Spatial-Temporal Multi-level Association for Video Object Segmentation

Deshui Miao[1†], Xin Li[2†], Zhenyu He[1,2 *],
Huchuan Lu[3], and Ming-Hsuan Yang[4,5]

[1] Harbin Institute of Technology, Shenzhen [2] Peng Cheng Laboratory
[3] Dalian University of Technology [4] University of California at Merced
[5] Yonsei University

Abstract. Existing semi-supervised video object segmentation methods either focus on temporal feature matching or spatial-temporal feature modeling. However, they do not address the issues of sufficient target interaction and efficient parallel processing simultaneously, thereby constraining the learning of dynamic, target-aware features. To tackle these limitations, this paper proposes a spatial-temporal multi-level association framework, which jointly associates reference frame, test frame, and object features to achieve sufficient interaction and parallel target ID association with a spatial-temporal memory bank for efficient video object segmentation. Specifically, we construct a spatial-temporal multi-level feature association module to learn better target-aware features, which formulates feature extraction and interaction as the efficient operations of object self-attention, reference object enhancement, and test reference correlation. In addition, we propose a spatial-temporal memory to assist feature association and temporal ID assignment and correlation. We evaluate the proposed method by conducting extensive experiments on numerous video object segmentation datasets, including DAVIS 2016/2017 val, DAVIS 2017 test-dev, and YouTube-VOS 2018/2019 val. The favorable performance against the state-of-the-art methods demonstrates the effectiveness of our approach. All source code and trained models will be made publicly available.

Keywords: Video object segmentation · Spatial-temporal information

1 Introduction

Semi-supervised Video Object Segmentation (VOS) aims to delineate and track the objects specified by the given masks within a video sequence [21, 25, 27, 31, 33, 39]. This practice holds significant promise across various applications, particularly as the prevalence of video content surges in domains like autonomous driving, augmented reality [17, 24, 37], and interactive video editing [17, 41, 47]. The core challenge of VOS lies in utilizing the limited initial information (one target sample) to accurately track and segment the target object that undergoes various visual changes and interacts with a complex video environment.

† Equal contribution. ∗ Corresponding author.

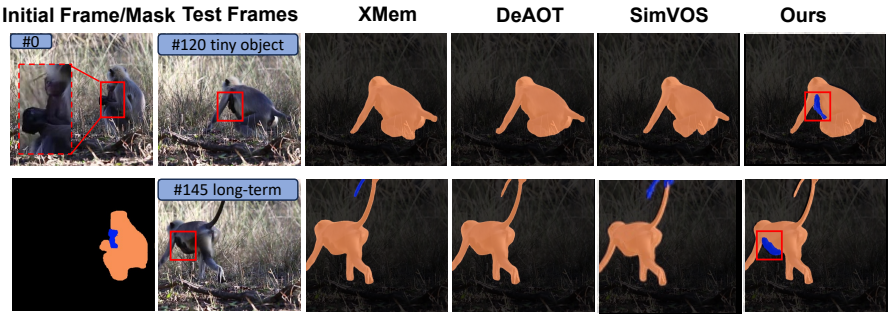

Fig. 1: Performance on challenging VOS scenarios with tiny objects and long-term changes. XMem [5], DeAOT [45], and SimVOS [36] do not work well on this scenario. Our method accurately predicts the mask of the 'baby monkey' (marked by the red box) over frames.

Early VOS methods primarily involve fine-tuning segmentation models [2, 31, 35] on annotated videos or developing pixel-wise matching maps [9, 30]. Recent approaches [5,6,25,44,45] mainly rely on matching-based frameworks, which build associations between the test frame features and memorized target features to infer the target state. The matching-based methods enable efficient temporal ID association by calculating the similarity between memorized features and the test feature, facilitating target association across multiple frames, and efficiently handling multi-object scenarios. To better model dynamic target feature learning, number VOS methods [36,50] explore integrating feature association into the feature extraction process to achieve a more comprehensive spatial feature correlation between reference and test frames. However, temporal feature association, which associates features after feature modeling, results in ineffective feature association when dealing with small object segmentation. Meanwhile, the method of spatial feature association suffers from modeling long-term associations.

The core points in comprehensive target feature learning are: 1) spatial information to ensure accurate prediction for small objects, and 2) temporal consistency that maintains the object's identity to distinguish a target from other objects and backgrounds. The above two challenges motivate us to design an efficient and sufficient spatial-temporal multi-level association method for the VOS task.

In this paper, we propose a spatial-temporal multi-level association (STMA) framework comprising a spatial-temporal multi-level feature association module (STML), a prediction part, and a spatial-temporal memory bank. During the feature extraction stage, object features, reference frames from the spatial memory, and the test frame are input into the STML module for asymmetric information interaction. Specifically, in the STML module, we decouple the self-attention into three substreams: object feature self-attention, reference object enhancement, and test reference correlation. The STML module can then learn dynamic target-aware features. Equipped with these features, we then perform

object-related semantic matching and ID correlation with the temporal memory bank. The temporal memory bank stores historical features related to the different targets, which serve as keys and values for ID assignment. The ID correlation module matches the target features with their respective IDs. It completes the transformation from one feature map to multiple target-related feature maps, ensuring each feature map contains a target with a unique ID. Especially, in the STML module, the target feature is comprehensively learned without differentiating between the IDs of different targets. In ID association, the IDs of different targets are obtained by comparing them with the IDs in temporal memory. We conduct extensive experiments on various public VOS datasets, including DAVIS 2016 & 2017 and YouTube-VOS 2018 & 2019. The favorable performance against the state-of-the-art methods on all these datasets demonstrates the effectiveness of the proposed algorithm, especially in handling challenging sequences with small targets or long-term duration.

The main contributions of this work are:

- We propose a spatial-temporal multi-level feature association module to facilitate efficient spatiotemporal target information interchange for video object segmentation.
- We develop a spatial-temporal memory bank to assist the STML module and ID assignment for long-term modeling. It retains information pertinent to the targets from previous frames, which is utilized to match, segregate, and enhance the features of each target in the test frame.
- We conduct extensive experiments to demonstrate the effectiveness of the proposed method. Our method achieves favorable performance against other state-of-the-art methods on extensive VOS datasets. Specifically, Our method gains significant improvement on DAVIS 2017 val (88.9% $\mathcal{J}\&\mathcal{F}$), DAVIS 2017 test (85.6% $\mathcal{J}\&\mathcal{F}$) and YouTube-VOS 2019 (86.3% $\mathcal{J}\&\mathcal{F}$) without any pre-training.

2 Related Work

We discuss the closely related methods from the aspects of feature association and memory construction.

Feature Association in VOS. Numerous methods [2,12,14,15,23,31,32,35,40,49] employ online learning at test time to update models and build connections between consecutive frames to transmit segmentation masks sequentially, which is not efficient in the fine-tuning step. Although some methods [4,42] improve the test efficiency, they are often susceptible to cumulative errors due to occlusions or tracking drift. Similar to STM [25], recent methods focus more on memory matching and propagation with a feature memory bank and matching strategy. In this pattern, feature matching and propagation are performed after feature extraction. XMem [5] meticulously crafts distinct memory storage mechanisms and proposes the Atkinson-Shiffrin model for matching, yielding impressive results. Zhang et al. [50] develop a correspondence-aware training framework that

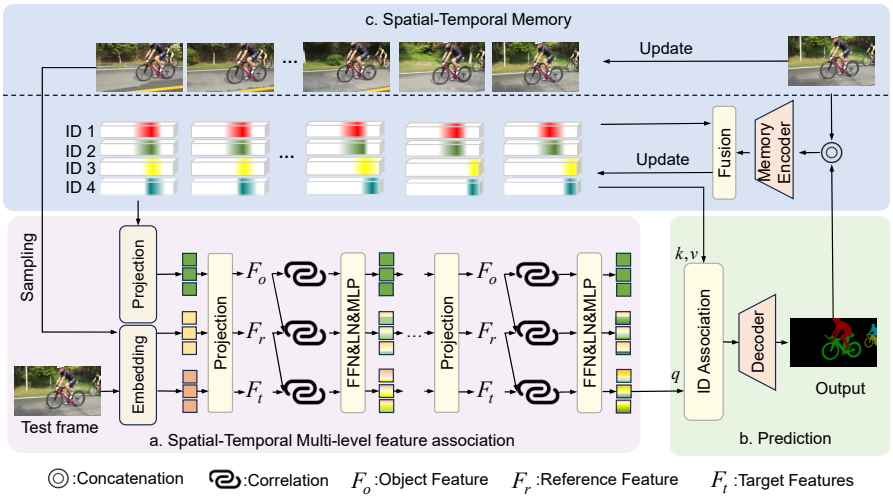

Fig. 2: Overall framework. It consists of a spatial-temporal multi-level (STML) feature association part, a prediction module, and a spatial-temporal memory. The STML module conducts simultaneous feature extraction and correlation. The spatial-temporal memory not only provides object features and reference frames for STML but also offers temporal feature information for ID association.

enhances propagation-based VOS solutions by explicitly fostering robust correspondence matching throughout the network learning process.

In addition, other approaches [20, 44, 45] leverage transformers for feature interaction and spatiotemporal information propagation. Aiming to achieve simultaneous multi-object tracking and enhance transformer efficacy within VOS, AOT [44, 45] proposes an ID propagation mechanism and advances the Long-Short-Term-Transformer (LSTT) [38] for VOS and achieve promising performance. However, the aforementioned methods, performing temporal feature matching after feature extraction, do not work well in handling small and faint objects, since the fine-grained details can be dismissed during the feature extraction process. Similar to extensive single object tracking methods [3, 10, 46], SimVOS [36] performs the joint extraction and association of target features between historical and current frames in ViT blocks. However, SimVOS processes every target sequentially during the training and testing phases, which is time-consuming and does not consider the long-term temporal association between the targets.

Memory construction in VOS. Recent advancements in the field have focused on incorporating a larger number of past frames into a feature memory bank to overcome limitations related to context understanding. STM [25] gains significant attention for this purpose and has inspired a range of subsequent research. However, a common challenge with most STM variants [5, 6, 50] is their struggle to process long videos without causing an excessive increase in the size of the feature memory bank. Although AOT [44, 45] introduces an adaptation of

the attention mechanism for transformers, it does not address the issue of GPU memory overload. Meanwhile, some approaches [12, 24] implement a localized feature memory window, which unfortunately overlooks the importance of capturing long-term context beyond this limited scope. To address the long-term VOS problem, XMem [5] designs an updating architecture that integrates several independent but intricately linked feature memory stores. SimVOS [36] facilitates information transfer between different frames by updating reference frames. To better model spatial-temporal information, we construct spatial-temporal memory and perform spatial-temporal multi-level feature association within a single framework to achieve both efficient feature representation and effective ID association. Furthermore, we introduce object features to enhance the discriminability between different target representations, thereby better distinguishing between multi-targets and enhancing the accuracy of ID association.

3 Proposed Algorithm

Our approach aims to learn a spatially-aware and dynamically-distinctive target representation for video object segmentation. To this end, we propose a spatial-temporal multi-level association VOS framework comprising a spatial-temporal memory to remember and update the target object information adaptively, a feature association part to generate target features, and a prediction component to generate the final results.

3.1 Overall Framework

Given the test frame $\mathbf{X}_t \in \mathbb{R}^{3\times H\times W}$ with n target objects, the goal of our method is to predict the target masks $\mathbf{M}_t \in \mathbb{R}^{H\times W}$ of $\mathbf{X}_t$ based on the given reference frames $\mathbf{X}_r \in \mathbb{R}^{m\times 3\times H\times W}$, where m is the number of reference frames. The proposed method takes the test frame feature, reference frame feature, and object feature as input. The features of the test and reference frames are generated using a linear model. Specifically, we convert each input frame into a string of 2D flat patches with dimension $N \times 3P^2$, resulting in the reconfigured reference sequences, namely, $\mathbf{X}_r \in \mathbb{R}^{m\times N\times 3P^2}$ and the test sequence $\mathbf{X}_t \in \mathbb{R}^{N\times 3P^2}$, where P^2 is the patch size and $N = HW/P^2$ is the number of patches. After employing the linear projection $\mathbf{E} \in \mathbb{R}^{3P^2\times C}$ to transform the 2D patches into 1D tokens of C dimensions and incorporating the sinusoidal positional embedding $\mathbf{P} \in \mathbb{R}^{N\times C}$, we get the reference feature $\mathbf{F}_r \in \mathbb{R}^{m\times N\times C}$, target feature $\mathbf{F}_t \in \mathbb{R}^{N\times C}$. The object features $\mathbf{F}_o \in \mathbb{R}^{n\times C}$ are generated using a projection layer and a position embedding operation.

Upon the features of the three inputs, the proposed method conducts spatial-temporal multi-level feature association to generate the correlated target features by leveraging the attention-based correlation. We then split the target feature from the above-generated feature and perform ID association to get the feature map of every target. Specifically, we calculate the query-key affinity matrix between $\mathbf{F}_t$ and $\mathbf{Fk} \in \mathbb{R}^{T\times(H//16)\times(W//16)\times C}$ from temporal memory to read out

the feature corresponding to the target feature, where T represents the length already stored in memory. By combining the similarity matrix with the memorized ID values $\mathbf{v} \in \mathbb{R}^{T\times(H//16)\times(W//16)\times C}$, we compute the readout features $\mathbf{F}_R$ related to the object targets and predict the final target masks using a decoder. The overall process on the t-th frame can be formulated as:

$$M_t = \phi\Big(P_{mem}\big(\mathrm{STML}(\mathbf{F}_t, \mathbf{F}_r, \mathbf{F}_o), \mathbf{M}_{target}\big), \mathbf{F}_t^{1/8}, \mathbf{F}_t^{1/4}\Big), \tag{1}$$

where $\mathbf{M}_{target}$ represents the memorized features, P_{mem} indicates the function of the ID association module, $\mathbf{F}_t^{1/4}$ and $\mathbf{F}_t^{1/8}$ are the multi-scale features of frame t, and ϕ is the decoder function, which predicts the target masks.

3.2 Spatial-Temporal Multi-Level Feature Association

The spatial-temporal multi-level feature association model is proposed to associate the target frame, reference frames from the spatial memory, and object features to learn comprehensive target-aware features for the subsequent target differentiation and mask prediction. As shown in Figure 3, the main component of the STML module is the spatial-temporal correlation block, which uses the Multi-Head Attention block [29] to facilitate the interaction between different types of input information. As directly computing the cross-attention between all three kinds of input features is inefficient, we decompose the calculation into the operations of object self-attention $\mathbf{A}^o_{attn}$, reference object enhancement $\mathbf{A}^r_{attn}$, and test reference correlation $\mathbf{A}^t_{attn}$.

For object features, we conduct self-attention to enhance the distinction between different targets. The self-attention of the object features $\mathbf{A}^o_{attn}$ is computed as:

$$\mathbf{A}^o_{attn} = f_{softmax}\left(\frac{q^o {k^o}^\top}{\sqrt{d}}\right) v^o, \tag{2}$$

where $f_{softmax}$ represents the operation of SoftMax, q^o, k^o, and v^o are generated by projecting $\mathbf{F}_o$. We then use the object features to enhance each reference feature, which is performed by correlating the object features with every reference feature. For each reference frame feature i, the asymmetric attention is defined as:

$$\begin{aligned} k^{r_i}_{tem} &= f_{concat}\left(W^K\left[\mathbf{F}^i_r\right], W^K\left[\mathbf{F}_o\right]\right), \\ v^{r_i}_{tem} &= f_{concate}\left(W^V\left[\mathbf{F}^i_r\right], W^V\left[\mathbf{F}_o\right]\right), \\ \mathbf{A}^{r_i}_{attn} &= f_{softmax}\left(\frac{q^{r_i} {k^{r_i}_{tem}}^\top}{\sqrt{d}}\right) v^{r_i}_{tem}, \end{aligned} \tag{3}$$

where $*_{tem}$ is a temporary features and f_{concat} is the operation to concatenate features. Our rationale for individually segmenting each reference frame feature during information correlation is to ensure the reference features remain uninfluenced by one another, preserving their discrete integrity.

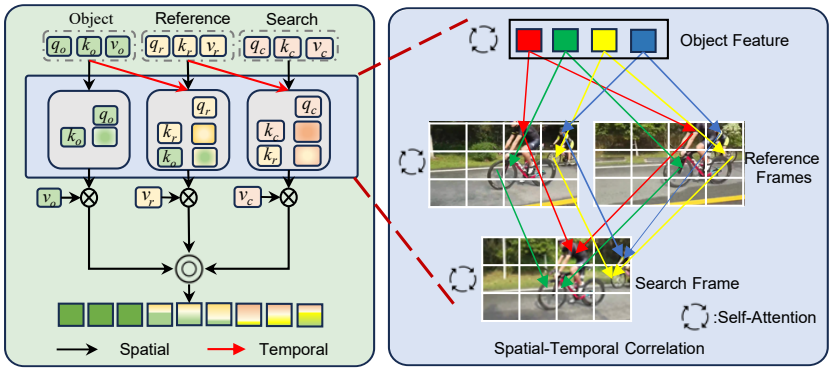

Fig. 3: Illustration of the proposed spatial-temporal correlation. Given two reference frames as examples. The object features conduct self-attention and the reference features perform attention both with themselves and object features. The target feature undergoes attention with both itself and the reference feature simultaneously.

Then, the asymmetric information propagation is performed between the feature of the target and references.

$$\begin{aligned} k_{tem}^{t} &= f_{concat}\left(W^{K}\left[\mathbf{F}_{t}\right], W^{K}\left[\mathbf{F}_{r}\right]\right), \\ v_{tem}^{t} &= f_{concat}\left(W^{V}\left[\mathbf{F}_{t}\right], W^{V}\left[\mathbf{F}_{r}\right]\right), \\ \mathbf{A}_{attn}^{t} &= f_{softmax}\left(\frac{q^{t} {k_{tem}^{t}}^{\top}}{\sqrt{d}}\right) v_{tem}^{t}. \end{aligned} \tag{4}$$

In this manner, the target feature only receives the information from the reference feature, and the references remain unchanged during this step. With the multi-level association block, the information can also propagate to the target feature.

The self-attention among object features serves to differentiate various objects. Each reference frame feature interacts solely with its respective frame and the object features, ensuring no information interference between reference frames while concurrently enhancing target data. In addition to self-attention, the test feature engages with reference features, assimilating valuable information from them. Such a strategic configuration is instrumental in sharpening the target details within reference features and enriches the target information in the test frame. The intuitive representation of STML in one block is shown in 3, where the correlation among three types of features can be strengthened by our mechanism.

3.3 Spatial-Temporal Memory

In this section, we give more details about the updation of spatial-temporal memory and how to utilize the temporal memory for object feature generation.

Table 1: Ablation study on the DAVIS 2017 val and test datasets.

Dataset	DAVIS 2017 val			DAVIS 2017 test			FPS
Variant	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	
W/O Obj	88.2	84.8	91.6	83.8	80.1	86.8	11.2
W/O Spatial Memory	87.8	84.8	90.9	83.5	80.4	86.6	22.1
W/O Temporal Memory	88.1	84.9	90.9	81.0	77.5	84.6	3.5
Joint Self-Attn	87.8	84.8	90.8	82.7	79.5	85.8	9.6
Full version (Proposed method)	**88.9**	**86.0**	**91.8**	**85.6**	**81.9**	**89.2**	10.7

Object Feature Generation. We propose a straightforward method to generate object ID features. Object features $\mathbf{F}_0 \in \mathbb{R}^{n\times C\times(H//16)\times(W//16)}$ is initialized as all zeros and then generated by interacting with the first frame and ground-truth mask by a feature enhancement layer. Object ID features are generated by pooling the memorized ID values while processing frame t.

$$\mathbf{F}_o = f_o\left(\mathbf{v}\right) \in \mathbb{R}^{n\times C}, \tag{5}$$

where f_o projects the memorized ID features like MaxPooling. With the object features, we model the propagation of frame t in the proposed STML module.
Spatial Memory Update. After completing the target prediction, the target frame is updated into the spatial memory according to a first-in-first-out strategy. To ensure that targets are not lost in long-term scenarios, spatial memory will not discard the frame that first provided the target.
Temporal Memory Update. After obtaining the prediction output of the current frame, we employ a simple fusion strategy to add the current frame into memory. The target feature of STML is updated as the key into the temporal memory. By concatenating the output masks of different targets with the target frame and encoding the features, the ID values of different targets are obtained and stored in the temporal memory. Similar to XMem [5], when the memory reaches its specified capacity, we update the elements in memory using the least-frequently-used (LFU) eviction algorithm.

3.4 Training and Inference

Training. Our method follows the three-step training scheme used by previous methods [5, 6, 50]. We first conduct the synthetic static pre-training on various static datasets [8, 19, 28, 34, 48]. Then we fine-tune the model on video datasets, including DAVIS [26] and YouTube-VOS [39]. To improve the performance, we also perform training on the BLK30k dataset [7]. We use the strategy of curriculum sampling, and the sampled sequence length is set to 5. Two of the past frames in spatial memory are randomly selected as the reference frames. The training loss uses a combination of bootstrapped cross entropy loss and Dice loss, with equal weights of 0.5 for each.
Inference. We use a first-in-first-out queue to memorize the reference frames in spatial memory for both DAVIS dataset [26] and YouTube-VOS dataset [39].

Table 2: Comparison over different backbone and pre-trained weights. It shows that our method achieves the best performance under the same backbone and pre-trained weights.

Methods	Backbone	Pretrained	D17 val	D17 test	Y19 val	FPS
XMem [5]	ConvViTb	MAE	85.7	82.1	84.7	22.3
DeAOTL [45]	ConvViTb	MAE	84.2	80.5	83.6	8.9
SimVOS-BS [36]	ViTb	MAE	87.1	82.2	79.3	7.4
SimVOS [36]	ViTb	MOCOV3	81.3	-	-	3.1
	ViTb	MAE	88.0	80.4	84.2	3.1
Ours	ViTb	MOCOV3	84.1	79.7	83.4	11.0
	ViTb	MAE	**88.2**	**82.1**	**85.0**	11.0
	ConvViTb	Scratch	83.6	77.0	82.3	10.7
	ConvViTb	MAE	**88.9**	**85.6**	**86.3**	10.7

Specifically, we keep the first frame unchanged in our memory. For the DAVIS dataset [26], the frame is added to the queue every 3rd frame, while the 5th frame is for the YouTube-VOS dataset. All the results are generated in the NVIDIA V100 GPU.

4 Experiments

In this section, we conduct experiments to evaluate the proposed method on numerous VOS datasets from multiple aspects. We first introduce the implementation details of our model. Then, we conduct ablation studies to validate the effectiveness of the spatial-temporal memory and STML module. Finally, we evaluate the overall performance of the proposed method against state-of-the-art approaches using both qualitative and quantitative measures. Visualized results on challenging sequences are provided for a comprehensive qualitative analysis.

4.1 Implementation Details

We use the ViT with different pre-trained weights to construct the STML module. The decoder consists of different residual upsampling blocks, which gradually fuse the multi-scale features to predict the final single-channel logits. We use AdamW [16] with learning rate $3e^{-5}$ and a weight decay of 0.05 as the optimizer. The pretrain iteration on the static dataset is 150k with batch size 16, and stage 1 is performed with 250k iterations. The main training stage on DAVIS and YouTube-VOS is conducted with 160k iterations with batch size 16. The learning rate reduces by 10 times after 100k iterations. More experimental settings can be found in the supplement material.

4.2 Ablation Study

In this section, we conduct experiments on different datasets to validate the effectiveness of our proposed approach. In Table 1, We train five different variants

Table 3: Quantitative comparisons on the DAVIS 2017, YouTube-VOS 2018 & 2019 dataset. The best two results are shown in red and blue color. In the table, * and † denote the models are pre-trained using the additional static image datasets and the large BL30K dataset, respectively.

Dataset	DAVIS 2016 val			DAVIS 2017 val			DAVIS 2017 test			YouTube-VOS 2018 val					YouTube-VOS 2019 val				
Method	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{G}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	$\mathcal{G}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
SST [11]	-	-	-	82.5	79.9	85.1	-	-	-	81.7	81.2	-	76.0	-	81.8	80.9	-	76.6	-
JOINT [24]	-	-	-	83.5	80.8	86.2	-	-	-	83.1	81.5	85.9	78.7	86.5	82.8	80.8	84.8	79.0	86.6
XMem [5]	-	-	-	84.5	81.4	87.6	79.8	76.3	83.4	84.3	83.9	88.8	77.7	86.7	84.2	83.8	88.3	78.1	86.7
VOS−Corr [50]	-	-	-	86.1	82.7	89.5	81.0	77.3	84.7	85.6	84.9	89.7	79.0	87.8	-	-	-	-	-
SimVOS-B [36]	92.9	91.3	94.4	88.0	85.0	91.0	80.4	76.1	84.6	-	-	-	-	-	84.2	83.1	87.1	79.1	84.7
STMA	91.6	90.3	93.0	**88.9**	**86.0**	**91.8**	**85.6**	**81.9**	**89.2**	86.1	**85.1**	**90.1**	**80.1**	**89.1**	**86.3**	**85.0**	**89.5**	**81.3**	**89.2**
STM [25] *	89.3	88.7	89.9	81.8	79.2	84.3	-	-	-	79.4	79.7	84.2	72.8	80.9	-	-	-	-	-
AFB-URR [22] *	-	-	-	76.9	74.4	79.3	-	-	-	79.6	78.8	83.1	74.1	82.6	-	-	-	-	-
CFBI [43] *	-	-	-	81.9	79.1	84.6	74.8	71.1	78.5	81.4	81.1	85.8	75.3	83.4	81.0	80.6	85.1	75.2	83
MiVOS [7]*	-	-	-	84.5	81.7	87.4	78.6	74.9	82.2	82.6	81.1	85.6	77.7	86.2	82.4	80.6	84.7	78.1	86.4
STCN [6] *	91.6	90.8	92.5	85.3	82.0	88.6	77.8	74.3	81.3	84.3	83.2	87.9	79.0	87.3	84.2	82.6	87.0	79.4	87.7
Swin-B-AOT-L [44] *	92.0	90.7	93.3	85.4	82.4	88.4	81.2	77.3	85.1	85.1	85.1	90.1	78.4	86.9	85.3	84.6	89.5	79.3	87.7
SwinB-DeAOT-L [45] *	92.9	91.1	94.7	86.2	83.1	89.2	82.8	78.9	86.7	86.3	85.4	90.7	80.1	89.0	86.4	85.4	90.3	80.5	89.3
XMem [5] *	91.5	90.4	92.7	86.2	82.9	89.5	81.0	77.4	84.5	85.7	84.6	89.3	80.2	88.7	85.5	84.3	88.6	80.3	88.6
VOS-Corr [50]	92.2	91.1	93.3	87.7	84.1	91.2	82.0	78.3	85.6	86.9	85.5	90.2	81.6	90.4	86.6	85.3	89.8	81.4	89.8
ISVOS [33] *	92.6	91.5	93.7	87.1	83.7	90.5	82.8	79.3	86.2	86.3	85.5	90.2	80.5	88.8	86.1	85.2	89.7	80.7	88.9
STMA *	**93.4**	**91.9**	**94.8**	**90.4**	**87.3**	**93.5**	**87.1**	**83.7**	**90.5**	86.4	85.0	89.6	81.5	89.3	**86.8**	**85.6**	90.0	**81.8**	89.5
XMem [5]†	92.0	90.7	93.2	87.7	84.0	91.4	81.2	77.6	84.7	86.1	85.1	89.8	80.3	89.2	85.8	84.8	89.2	80.3	88.8
ISVOS [33] †	92.8	91.8	93.8	88.2	84.5	91.9	84.0	80.1	87.8	86.7	86.1	90.8	81.0	89.0	86.3	85.2	89.7	81.0	89.1
STMA †	**93.5**	**92.1**	**94.9**	**90.8**	**87.8**	**93.9**	**87.7**	**84.4**	**91.0**	**87.0**	85.7	90.2	**82.0**	**90.0**	**87.1**	**85.6**	**90.0**	**82.4**	**90.1**

of our method on the YouTube-VOS and DAVIS datasets with ConvVitb as the STML module. In Table 2, performance with different backbone and pre-trained weights is compared.

W/O Obj, which removes the object features from the STML module. The reference features conduct self-attention within themselves. The target feature employs self-attention and cross-attention with reference features.

W/O Spatial Memory, which removes the reference features from the input of the STML module. We conduct cross-attention between target features and object features.

W/O Temporal Memory, which removes the temporal memory and ID association module. The ID classification of different targets is similar to SimVOS, which adds the mask of every target to the corresponding frames. We concatenate the mask of each target to the corresponding frames and follow the SimVOS to perform the VOS task.

Joint Self-Attn conducts a straightforward self-attention by concatenating all three inputs, object features, reference features, and target features.

Effect of object features. Without object features, the performance on three different validation sets all decreases. Specifically, using object features gains 0.7% and 2.5% in $\mathcal{J}\&\mathcal{F}$ on the DAVIS 2017 val and DAVIS 2017 test, respectively. The results validate the benefits of enhancing the target representation in reference and test frames by utilizing object features in the STML module.

Effect of the STML module. We present the performance of different as-

Table 4: Long-term performance on the LVOS test set. The results are obtained with or without fine-tuning the training set of the LVOS dataset.

Method		AFB-URR [22]	CFBI [43]	STCN [6]	RDE [18]	XMem [5]	LWL [1]	AOT-L [44]	AOT-B [44]	DDMem [13]	**Ours**
Without Finetuning	$\mathcal{J}\&\mathcal{F}$	39.9	44.8	45.8	49.0	49.5	50.7	54.1	54.4	55.0	**56.8**
	$\mathcal{J}$	36.2	40.5	41.6	44.4	45.2	46.5	48.7	49.3	49.9	**52.7**
	$\mathcal{F}$	43.6	49.0	50.0	53.5	53.7	54.8	59.5	59.4	60.2	**60.9**
Finetuning	$\mathcal{J}\&\mathcal{F}$	40.8	44.8	48.3	50.2	50.9	50.8	54.7	54.5	55.7	**57.4**
	$\mathcal{J}$	37.5	40.2	44.0	45.7	46.5	46.4	49.2	49.2	50.3	**53.3**
	$\mathcal{F}$	44.1	49.4	52.5	54.6	55.3	55.2	60.2	59.8	61.2	**61.5**

sociations among object, reference, and test features. First, we try to remove the object features or reference features. Both decrease the performance on the DAVIS dataset, indicating that our STML module achieves sufficient feature correlation and target association. Then, we explore a straightforward way to jointly perform self-attention by concatenating the three features, which gains no improvement on both datasets and even gets lower $\mathcal{J}\&\mathcal{F}$ in the test set. Although this way generates sufficient correlation, it utilizes extra background noises than our proposed method. Especially on the DAVIS 2017 test set, our method gains 2.0+% in $\mathcal{J}\&\mathcal{F}$. These comparisons demonstrate the efficiency of STML in effectively managing the association of target features, thereby making it easier for the model to distinguish each object.

Effect of backbone and pre-trained weights. Table 2 shows the comparison of our methods with other SOTA methods on different backbones and pre-trained weights. We replace the backbone of XMem and DeAOT with ConvViTb and train them on video datasets. The results show that our method outperforms other methods substantially with the same backbone. We also confirm the importance of pre-trained weights. Compared to methods using transformers for ID association, our method achieved favorable inference speeds, especially when compared with SimVOS.

4.3 State-of-the-Art Comparison

We evaluate the proposed method by comparing it with the state-of-the-art methods from both quantitative and qualitative analyses. All the models are trained with ConvViTb as the STML module. The experiments are conducted on five datasets, including DAVIS2016 val set, DAVIS 2017 val & test set, and YouTube-VOS 2018 & 2019 val set. We discuss the detailed results on every dataset below.

Quantitative Analysis.

Table 3 compares our method with state-of-the-art models on five benchmarks. **DAVIS 2016** [26] is a classical VOS dataset that offers 50 high-quality, single-object video sequences with frame-by-frame pixel-level annotations. After pertaining, our method gains significant improvement (93.5% $\mathcal{J}\&\mathcal{F}$) on the val set.

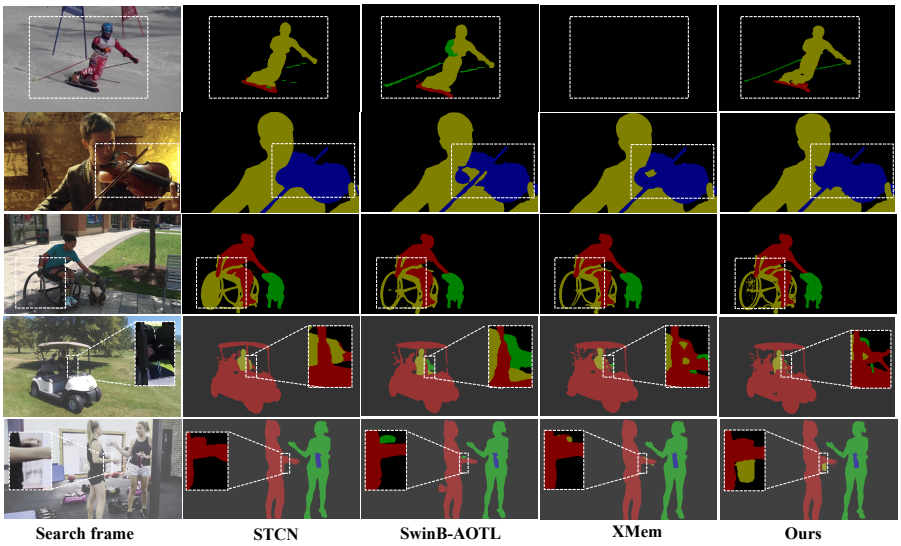

Fig. 4: Visualized results on sequences with small and faint objects. It shows that our method generates finer masks compared to the state-of-the-art methods.

Extensive information enhancement of targets in the STML module makes distinguishing and identifying target objects easier during model prediction.

DAVIS 2017 [26], an extension of DAVIS2016, is a benchmark that offers densely annotated, high-quality, full-resolution videos with multiple objects of interest. The val set has 30 videos, and the test set has more videos in challenging scenarios. Table 3 shows our method achieves gains better performance in both DAVIS 2017 [26] val and test-dev set. Even without pre-training on static images, our model still shows favorable results, 88.9% on val and 85.6% on test-dev. The improvements validate the efficacy of our model in multi-object and small-target scenarios.

YouTube-VOS 2018 [39] has 3471 videos with 65 categories for training and 474 videos for validation. In the validation, there are 26 categories that the model has not seen in its training, enabling us to evaluate its generalization ability for class-agnostic targets. Table 3 illustrates our method achieves enhanced performance compared to previous state-of-the-art models. Specifically, our model performs better in unseen categories, with 82.0% in $\mathcal{J}_u$ and 90.0% in $\mathcal{F}_u$, indicating that our method possesses a more powerful generalization ability.

YouTube-VOS 2019 is an extension of YouTube-VOS 2018, featuring a greater number of masked targets and including more challenging sequences in its val set. In Table 3, our method achieves competitive performance against the state-of-the-art methods. Notably, our model gets favorable performance without any pre-training stages.

LVOS [13] is a new benchmark for evaluating video object segmentation algorithms in long-term challenging sequences with diverse scenarios and pixel-level

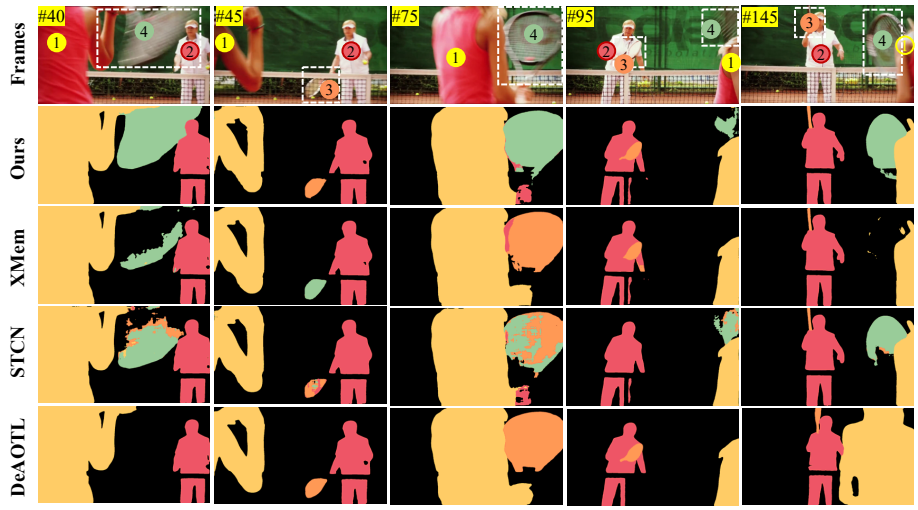

Fig. 5: Visualized results on sequences with complicated ID connections. The proposed method performs well in tracking the tennis rackets, which demonstrates excellent performance in terms of ID propagation.

annotations. We evaluate our method on the LVOS [13] test set without fine-tuning. Table 4 shows that our method achieves **56.8%** $\mathcal{J}\&\mathcal{F}$ and performs better than previous approaches, which demonstrates the robustness of our memory bank and object features when dealing with long-term scenarios.

Qualitative Analysis.
We visualize some challenging sequences, including scenarios involving small and faint targets, as well as those where targets closely resemble the background. In Figure 4, our approach shows favorable segmentation performance against the state-of-the-art methods. For instance, our model accurately segments targets and provides detailed results in a video that includes wheelchairs, golf clubs, and transparent bottles. This is attributed to our proposed spatial-temporal multi-level feature association, which enables our model to capture more detailed target information by sufficient interaction, thereby facilitating precise segmentation in these challenging scenarios Additionally, incorporating object features in our model ensures consistent maintenance of the target ID throughout extended video sequences, such as in the case of skating. Figure 5 displays visualized results from different methods within a video sequence. Other methods encounter issues with ID misalignment when distinguishing between two badminton rackets. Although SwinB-DeAOTL [44] accurately segments target 4 at $75th$ frame, the ID assignment is imprecise. In comparison, our method distinguishes different objects with the same attributes, further validating the effectiveness of our framework.

Fig. 6: Failure case. It shows that our approach incorrectly associates the goose with ID 1 with another goose due to occlusion and insufficient discriminative information.

4.4 Limitations

Our method enhances the segmentation precision and the accuracy of ID matching for small targets through multi-object feature correlation and ID association, achieving efficient video object segmentation. However, in more complex scenes, such as those depicted in Figure 6, it becomes challenging to distinguish background distractions from the target due to their high similarity in appearance and location. To be more specific, Figure 6 shows that our method erroneously propagates the IDs when targets 1 and 2 are obscured by similar objects. This may be due to the severe loss of discriminative features when the target is occluded, leading to issues of ID-switch and difficulty in correctly distinguishing between the target and background distractions. A potential direction to improve this issue could be incorporating additional information, such as motion information or textual descriptions, into the modeling process.

5 Conclusion

In this paper, we present a spatial-temporal multi-level association framework to achieve efficient video object segmentation. Our proposed method enhances efficiency while ensuring sufficient interaction of feature information by separating the tasks of feature extraction and interaction modeling into three distinct components: object self-attention, reference target enhancement, and target object correlation. In the spatial-temporal multi-level feature association module, we enhance the feature information related to the multiple targets through asymmetric interactions involving three types of information. Then, we separate the features of each target through ID propagation from the enhanced feature map for decoding. Our methodology can be described as the augmentation of features for multiple targets, followed by assigning unique identification to each target. The favorable performance on all the tested datasets against the state-of-the-art methods demonstrates the effectiveness of the proposed algorithm.

Acknowledgements

The paper is supported by the Guangdong Basic and Applied Basic Research Foundation (Grant No.2024A1515011292), the Major Key Project of PCL (PCL2023A08), the National Natural Science Foundation of China (Grant No.62172126), and the Shenzhen Research Council (Grant No.JCYJ20210324120202006).

References

1. Bhat, G., Lawin, F.J., Danelljan, M., Robinson, A., Felsberg, M., Van Gool, L., Timofte, R.: Learning what to learn for video object segmentation. In: ECCV. pp. 777–794. Springer (2020) 11
2. Caelles, S., Maninis, K., Pont-Tuset, J., Leal-Taixé, L., Van Gool, L.: One-shot video object segmentation. In: CVPR. pp. 221–230 (2017) 2, 3
3. Chen, B., Li, P., Bai, L., Qiao, L., Shen, Q., Li, B.: Backbone is all you need: A simplified architecture for visual object tracking. In: ECCV (2022) 4
4. Chen, Y., Pont-Tuset, J., Montes, A., Van Gool, L.: Blazingly fast video object segmentation with pixel-wise metric learning. In: CVPR. pp. 1189–1198 (2018) 3
5. Cheng, H.K., Schwing, A.G.: Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: ECCV (2022) 2, 3, 4, 5, 8, 9, 10, 11
6. Cheng, H.K., Tai, Y.W., Tang, C.K.: Rethinking spacetime networks with improved memory coverage for efficient video object segmentation. In: NeurIPS. pp. 11781–11794 (2021) 2, 4, 8, 10, 11
7. Cheng, H., Tau, Y., Tang, C.: Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In: CVPR (2021) 8, 10
8. Cheng, H.K., Chung, J., Tai, Y.W., Tang, C.K.: Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In: CVPR. pp. 8890–8899 (2020) 8
9. Cheng, J., Tsai, Y., Hung, W., Wang, S., Yang, M.: Fast and accurate online video object segmentation via tracking parts. In: CVPR (2018) 2
10. Cui, Y., Jiang, C., Wang, L., Wu, G.: Mixformer: End-to-end tracking with iterative mixed attention. In: CVPR. pp. 13608–13618 (2022) 4
11. Duke, B., Ahmed, A., Wolf, C., Taylor, G.W.: Sstvos: Sparse spatiotemporal transformers for video object segmentation. In: CVPR. pp. 5912–5921 (2021) 10
12. Ge, W., Lu, X., Shen, J.: Video object segmentation using global and instance embedding learning. In: CVPR (2021) 3, 5
13. Hong, L., Chen, W., Liu, Z., Zhang, W., Guo, P., Chen, Z., Zhang, W.: Lvos: A benchmark for long-term video object segmentation. In: ICCV (2023) 11, 12, 13
14. Hu, P., Wang, G., Kong, X., Kuen, J., Tan, Y.P.: Motion-guided cascaded refinement network for video object segmentation. In: CVPR (2018) 3
15. Hu, Y.T., Huang, J.B., Schwing, A.: Maskrnn: Instance level video object segmentation. In: NeurIPS (2017) 3
16. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 9
17. Li, L., Ji, W., Wu, Y., Li, M., Qin, Y., Wei, L., Zimmermann, R.: Panoptic scene graph generation with semantics-prototype learning. arXiv preprint arXiv:2307.15567 (2023) 1
18. Li, M., Hu, L., Xiong, Z., Zhang, B., Pan, P., Liu, D.: Recurrent dynamic embedding for video object segmentation. In: CVPR (2022) 11
19. Li, X., Wei, T., Chen, Y.P., Tai, Y.W., Tang, C.K.: Fss-1000: A 1000-class dataset for few-shot segmentation. In: CVPR (June 2020), 5 8
20. Li, X., Loy, C.C.: Video object segmentation with joint re-identification and attention-aware mask propagation. In: ECCV (2018) 4
21. Liang, Y., Li, X., Jafari, N., Chen, J.: Video object segmentation with adaptive feature bank and uncertain-region refinement. In: NeurIPS (2020) 1

22. Liang, Y., Li, X., Jafari, N., Chen, J.: Video object segmentation with adaptive feature bank and uncertain-region refinement. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H.T. (eds.) NeurIPS. vol. 33, pp. 3430–3441. Curran Associates, Inc. (2020) 10, 11
23. Luiten, J., Voigtlaender, P., Leibe, B.: Premvos: Proposal-generation, refinement and merging for video object segmentation. In: CVPR. pp. 565–580 (2018) 3
24. Mao, Y., Wang, N., Zhao, W., Li, H.: Joint inductive and transductive learning for video object segmentation. In: ICCV (2021) 1, 5, 10
25. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: ICCV. pp. 9226–9235 (2019) 1, 2, 3, 4, 10
26. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017) 8, 9, 11, 12
27. Seeing, H., Oh, S., Lee, J., Lee, S., Lee, S., Kim, E.: Hierarchical memory matching network for video object segmentation. In: ICCV (2021) 1
28. Shi, J., Yan, Q., Xu, L., Jia, J.: Hierarchical image saliency detection on extended cssd. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(4), 717–729 (2016), 5 8
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 6
30. Voigtlaender, P., Chai, Y., Schroff, F., Adam, H., Chen, L.: Feelvos: Fast end-to-end embedding learning for video object segmentation. In: CVPR. pp. 9481–9490 (2019) 2
31. Voigtlaender, P., Leibe, B.: Online adaptation of convolutional neural networks for video object segmentation. In: BMVC (2017) 1, 2, 3
32. Wang, H., Jiang, X., Ren, H., Hu, Y., Bai, S.: Swiftnet: Real-time video object segmentation. In: CVPR (2021) 3
33. Wang, J., Chen, D., Wu, Z., Luo, C., Tang, C., Dai, X., Zhao, Y., Xie, Y., Yuan, L., Jiang, Y.G.: Look before you match: Instance understanding matters in video object segmentation. In: CVPR (2023) 1, 10
34. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: CVPR (July 2017), 5 8
35. Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.: Fast online object tracking and segmentation: A unifying approach. In: CVPR (2019) 2, 3
36. Wu, Q., Yang, T., Wu, W., Chan, A.: Scalable video object segmentation with simplified framework. In: ICCV (2023) 2, 4, 5, 9, 10
37. Xia, Y., Xu, Y., Li, S., Wang, R., Du, J., Cremers, D., Stilla, U.: Soe-net: A self-attention and orientation encoding network for point cloud-based place recognition. In: CVPR (2021) 1
38. Xu, M., Xiong, Y., Chen, H., Li, X., Xia, W., Tu, Z., Soatto, S.: Long short-term transformer for online action detection. NeurIPS **34**, 1086–1099 (2021) 4
39. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. In: ECCV (2018) 1, 8, 12
40. Xu, X., Wang, J., Ming, X., Lu, Y.: Towards robust video object segmentation with adaptive object calibration. In: ACM MM (2022) 3
41. Yang, J., Peng, W., Li, X., Guo, Z., Chen, L., Li, B., Ma, Z., Zhou, K., Zhang, W., Loy, C.C., et al.: Panoptic video scene graph generation. In: CVPR. pp. 18675–18685 (2023) 1
42. Yang, L., Wang, Y., Xiong, X.: Efficient video object segmentation via network modulation. In: CVPR. pp. 6499–6507 (2018) 3

43. Yang, Z., Wei, Y., Yang, Y.: Collaborative video object segmentation by foreground-background integration. In: ECCV (2020) 10, 11
44. Yang, Z., Wei, Y., Yang, Y.: Associating objects with transformers for video object segmentation. In: NeurIPS (2021) 2, 4, 10, 11, 13
45. Yang, Z., Yang, Y.: Decoupling features in hierarchical propagation for video object segmentation. In: NeurIPS (2022) 2, 4, 9, 10
46. Ye, B., Chang, H., Ma, B., Shan, S., Chen, X.: Joint feature learning and relation modeling for tracking: A one-stream framework. In: ECCV (2022) 4
47. Yoo, J.S., Lee, H., Jung, S.W.: Video object segmentation-aware video frame interpolation. In: ICCV. pp. 12322–12333 (2023) 1
48. Zeng, Y., Zhang, P., Zhang, J., Lin, Z., Lu, H.: Towards high-resolution salient object detection. In: ICCV (October 2019), 5 8
49. Zhang, Y., Wu, Z., Peng, H., Lin, S.: A transductive approach for video object segmentation. In: CVPR (2020) 3
50. Zhang, Y., Li, L., Wang, W., Xie, R., Song, L., Zhang, W.: Boosting video object segmentation via space-time correspondence learning. In: CVPR (2023) 2, 3, 4, 8, 10