Class-Agnostic-Pairwise-Affinity for Semi-supervised Video Object Segmentation

Hui Zhou[1] *, Yong Liu[1,2] *, Jisheng Dang[3], Huicheng Zhen[3], Qiang Zhou[2],
Shanglin Li[1,2], Yixing Zhu[2], Yansong Tang[1], Yujiu Yang[1] †

[1]Shenzhen International Graduate School, Tsinghua University [2]ByteDance Inc.
[3]School of Computer Science and Engineering, Sun Yat-sen University.
{zhouh21,liu-yong20}@mails.tsinghua.edu.cn,
{yang.yujiu}@sz.tsinghua.edu.cn

Abstract

In this paper, we introduce Class-Agnostic-Pairwise-Affinity (CAPA) for semi-supervised video object segmentation. STM-based methods achieved superior performance in this field, but were still difficult to distinguish similar objects due to pixel matching mechanism. Different from previous STM-based methods which mostly rely on affinity between the query frame and past frames, the CAPA module introduced in this paper is based on pairwise affinity prediction in the query frame. The module can separate similar objects even when they are close to each other. Through model ensembling and multi-scale testing, we achieve outstanding performance on the YouTube-VOS 2022 Challenge with J&F score of 85.7%.

1. Introduction

Video object segmentation (VOS) has been received extensive attention with its widespread application in video editing, autonomous driving, etc.

Among all the VOS scenarios, semi-supervised video object segmentation (Semi-VOS) is widely researched and also play a important role in other VOS tasks. [1] In Semi-supervised Video Object Segmentation, the mask of one or multiple objects of interest are given in the first frame of a video, and the algorithm should produce the segmentation masks of these objects in the subsequent frames.

There are several streams of literature studying the Semi-supervised Video Object Segmentation task. [1] Matching-based methods which calculate the pixel-level similarity between the query frame and the past frames dominate this field. With the guidance of middle frames, Space-Time Memory Network (STM) [3] shows effectiveness in adapting to changes in the appearance of targets. The Semi-VOS task has made great progress since STM was proposed. Yet origin STM suffers from huge memory consumption and heavy computation. What's worse, post-processing is also required in multi-target scenarios since STM could only process a single target at once. Subsequently, AOT [7], [8]improved the multi-target processing paradigm by associating multi objects simultaneously with transformer and identification mechanism. And more accurate segmentation has been achieved with its hierarchical matching.

However, the scenarios where there are several objects similar and close to each other were still hard to process. For instance, in Fig. 4, there are two tigers not only very similar in appearance but also close to each other. Suppose we are interested in the right one. In the next frame, another tiger appears in the position of the target and the target is occluded. STM-based methods (including AOT) showed poor performance to distinguish the two tigers due to they were solely based on appearance even with short-term attention.

To remedy this, our intuition is to distinguish multiple objects based on the current frame, rather than just matching them according to previous frames. A natural solution might be to use classic detection or instance segmentation algorithms since they can distinguish different instances. Regrettably, most of these methods are limited to training classes. Solving unseen categories in the YouTube-VOS challenge is often intractable. So we couldn't directly apply them to our task. Inspired by [5], we propose a Class-Agnostic-Pairwise-Module (CAPA) that predicts whether adjacent points on the feature map belong to the same instance to provide auxiliary information for distinguishing between instances. As a result, our method can effectively suppress the error in the scenarios with multiple objects similar and close to each other.

*Equal contribution.

†Yujiu Yang is the corresponding author.

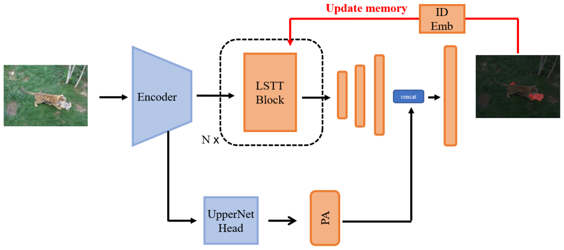


Figure 1. The overview of our framework, notice that the LSTT block repeat several times

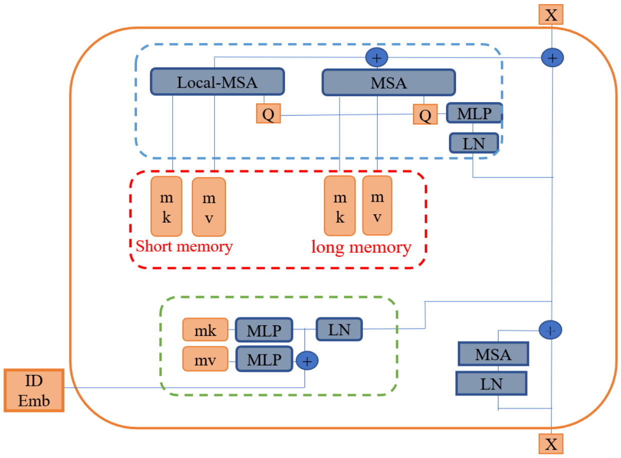


Figure 2. LSTT block. MSA,LN,MLP denote multi-head self attention, layer normalization. and a single linear layer respectively. The red, blue and green dotted box indicate memory bank, memory readding process and memory updating process. Local-MSA is the short-term attention from [8]

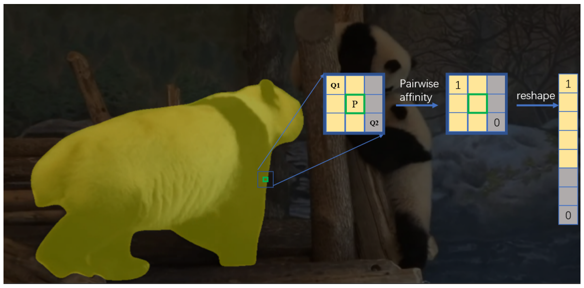


Figure 3. Pairwise Prediction

2. Method

The existing mechanisms can not effectively distinguish instances in the scene with several objects similar and close to each other. To address this problem, we should distinguish multiple objects based on the current frame, instead of matching them according to previous frames. Inspired by [5], we propose a Class-Agnostic-Pairwise-Module (CAPA). CAPA predicts whether adjacent points on the feature map belong to the same instance, thus providing auxiliary information for distinguishing instances.

Given a particular frame, the goal of CAPA is to predict whether adjacent pixels in this frame belong to the same instance. Note the frame as $I \in R^{3\times H\times W}$. For pixel $P = (i, j)$ in the frame I, and its adjacent points $Q \in \{(i+1, j), (i+1, j+1), (i, j+1), (i, j-1), (i-1, j), (i-1, j-1), (i-1, j+1), (i+1, j-1)\}$, the target output value of the CAPA module is 1 if P and Q belong to the same instance, otherwise 0.

Take Fig. 3 as example, the center pixel (which is illustrated in the green boundary box) of the blue 3x3 block is P, and two of its adjacent pixels are $Q1$ (at the left top corner) and $Q2$ (at the right bottom corner). P and $Q1$ belong to the same instance (the bigger panda), so the prediction result of the corresponding position of $Q1$ should be 1. P and $Q2$ (at the background) do not belong to the same instance, so the corresponding position of $Q2$ is predicted to be 0. After predicting all pixel points, we finally get the result of Pairwise Affinity as $PA \in \{0, 1\}^{8\times H\times W}$.

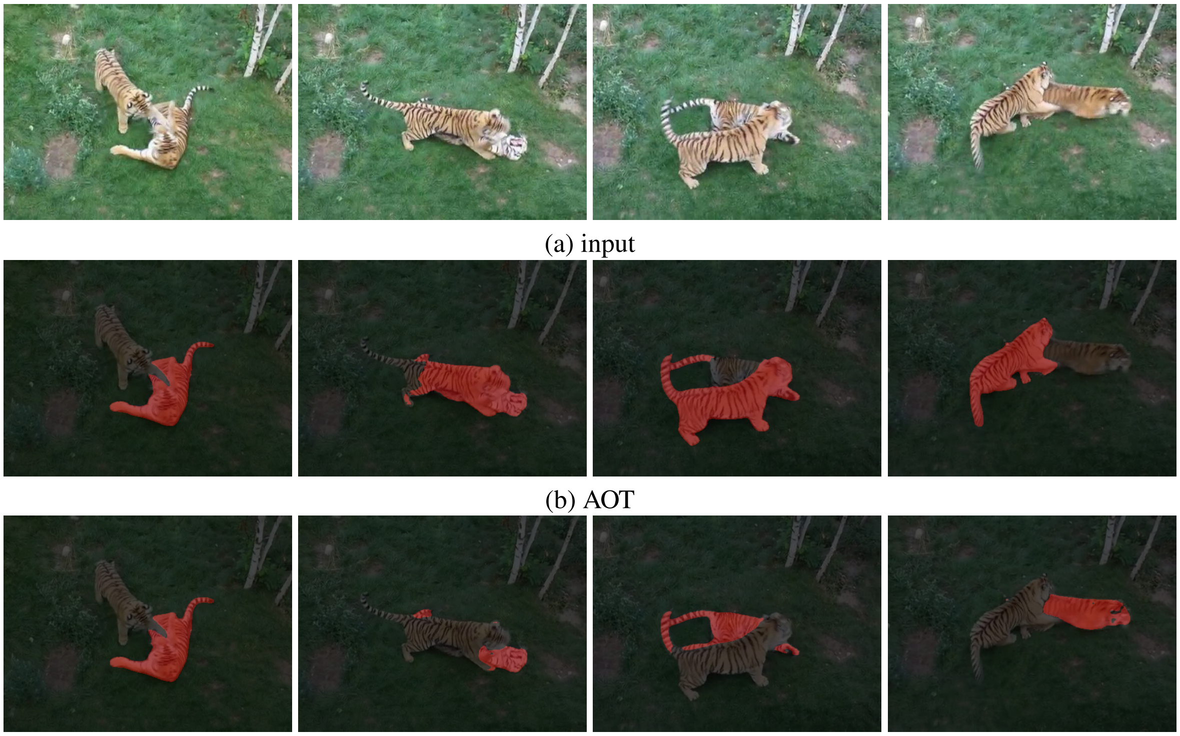
To demonstrate the effectiveness of our CAPA module, we apply our CAPA method to AOT, a state-of-the-art VOS method. Note that CAPA can also be easily used in other STM-based algorithms.

As shown in Fig. 1, the red arrow represents the process of updating memory with a known mask, and the black arrow represents the process of inferring the target mask of the current image based on memory. When the target object appears for the first time, the Identification Mechanism of AOT builds the initial identification feature. Then, the identification feature is used together with the feature of the query image as the input of the multi-layer LSTT layer, and the output of each layer is stored in memory. At the stage of inference, the interaction between the feature of the input frame and the features from memory will be feed through multiple LSTT layers to obtain input feature to the decoder. Finally, the mask of the current frame is obtained from the decoder with the decoder input feature. To use our CAPA method, as shown in Fig. 1, it is easily applied as an additional branch with AOT. In order to add auxiliary information, we feed the feature predicted by CAPA into the AOT decoder. The final result mask is generated by the combination of CAPA feature and the original output feature of multiple LSTT layers.

3. Experiments

3.1. Implementation

To illustrate the effectiveness of our CAPA method, we add the pairwise module to AOTv2 [7]. Swin-B [2] is



(a) input

(b) AOT

(c) ours

Figure 4. Compare our method to AOT.

adopted as our backbone. The CAPA use UperNet [6] as pairwsie affinity head, which is consistent with [5].

We trained CAPA part as [5] and froze the weight of this module in the later training process. For the complete model training, we followed AOT and performed the pre-training on several static image datasets. But we found that models trained solely on the above datasets performed poorly in certain categories. To remedy this, we also performed pre-training on two additional video segmentation datasets which are UVO [5]and OVIS [4]. UVO is an open-world video segmentation dataset containing a large number of new categories. OVIS focuses on multi-object video segmentation in occluded conditions.

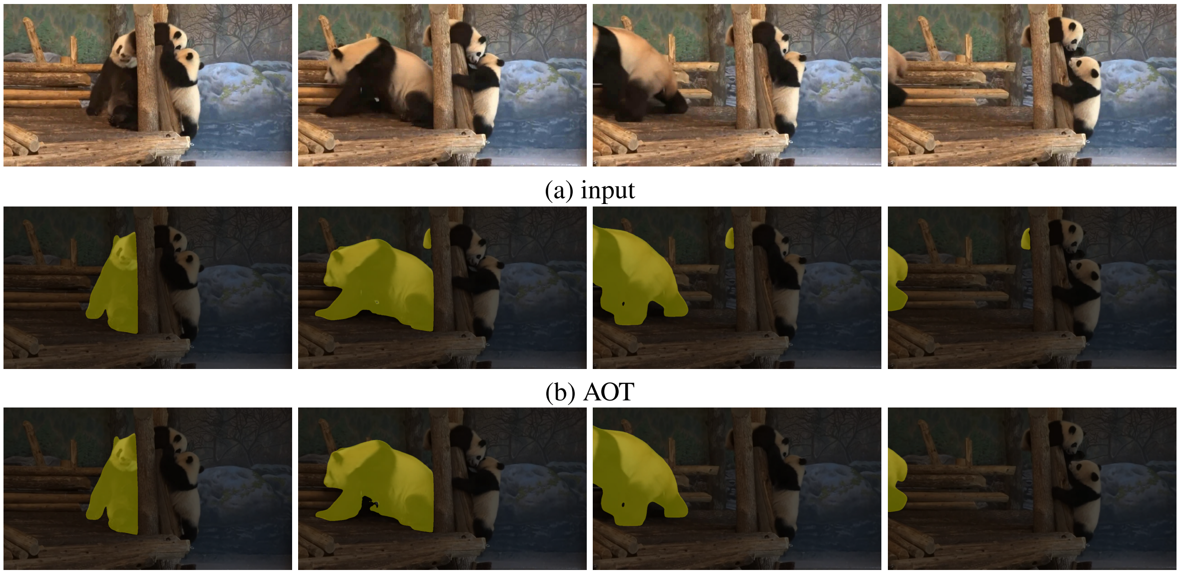
To further boost performance, we also apply model ensembling strategy. The final decision is based on frame-level mask voting mechanisms. In this mechanism, the output mask of each model constitutes candidates. The IoU of candidate M_1 and M_2 is just the voting score of M_1 from M_2 (also the voting score of M_2 from M_1). The sum of the voting score of each candidate counts for the final decision. The candidate with the highest score is chosen. In addition, multi-scale (0.75,1,1.25,1.4) testing and random flip are used to improve model performance.

3.2. Ablation Study

We perform several ablation studies to verify the effectiveness of our CAPA module and extra data in the pre-training. We started with pure AOTv2. And then add UVO and OVIS as additional datasets. Later, we put on CAPA module and finally ensemble these three models. As shown in Tab. 1, without additional datasets and CAPA, AOTv2 got 84.2 on J&F metrics. With additional datasets, the score was up to 85.1. This may be caused by additional datasets for pre-training so that the model can predict better on previously poorly handled categories. After adding the CAPA module on this basis, we observed the model performed much better in many scenarios especially there are several similar object like Figs. 4 and 5. While the overall score dropped slightly to 84.7 probably because the decoder is not the appropriate place for incorporating the output of CAPA module, the CAPA handle difficult scenarios as Figs. 4 and 5. Therefore, model ensembling with CAPA and origin AOTv2 can further improve overall score.

4. Conclusion

Distinguishing between similar targets still is one of the difficult scenes of VOS task, in this paper, by introducing instance auxiliary information, our method eases the problem. Solving complex scenes with several objects similar



(a) input

(b) AOT

(c) ours

Figure 5. Compare our method to AOT.

id	setting	extra data	CAPA	J&F Score
1	AOTv2			84.2
2	AOTv2	✓		85.1
3	AOTv2	✓	✓	84.7
	ensemble 1,2,3			85.7

Table 1. Ablation study

and close to each other is often intractable. Traditional STM methods only focus on similarity matching between frames thus perform poorly in distinguishing similar objects within the same frame. By introducing short-term attention, AOT method alleviates this problem by local space-time continuity for simple motion scenes. In this paper, we introduced a Class-Agnostic-Pairwise-Module (CAPA) module to alleviate the problem from a new perspective. Our module predicts whether adjacent points on the feature map belong to the same instance to provide auxiliary information for distinguishing between instances. Through model ensembling and multi-scale testing, our method achieves 85.7% J&F score on the YouTube-VOS 2022 test dataset.

References

[1] Mingqi Gao, Feng Zheng, James J. Q. Yu, Caifeng Shan, Guiguang Ding, and Jungong Han. Deep learning for video object segmentation: A review. *Artificial Intelligence Review*, Apr. 2022. 1

[2] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[3] 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

[4] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip H. S. Torr, and Song Bai. Occluded Video Instance Segmentation: A Benchmark, May 2022. 3

[5] Weiyao Wang, Matt Feiszli, Heng Wang, Jitendra Malik, and Du Tran. Open-World Instance Segmentation: Exploiting Pseudo Ground Truth From Learned Pairwise Affinity. *arXiv:2204.06107 [cs]*, Apr. 2022. 1, 2, 3

[6] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified Perceptual Parsing for Scene Understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 3

[7] Zongxin Yang, Jiaxu Miao, Xiaohan Wang, Yunchao Wei, and Yi Yang. Associating Objects with Scalable Transformers for Video Object Segmentation. Mar. 2022. 1, 2

[8] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating Objects with Transformers for Video Object Segmentation. In *Advances in Neural Information Processing Systems*, volume 34, pages 2491–2502. Curran Associates, Inc., 2021. 1, 2