [6] and VOT-LT 2019 [9]. There are 85 classes in LaSOT and about 20 classes in VOT-LT 2019. Then we carefully select a set of categories based on the following rules: (1) The resolution of videos are larger than 720p, (2) The video is representative enough to include at least one attribute demonstrated in Table 2 of paper. (3) Class of the video is relative to daily life. (4) The total number of videos with this category should be greater than ten. Based on the four rules, we choose 27 categories. Because VOT-LT and LaSOT are single-object tracking dataset, LVOS is a mutliple-object. For target selection, we may follow the target object in VOT-LT and LaSOT, or select different objects as targets.

For the annotation process, because all the masks are obtained by models, we need two-pass manual corrections. During Step 1 1 FPS automatic sgmentation, we utilize the box of target object in each frame to get segmentation. If the target object of a video is the same as that in LaSOT or VOT-LT, we use the gropundtruth boxes. Otherwise, we adopt tracking model to obtain the box of target object in each frame.

2. Training Strategy

Following [20, 10], we divide the training stage into two phases: (1) pretraining on static image datasets [4, 5, 12, 17, 7] by applying data augmentation such as synthetic deformation with the initial learning rate of 4×10^{-4} and a weight decay of 0.03 for 100,100 steps. (2) main training on the VOS datasets [15, 18] with the initial learning rate of 2×10^{-4} and a weight decay of 0.07 for 100,100 steps. AdamW [13] optimizer is adopted for optimization. The batch size is set as 16. Dice loss [14] and bootstrapped cross entropy loss with equal weighting is used.

Methods	Backbone	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	FPS
CFBI[19]	ResNet101[8]	81.9	79.1	84.6	5.9
LWL[1]	ResNet50[8]	81.6	79.1	84.1	13.2
STCN[3]	ResNet50[8]	<u>85.4</u>	<u>82.2</u>	<u>88.6</u>	20.2
RDE[10]	ResNet50[8]	84.2	80.8	87.5	27.0
XMem[2]	ResNet50[8]	**86.2**	**82.9**	**89.5**	22.6
AOT-B[20]	MobileNet-V2[16]	82.5	79.7	85.2	**29.6**
AOT-L[20]	MobileNet-V2[16]	83.8	81.1	86.4	18.7
DDMemory	MobileNet-V2[16]	84.2	81.3	87.1	<u>28.1</u>

Table 1: Comparisons with state-of-the art models on DAVIS 2017 validation set[15]. Bold and underline denote the best and second-best respectively in each column.

3. Results on Short-term Videos Validation Sets

We compare our DDMemory with state-of-the-art VOS models on short-term videos validation datasets (DAVIS 2017 [15] and YouTube-VOS 2018 [18]) in Table 1 and 2. We re-time these models on our hardware (one V100 GPU) for a fair comparison. DDMemory exceeds the majority of models and maintains an efficient speed. Despite having higher performance than DDMemory, XMem and STCN employ a stronger backbone ResNet50 [8], while DDMemory only uses MobileNet-V2 [16]. Although the segmentation accuracy in short-term videos can be improved by the global temporal information, but the improvement on short-term videos validation sets is not very obvious. The reason may be that the length of the videos is relatively short.

Methods	Backbone	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	FPS
CFBI[19]	ResNet101[8]	81.4	81.1	85.8	75.3	83.4	4.0
LWL[1]	ResNet50[8]	81.5	80.4	84.9	76.4	84.4	-
STCN[3]	ResNet50[8]	83.0	81.9	86.5	77.9	85.7	13.2
RDE[10]	ResNet50[8]	81.9	81.1	85.5	84,8	76.2	17.7
XMem[2]	ResNet50[8]	**85.7**	**84.6**	**89.3**	**80.2**	**88.7**	11.8
AOT-B[20]	MobileNet-V2[16]	83.5	82.6	87.5	77.7	86.0	**20.5**
AOT-L[20]	MobileNet-V2[16]	83.8	82.9	87.9	77.7	86.5	16.0
DDMemory	MobileNet-V2[16]	84.1	83.5	88.4	78.1	86.5	18.7

Table 2: Comparisons with state-of-the art models on YouTubeVOS-2018 validation set [18]. Bold and underline denote the best and second-best respectively in each column.

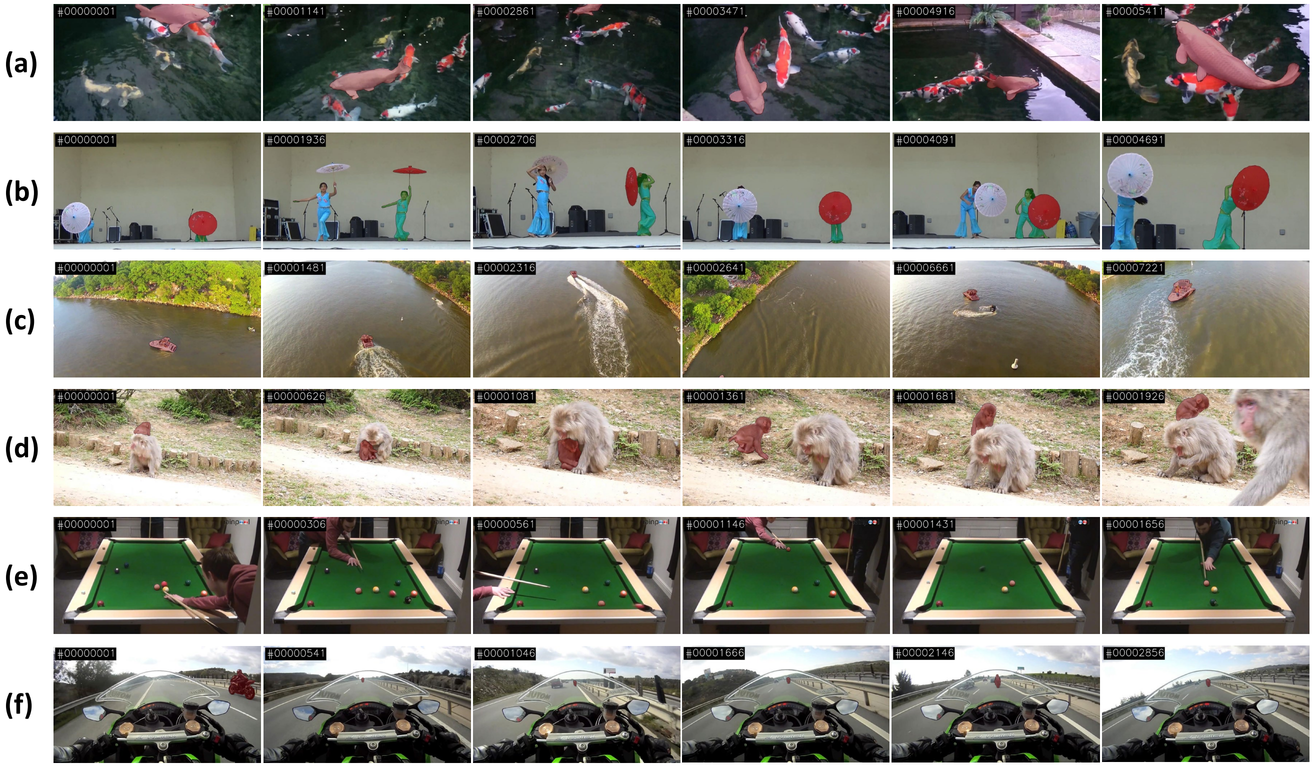

Figure 1: Qualitative results on LVOS validation and test set. DDMemory performs well on long-term videos.

4. Additional Qualitative Results

We show more qualitative results on LVOS in Figure 1. As demonstrated, our DDMemory can handle many challenging long-term VOS attributes, such as long-term reappearance, similar objects, occlusion, fast and complex occlusions, low resolution, and scale variation, etc. In row (a), DDMemory successfully distinguishes the white goldfish with other similar fishes. In row (b), the two people and umbrellas are not confused with each other in spite of occlusion. In row (c), DDMemory can re-detect the boat after long-term and frequent disappearance. In row (d), the small white ball is similar to other balls, and DDMemory still succeeds in tracking and segmenting it. In row (f), DDMemory tracks the motorcycle well despite the fast motion and large scale variation.

R	G	L	FPS	GPU	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
✓			57.4	0.52	44.2	39.0	49.4
	✓		55.2	0.62	42.7	37.4	48.0
		✓	43.5	0.68	18.3	17.1	19.6
✓	✓		46.7	0.78	47.8	42.4	53.3
✓		✓	35.6	0.82	57.9	53.0	62.8
	✓	✓	35.1	0.76	54.9	51.1	58.7
✓	✓	✓	30.3	0.88	61.9	56.3	67.4

Table 3: Ablation study on LVOS validation set. R, G, and L denote Mem_R, Mem_G, and Mem_L, respectively.

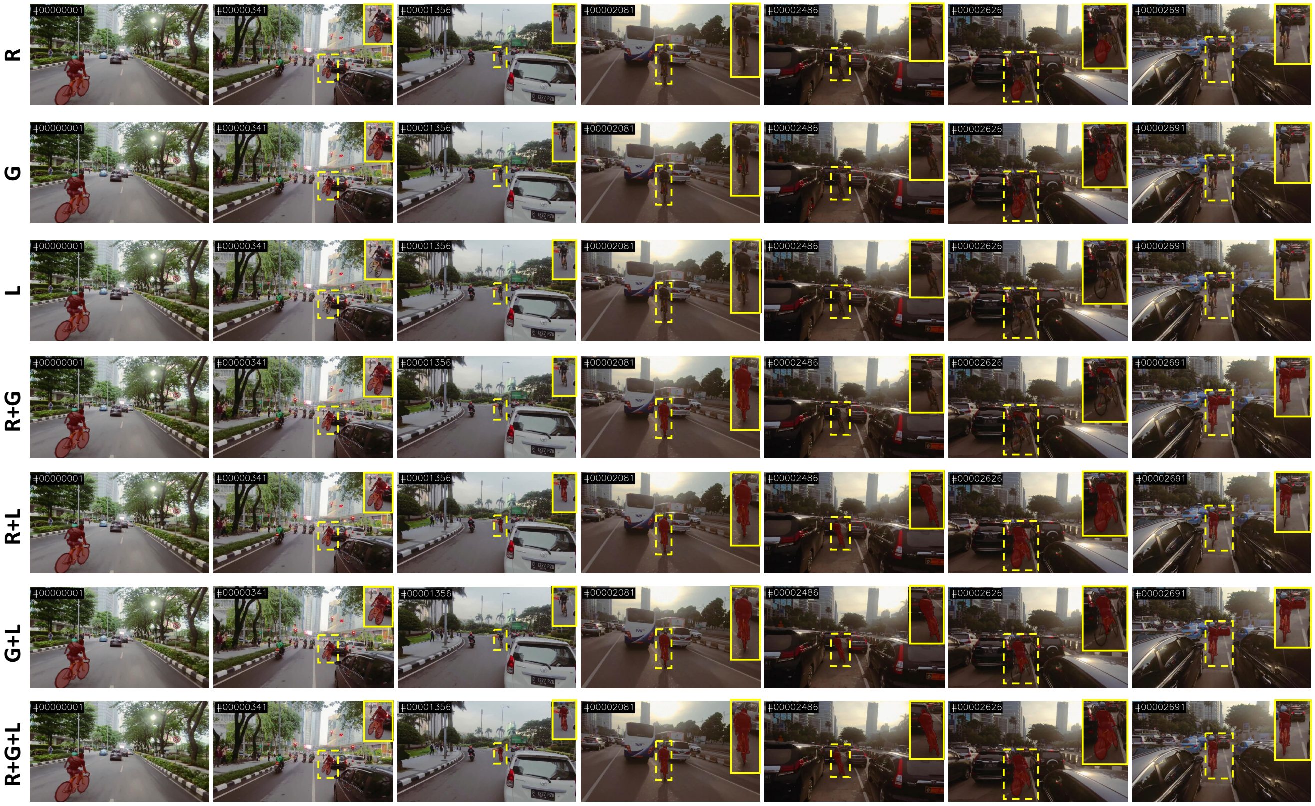

Figure 2: Ablation study. We visualize the results of different combinations of three memory banks on the same video. Best viewed in color.

5. More Analysis About Diverse Dynamic Memory

We conduct an ablation study on the role of each memory bank in Table 3. To more clearly illustrate each memory bank's impact, we visualize the results of different combinations of three memory banks on the same video in Figure 2. This video is about a man riding a bike through the streets. The man frequently gets occluded by cars. Moreover, there are many other challenges in this video, such as background clutter (there are many similar people on the street), fast motion (the man is moving quickly), low resolution (sometimes the bounding-box of this man is small), and significant appearance change (the

Attr	AFB-URR [11]	RDE [10]	CFBI [19]	AOT-B [20]	AOT-L [20]	STCN [3]	XMem [2]	LWL [1]	Ora B	Ora M	Ora B+M
FM	34.1	48.4	45.3	54.3	55.3	42.5	46.7	48.2	73.8	82.6	85.5
OCC	34.5	48.2	46.1	50.6	52.1	43.3	47.6	50.3	72.5	79.1	83.6
OV	42.2	53.4	47.6	54.4	55.2	51.5	53.8	48.6	74.2	79.9	82.8
SV	33.1	48.4	45.7	48.3	50.4	41.5	43.7	47.2	66.8	76.6	80.5
AC	41.6	51.9	45.9	53.1	55.7	48.2	48.9	52.4	75.7	82.2	84.2
LRA	33.1	41.4	39.9	44.3	45.3	37.5	40.7	45.2	63.8	74.6	78.5
CTC	36.9	41.6	40.4	44.5	45.7	39.7	45.1	46.1	64.4	75.5	77.7

Table 4: Attribute-based aggregate performance. For each method, we just show $\mathcal{J}$. Ora B, Ora M, Ora B+M denote oracle box, oracle mask and oracle box + mask in oracle experiments, respectively.

appearance of this man changes a lot over the time). This video is extremely challenging. In the first row, we just use the reference memory Mem_R, despite the re-detection after occlusion, the reference memory is sensitive to large appearance changes. In the second row, only global memory Mem_G is enabled. Global memory is rich in temporal information so Mem_G can handle occlusion, too. Because of the error accumulation, there are still many segmentation defects. In the third row, only local memory Mem_L is used, and it is obvious that the model loses track after the first occlusion. In the fourth row, we utilize the reference and global memory. Although the model is better at handling changes in appearance, it still has trouble precisely segmenting the target. In the fifth row, we combine reference and local memory. The local memory boosts the contour accuracy to a large descent. In the sixth row, global memory and local memory banks are used. Compared to the fifth raw, the segmentation accuracy is a little worse. In the final row, we combine the three complementary memory banks. DDMemory tracks and segments target objects successfully. The visual results demonstrate the role of the three memory banks. The reference memory Mem_R is responsible for the re-detection after occlusion or out-of-view and is sensitive to large appearance changes. The local memory Mem_L provides location cues and appearance prior. The global memory Mem_G encodes the long-term temporal information as a complement to the other two memory features. For long-term VOS, all three memory banks are essential and complementary.

6. Oracle Experiments

For oracle box, we convert groundtruth mask into box and only search target in the groundtruth box area. For oracle mask, we search target in whole image and use groundtruth mask to update Mem_G and Mem_L. For oracle box and mask, we search target in the groundtruth box area and use groundtruth mask to update Mem_G and Mem_L

7. Attribute-based Evaluation

We report performance of more models in Table 4 on validation set characterized by the most informative attributes. Scale variation has a more pronounced negative impact on short-term visual object segmentation (VOS) performance than other challenges, particularly for models that employ online adaption (OD) or compressed memory (C) feature banks. Additionally, specific long-term challenges have an even greater impact on accuracy. Visual object segmentation (VOS) models may lose track of the target object when it becomes small in size. Models that always keep the first frame in memory can re-detect the target object. However, models that employ online adaption (OD) or compressed memory (C) feature banks may mistake background objects for the target object, or they may be unable to restore detection due to the lack of guidance from the first frame. Therefore, the ability to recover a disappeared object, distinguish the target object from similar background objects, detect small objects, and model long-term historical information is crucial for robust LVOS.

References

[1] Goutam Bhat, Felix Järemo Lawin, Martin Danelljan, Andreas Robinson, Michael Felsberg, Luc Van Gool, and Radu Timofte. Learning what to learn for video object segmentation. In *European Conference on Computer Vision*, pages 777–794. Springer, 2020.

[2] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. *arXiv preprint arXiv:2207.07115*, 2022.

[3] 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.

[4] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582, 2014.

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[6] 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.

[7] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, pages 991–998. IEEE, 2011.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kamarainen, Luka ˇCehovin Zajc, Ondrej Drbohlav, Alan Lukezic, Amanda Berg, et al. The seventh visual object tracking vot2019 challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[10] Mingxing Li, Li Hu, Zhiwei Xiong, Bang Zhang, Pan Pan, and Dong Liu. Recurrent dynamic embedding for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1332–1341, 2022.

[11] Yongqing Liang, Xin Li, Navid Jafari, and Jim Chen. Video object segmentation with adaptive feature bank and uncertain-region refinement. *Advances in Neural Information Processing Systems*, 33:3430–3441, 2020.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[14] Sebastian Nowozin. Optimal decisions from probabilistic models: the intersection-over-union case. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 548–555, 2014.

[15] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2663–2672, 2017.

[16] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[17] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):717–729, 2015.

[18] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 585–601, 2018.

[19] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *European Conference on Computer Vision*, pages 332–348. Springer, 2020.

[20] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34:2491–2502, 2021.