# Alignment Before Aggregation: Trajectory Memory Retrieval Network for Video Object Segmentation

Rui Sun[1*] Yuan Wang[1*] Huayu Mai[1] Tianzhu Zhang[1,2,3†] Feng Wu[1,2]
[1] University of Science and Technology of China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[3]Deep Space Exploration Lab
{issunrui,wy2016,mai556}@mail.ustc.edu.cn, {tzzhang, fengwu}@ustc.edu.cn

## Abstract

*Memory-based methods in semi-supervised video object segmentation task achieve competitive performance by performing dense matching between query and memory frames. However, most of the existing methods neglect the fact that videos carry rich temporal information yet redundant spatial information. In this case, direct pixel-level global matching will lead to ambiguous correspondences. In this work, we reconcile the inherent tension of spatial and temporal information to retrieve memory frame information along the object trajectory, and propose a novel and coherent Trajectory Memory Retrieval Network (TMRN) to equip with the trajectory information, including a spatial alignment module and a temporal aggregation module. The proposed TMRN enjoys several merits. First, TMRN is empowered to characterize the temporal correspondence which is in line with the nature of video in a data-driven manner. Second, we elegantly customize the spatial alignment module by coupling SVD initialization with agent-level correlation for representative agent construction and rectifying false matches caused by direct pairwise pixel-level correlation, respectively. Extensive experimental results on challenging benchmarks including DAVIS 2017 validation / test and Youtube-VOS 2018 / 2019 demonstrate that our TMRN, as a general plugin module, achieves consistent improvements over several leading methods.*

## 1. Introduction

Semi-supervised Video Object Segmentation (VOS) is a fundamental task to perform pixel-wise classification of a set of class-agnostic objects in video sequences. It has been widely applied to autonomous driving [61], video editing [31], augmented reality [30], *etc*. Since the object mask is only given in the first frame without any other prior information assumptions, how to fully exploit limited and semantic-agnostic information to perform accurate segmentation in the subsequent frames is thus extremely challenging.

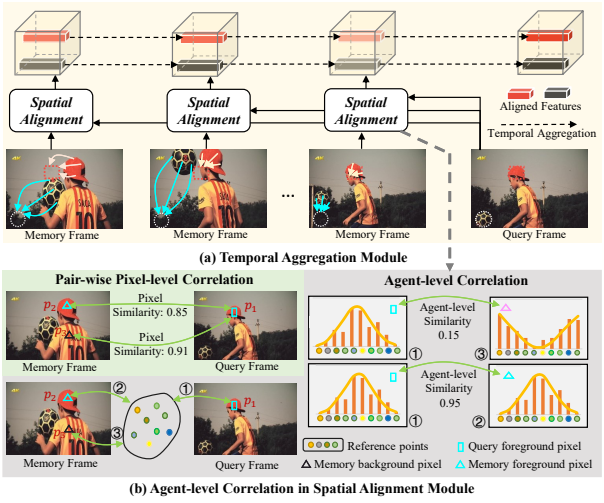



Figure 1: Illustration of our motivation. (a) shows TMRN reconciles the inherent tension of spatial and temporal information to retrieve memory frame information along the object trajectory. (b) shows we carefully design a spatial alignment module through a set of representative reference features (agents) to rectify direct pairwise pixel-level correlation caused by distractors with similar appearance.

Recently, memory-based methods [56, 37, 7, 38, 32] dominate this field credited to their simplicity yet competitive performance. The core idea of the memory-based methods is to perform dense matching between *query* (*i.e*., current frame) and *memory* (*i.e*., past frames with given or segmented masks), and to retrieve the constructed memory bank in a pixel-level manner. Despite their promising results,

*Equal contribution
†Corresponding author

these methods neglect the fact that videos carry rich temporal information (*e.g*., the target object *ball* moves over time in Figure 1 (a)) yet redundant spatial information. In this case, direct pixel-level global matching forces each query pixel to retrieve all memory pixels equivalently across space and time, leading to ambiguous correspondences that suffer from superfluous spatial information, and are fragile to the movement of objects and cameras ascribed to contempt for temporal information (*i.e*., trajectory). To make matters worse, the temporal information will be further diluted when the memory frames gradually increase as the video progresses, leading to sub-optimal results. Therefore, it is highly desirable to characterize the temporal correspondence from the VOS task, that is, aggregate the trajectory features of the target object *ball* from all relevant memory frames for segmentation.

In this paper, we aim to reconcile the inherent tension of spatial and temporal information to retrieve memory frame information along the object trajectory. Specifically, we design a novel **T**rajectory **M**emory **R**etrieval **N**etwork (**TMRN**) that can be applied as a generic plugin, including a spatial alignment module and a temporal aggregation module to equip with the trajectory information for robust VOS. In specific, as shown in Figure 1 (a), we enable each query pixel to independently retrieve the pixels in each memory frame to seek the location of the counterpart trajectory, and obtain spatially *aligned* memory pixel features (resides in the same spatial position as the corresponding query ones). Then the resultant aligned memory pixels are pooled through the temporal *aggregation* module to reason about inter-frame connections (*i.e*., the contribution of each memory frame) in an adaptive manner, please refer to figure 7. However, directly employing pairwise pixel-level correlation in the spatial alignment module tends to struggle to distinguish the objects with similar appearances (*e.g*., color), increasing the risk of false matches. As shown in Figure 1 (b), due to the distractor *jersey* with similar appearances in memory frame, $p_1$ in the query frame situated on the *cap* is erroneously closer to $p_3$ located in the *jersey* than counterpart $p_2$ in the memory frame.

To mitigate the false matching problem, we carefully design the spatial alignment module through a set of representative reference features (referred to as *agents*) to rectify direct pairwise pixel-level correlation. The main idea is, for each pixel from the query or memory frame, we can obtain the agent-level correlation (*i.e*., a likelihood vector) by comparing this pixel with a set of agents. In essence, the agent-level correlation reflects the consensus among representative agents with a broader receptive field, thus it encodes the relative semantic comparability of the agents that can be relied upon. Intuitively, each pair of pixels with true correspondence (*e.g*., the $p_1$-$p_2$ pair in Figure 1 (b)) from the query and memory frame should be not only visually similar to each other (*i.e*., high pairwise pixel-level correlation), but also holding consensus to any other agents (*i.e*., similar agent-level correlation pair). Based on this correlation consistency in spatial alignment module, false matches caused by similar vision but dissimilar agent correlations will be suppressed (*e.g*., the point $p_1$-$p_3$ pair in Figure 1 (b)), ensuring that true pixel-level correlations between query-memory frame enjoy higher weights in pursuit of spatially well-aligned memory features.

However, it is non-trivial to attain the appropriate agents without any supervision signals for training. Intuitively, the agents should resonate favorably with diverse semantic cues from both query and memory pixels with a wide range of semantic contrast descriptive. In other words, the matching between query-memory pixels in the spatial alignment module based on agent-level correlation should preserve as much critical information as possible in the original pixel-level correlation. Therefore, we take advantage of the singular value decomposition (SVD) to obtain diverse and complementary agents benefiting from the inbuilt rapid decay properties of the singular value, considering the sum of the squares of the singular values after singular value decomposition can be regarded as the energy of the matrix (*i.e*., the representative information contained in the original pixel-level correlation).

In this work, our contributions can be summarized as follows: (1) We design a novel and coherent Trajectory Memory Retrieval Network (TMRN) that can be applied as a generic plugin, including a spatial alignment module and a temporal aggregation module to equip with the trajectory information in VOS. To the best of our knowledge, this is the first work to characterize the temporal correspondence which is in line with the nature of video in a data-driven manner. (2) We elegantly customize the spatial alignment module by coupling SVD initialization with agent-level correlation for representative agents construction and rectifying false matches caused by direct pairwise pixel-level correlation, respectively. (3) Extensive experimental results on challenging benchmarks including DAVIS 2017 validation / test and Youtube-VOS 2018 / 2019 demonstrate that our TMRN, as a general plugin module, achieves consistent improvements over several leading methods.

## 2. Related Work

In this section, we introduce several lines of research in semi-supervised VOS, and describe the memory-based methods in detail.

**Semi-supervised Video Object Segmentation.** Existing VOS methods can be roughly categorized into two categories attributed to the development of deep learning [47, 41, 46, 27, 42, 26, 40]: online-learning methods and offline-learning methods. For online-learning methods [1, 8, 10, 43, 28, 48], the optimal parameters are derived by fine-tuning the model for each video sequence in the inference stage. Xiao *et*

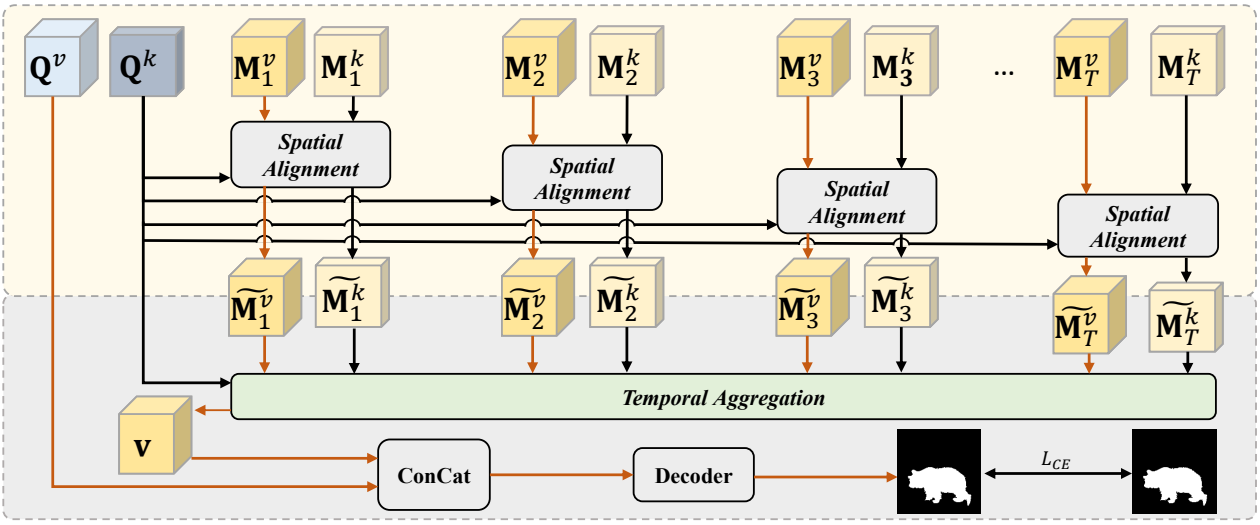



Figure 2: Illustration of the proposed TMRN. TMRN is mainly composed of a spatial alignment module and a temporal aggregation module to equip with the trajectory information, and enables each query pixel to independently retrieve the pixels in each memory frame to seek the location of the counterpart trajectory, and obtain spatially aligned memory pixel features. Then the resultant aligned memory pixels are pooled through the temporal aggregation module to reason about inter-frame connections.

*al.* [48] attempt to make a base segmentation model adapt to new videos by training a meta-learner. However, the online-learning methods require considerable time and are inappropriate for most practical applications.

The offline-learning paradigm aims to make the model trained on the whole training sequences be able to segment any input video without fine-tuning. Generally, there are two common solutions, including mask propagation and pixel-wise matching. Mask propagation based methods [17, 2, 60, 31] leverage the temporal motion consistency to propagate the segmentation mask to the current frame. However, since the propagation is conducted in a short-time interval, these methods are prone to error accumulation under certain conditions such as occlusion. And for the matching-based methods [15, 16, 58, 3], they calculate the correspondences between the current frame and the reference ones for segmentation. CFBI [54] utilizes foreground-background integration for segmentation. Recently, STM [32] demonstrates promising results and is a pioneer work for memory-based methods. Our work follows the memory-based methods due to its simplicity yet competitive results, and attempts to characterize the temporal correspondence which is in line with the nature of video, mitigating the inherent limitations of existing methods.

**Memory-based Video Object Segmentation.** The typical pipeline for memory-based approaches is that given a ground-truth mask at the first frame, we can extract a query frame feature which is compared with the memory features in the constructed memory bank to obtain the correspondences for mask prediction. A series of works [57, 53, 25, 49, 14] aim to improve segmentation performance in following aspects. (1) Apply the memory mechanism to other tasks such as interactive VOS [6, 33] or video object tracking [11]. (2) Reduce the size of the memory bank for a faster inference [22, 44, 19, 9]. (3) Make the model can segment multiple target objects simultaneously [55, 12, 56]. For instance, AOT [55] introduces a association mechanism to segment multiple objects simultaneously. (4) Conduct more reasonable ways for effective and robust memory readout [37, 23, 24, 34]. For example, STCN [7] utilizes the negative squared Euclidean distance instead of inner-product to compute the affinities for exploiting the rich memory information. XMem [5] incorporates multiple independent yet deeply-connected feature memory stores. However, these methods neglect the fact that videos carry rich temporal information (trajectory) yet redundant spatial information. Some methods [49, 57] attempt to explicitly model trajectories by introducing external knowledge at the cost of considerable model complexity, besides, they tend to accumulate errors as the video progresses due to noisy optical flow. In contrast, we characterize the temporal correspondence which is in line with the nature of video in a data-driven manner.

## 3. Our Method

### 3.1. Overview

Inspired by STM [32], the memory-based methods show superiority in VOS task and enjoy a dominant position. Typi-

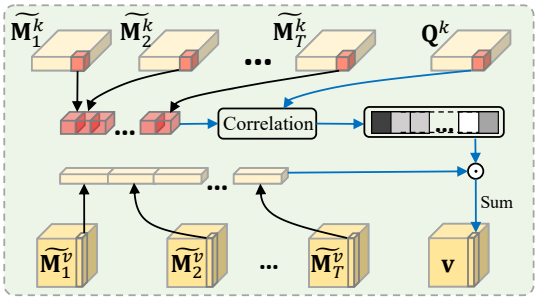



Figure 3: Illustration of the temporal alignment module. The resultant aligned memory pixels from spatial alignment module are pooled to reason about inter-frame connections (*i.e.*, the contribution of each memory frame) in an adaptive manner.

cally, they construct a memory to store the processed frames with the predicted or given masks. The current frame is segmented by retrieving information from the memory. In specific, each memory frame with corresponding mask is encoded as $\mathbf{M}_t^k \in \mathbb{R}^{h\times w\times C_k}$ and $\mathbf{M}_t^v \in \mathbb{R}^{h\times w\times C_v}$, where $h$ and $w$ denote the height and width of the feature map, $C_k$ and $C_v$ denote the channel number of the key feature map and value feature map, respectively. In this way, we can get the memory key $\mathcal{M}^k = \{\mathbf{M}_t^k\}_{t=1}^T \in \mathbb{R}^{T\times h\times w\times C_k}$ and memory value $\mathcal{M}^v = \{\mathbf{M}_t^v\}_{t=1}^T \in \mathbb{R}^{T\times h\times w\times C_v}$ containing $T$ memory frames (suppose the desired segmented frame is at $T+1$ time). The query frame (*i.e.*, current frame) is also encoded as $\mathbf{Q}^k \in \mathbb{R}^{h\times w\times C_k}$ and $\mathbf{Q}^v \in \mathbb{R}^{h\times w\times C_v}$.

### 3.2. Trajectory Memory Retrieval

In typical memory retrieval operation, all memory pixels across space and time are treated equivalently:

$$s_{i,j} = \frac{\exp(\beta_{i,j})}{\sum_{j=1}^{Thw}\exp(\beta_{i,j})},\ \beta_{i,j} = \mathrm{sim}(\mathbf{Q}_i^k, \mathcal{M}_j^k), \quad (1)$$

$$\mathbf{v}_i = \sum_{j=1}^{Thw} s_{i,j}\mathcal{M}_j^v, \quad (2)$$

where $i = 1, 2, ..., hw$ and $\mathrm{sim}(\cdot,\cdot)$ denotes similarity function. A decoder takes the concatenation of the retrieved value $\mathbf{v}$ and $\mathbf{Q}^v$ as input and outputs the predicted mask for current frame. Note that the above correlations are normalized across both space and time, leading to ambiguous correspondences that suffer from superfluous spatial information.

We are dedicated to improve the memory retrieval operation and we argue that each memory frame should be spatially aligned with the current frame before temporal aggregation as illustrated in Figure 2. In specific, we align each memory frame $\mathbf{M}^\star$, $\star \in \{k, v\}$ (omit the subscript $t$ for convenience) with query frame $\mathbf{Q}^k$ by:

$$s_{i,j} = \frac{\exp(\beta_{i,j})}{\sum_{j=1}^{hw}\exp(\beta_{i,j})},\ \beta_{i,j} = \mathrm{sim}(\mathbf{Q}_i^k, \mathbf{M}_j^k), \quad (3)$$

$$\widetilde{\mathbf{M}}_i^\star = \sum_{j=1}^{hw} s_{i,j}\mathbf{M}_j^\star, \quad (4)$$

where $\widetilde{\mathbf{M}}^\star$ denote the spatially aligned memory feature. Note that the correlations are normalized spatially and unrelated to time. Intuitively, this operation seeks the correspondence location of $\mathbf{Q}_i^k$ in memory frame $t$ and reconstruct the memory $\widetilde{\mathbf{M}}_i^\star$ using the feature of these locations, which is why it is called spatial alignment. Then $\mathbf{v}$ can be obtained by temporal aggregation as shown in Figure 3:

$$s_{i,t} = \frac{\exp(\beta_{i,t})}{\sum_{t=1}^{T}\exp(\beta_{i,t})},\ \beta_{i,t} = \mathrm{sim}(\mathbf{Q}_i^k, \widetilde{\mathbf{M}}_{i,t}^k), \quad (5)$$

$$\mathbf{v}_i = \sum_{t=1}^{T} s_{i,t}\widetilde{\mathbf{M}}_{i,t}^v. \quad (6)$$

Note that this aggregation operation is performed alone the temporal axis at one specific spatial position $i$. Thanks to the previous spatially alignment operation, such temporal aggregation is implicitly equivalent to retrieving the memory along the trajectory.

However, the direct pairwise pixel-level correlation $s_{i,j}$ calculated by Equation 3 is fragile to the distractor and at risk of false matches. To mitigate the false matching problem, we carefully design agent-level correlation mechanism to rectify the pixel-level correlation. It is non-trivial to attain the appropriate agents without any supervision signals and we resort to the singular value decomposition. The improved spatial alignment is shown in Figure 4 and detailed in the following section.

### 3.3. Spatial Alignment

**Agents Initialization.** Intuitively, the agents should resonate favorably with diverse semantic cues from both query and memory pixels. we take advantage of Singular Value Decomposition (SVD) to implement principal component analysis on the basis of original pixel-wise correlation matrix $\mathbf{S} = \{s_{i,j}\}_{i,j=1}^{hw} \in \mathbb{R}^{hw\times hw}$. Specifically, we decompose the $\mathbf{S}$ via SVD and only keep the largest $K$ singular values:

$$\mathbf{S} \overset{\substack{\mathrm{SVD}\\ \mathrm{Top}\text{-}K}}{=\!=} \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}, \quad (7)$$

where $\mathbf{U} \in \mathbb{R}^{hw\times K}$, $\mathbf{\Sigma} \in \mathbb{R}^{K\times K}$, $\mathbf{V}^\mathsf{T} \in \mathbb{R}^{K\times hw}$. Benefiting from the inbuilt rapid decay properties of the singular value, keeping the largest $K$ singular values is enough to retain the representative information contained in the original pixel-level correlation matrix.

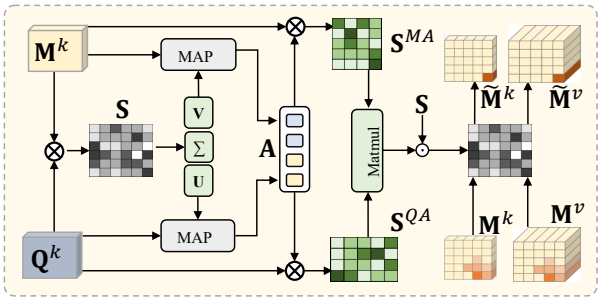



Figure 4: Illustration of the spatial alignment module. In this module, We elegantly customize the spatial alignment module by coupling SVD initialization with agent-level correlation for representative agents construction and rectifying false matches caused by direct pairwise pixel-level correlation, respectively.

For the left singular matrix $\mathbf{U}$, it can be seen as $K$ orthogonal bases in the space of query feature $\mathbf{Q}^k$, and then we explicitly map the $\mathbf{Q}^k$ in the form of linear transformation to get $K$ agents:

$$\mathbf{A}^Q = \mathbf{U}^\top \mathbf{Q}^k. \tag{8}$$

Another $K$ agents in the space of memory feature can be obtained in the same way:

$$\mathbf{A}^M = \mathbf{V}^\top \mathbf{M}^k. \tag{9}$$

**Agent-level Correlation.** After getting the agents $\mathbf{A} = [\mathbf{A}^Q, \mathbf{A}^M] \in \mathbb{R}^{2K \times C_k}$ with a wide range of semantic contrast descriptive, we can calculate the query-agent correlation $\mathbf{s}_i^{QA} \in \mathbb{R}^{1 \times 2K}$ and the memory-agent correlation $\mathbf{s}_j^{MA} \in \mathbb{R}^{1 \times 2K}$ by:

$$\mathbf{s}_i^{QA} = \text{softmax}(\frac{\mathbf{Q}_i^k \mathbf{A}^\top}{\sqrt{C_k}}). \tag{10}$$

$$\mathbf{s}_j^{MA} = \text{softmax}(\frac{\mathbf{M}_j^k \mathbf{A}^\top}{\sqrt{C_k}}). \tag{11}$$

Intuitively, each pair of pixels with true correspondence from the query and memory frame should be not only has high pairwise pixel-level correlation $s_{i,j}$, but also holding similar agent-level correlation. Thus, we calculate the similarity between the agent-level correlations by:

$$c_{i,j} = \mathbf{s}_i^{QA} \mathbf{s}_j^{MA\top}, \tag{12}$$

which is used to rectify the direct pairwise pixel-level correlation. Then the aligned memory frame feature can be obtained similar to Equation 4:

$$\widetilde{\mathbf{M}}_i^\star = \sum_{j=1}^{hw} c_{i,j} \cdot s_{i,j} \mathbf{M}_j^\star. \tag{13}$$

The subsequent processing is consistent with the Section 3.2.

# 4. Experiments

In this section, we construct experiments on widely used multi-object benchmarks including DAVIS 2017 validation / test [36, 35] and Youtube-VOS 2018 / 2019 [50] to evaluate our TMRN. For DAVIS, we follow the official metrics and adopt the region similarity $\mathcal{J}$, the contour accuracy $\mathcal{F}$ and the averaged score $\mathcal{J}\&\mathcal{F}$ for comparison. For YouTube-VOS, we measure the area similarity ($\mathcal{J}_\mathcal{S}$, $\mathcal{J}_\mathcal{U}$) and the contour accuracy ($\mathcal{F}_\mathcal{S}$, $\mathcal{F}_\mathcal{U}$) for the seen object categories and the unseen ones separately, and finally the averaged overall score $\mathcal{G}$ can be attained. Note that we use the official evaluation servers or toolkits to obtain all the scores.

## 4.1. Implementation Details

Our TMRN can be integrated into existing VOS methods as a generic plugin, and we verify the effectiveness of our model on representative three baselines, including STM [32], XMem [5] and STCN [7]. In specific, for STM [32] and STCN [7], we prepend the TMRN to improve the memory reading module (MRM), while for XMem [5], our model is inserted into the working memory reading mechanism. All the rest of the network architecture including memory frame encoder and query frame encoder, and training settings are exactly the same as the baselines. For XMem and STCN, TMRN is implemented on the memory features and query features extracted by ResNet50 and ResNet18 [13] with stride 16 respectively. While both the memory and query features of STM are encoded by ResNet50. All baselines undergo two-stage training, including static image pretraining [45, 39, 59, 4, 20] and video data main training [36, 51]. During inference, we construct memory frames with a sampling interval of 5. Please refer to supplementary material for more details.

## 4.2. Comparison with State-of-the-art Methods

**Quantitative Results.** We verify the effectiveness of TMRN on the DAVIS 2017 val / test [36, 35] and Youtube-VOS 2018 / 2019 [50] sets. (1) **DAVIS** is a densely annotated video object segmentation, The validation and test sets contain 60 and 30 videos, respectively. Table 1 tabulates the performance comparison with and without TMRN on three baseline methods. we consistently observe that our TMRN achieves consistent improvements over all baselines for all metrics, which strongly proves the effectiveness of our method. In specific, STM [32] with TMRN significantly outperforms the corresponding baseline (STM), achieving a large margin of 2.2%/3.1% in $\mathcal{J}\&\mathcal{F}$ for DAVIS 17 val / test. Besides, the introduction of TMRN has a clear lead of 1.0% in $\mathcal{J}\&\mathcal{F}$ for DAVIS 17 val compared to the best memory-based method (XMem [5]). (2) **YouTube-VOS** is a large-scale benchmark for multi-object VOS which provides more training and validation data than DAVIS. For the 2018 version, its validation

Table 1: The quantitative evaluation on multi-object benchmarks, including Youtube-VOS 2018 / 2019 [50] and DAVIS 2017 validation / test [36, 35]. The best results are shown in bold.

| | YouTube-VOS 2018 Val | | | | | YouTube-VOS 2019 Val | | | | | DAVIS-17 test | | | DAVIS-17 val | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | $\mathcal{G}$ | $\mathcal{J}_S$ | $\mathcal{F}_S$ | $\mathcal{J}_U$ | $\mathcal{F}_U$ | $\mathcal{G}$ | $\mathcal{J}_S$ | $\mathcal{F}_S$ | $\mathcal{J}_U$ | $\mathcal{F}_U$ | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ |
| KMN[ECCV20] [37] | 81.4 | 81.4 | 85.6 | 75.3 | 83.3 | - | - | - | - | - | 77.2 | 74.1 | 80.3 | 82.8 | 80.0 | 85.6 |
| CFBI[ECCV20] [54] | 81.4 | 81.1 | 85.8 | 75.3 | 83.4 | 81.0 | 80.6 | 85.1 | 75.2 | 83.0 | 76.6 | 73.0 | 80.1 | 81.9 | 79.3 | 84.5 |
| JOINT[ICCV2021] [29] | 83.1 | 81.5 | 85.9 | 78.7 | 86.5 | 82.8 | 80.8 | 84.8 | 79.0 | 86.6 | - | - | - | 78.6 | 76.0 | 81.2 |
| HMMN[ICCV21] [38] | 82.6 | 82.1 | 87.0 | 76.8 | 84.6 | 82.5 | 81.7 | 86.1 | 77.3 | 85.0 | 78.6 | 74.7 | 82.5 | 84.7 | 81.9 | 87.5 |
| AOT[NIPS21] [55] | 84.5 | 84.3 | 89.3 | 77.9 | 86.4 | 84.5 | 84.0 | 88.8 | 78.4 | 86.7 | 81.2 | 77.3 | 85.1 | 85.4 | 82.4 | 88.4 |
| RPCM[AAAI22] [52] | 84.0 | 83.1 | 87.7 | 78.5 | 86.7 | 83.9 | 82.6 | 86.9 | 79.1 | 87.1 | 79.2 | 75.8 | 82.6 | 83.7 | 81.3 | 86.0 |
| SITVOS[AAAI22] [18] | 81.3 | 79.9 | 76.4 | 84.3 | 84.4 | - | - | - | - | - | - | - | - | 83.5 | 80.4 | 86.5 |
| AOC[MM22] [53] | 84.0 | 82.7 | 78.8 | 87.4 | 87.1 | 84.1 | 82.7 | 79.4 | 76.9 | 87.2 | 79.3 | 74.7 | 83.9 | 83.8 | 81.7 | 85.9 |
| PerClip[CVPR22] [34] | 84.6 | 83.0 | 88.0 | 79.6 | 87.9 | 84.6 | 82.6 | 87.3 | 80.0 | 88.3 | - | - | - | 86.1 | 83.0 | 89.2 |
| GSFM[ECCV22] [23] | 83.8 | 82.8 | 87.5 | 78.3 | 86.5 | - | - | - | - | - | 77.5 | 74.0 | 80.9 | 86.2 | 83.1 | 89.3 |
| RDE[CVPR22] [19] | - | - | - | - | - | 83.3 | 81.9 | 86.3 | 78.0 | 86.9 | 78.9 | 74.9 | 82.9 | 86.1 | 82.1 | 90.0 |
| SWEM[CVPR22] [22] | 82.8 | 82.4 | 86.9 | 77.1 | 85.0 | - | - | - | - | - | - | - | - | 84.3 | 81.2 | 87.4 |
| QDMN[ECCV22] [24] | 83.8 | 82.7 | 87.5 | 78.4 | 86.4 | - | - | - | - | - | 81.9 | 78.1 | 85.4 | 85.6 | 78.1 | 85.4 |
| TBD[ECCV22] [9] | 80.5 | 79.4 | 75.5 | 83.8 | 83.2 | - | - | - | - | - | 69.4 | 66.6 | 72.2 | 80.0 | 77.6 | 82.3 |
| STM [32] | 79.4 | 79.7 | 84.2 | 72.8 | 80.9 | 79.2 | 79.6 | 83.6 | 73.0 | 80.6 | 72.2 | 69.3 | 75.2 | 81.8 | 79.2 | 84.3 |
| STM w/ TMRN | **81.5** | **81.5** | **86.5** | **74.8** | **83.2** | **81.3** | **81.9** | **85.6** | **75.3** | **82.4** | **75.3** | **72.0** | **78.7** | **84.0** | **81.1** | **86.8** |
| XMem [5] | 85.7 | 84.6 | 89.3 | 80.2 | 88.7 | 85.5 | 84.3 | 88.6 | 80.3 | 88.6 | 81.0 | 77.4 | 84.5 | 86.2 | 82.9 | 89.5 |
| XMem w/ TMRN | **86.4** | **85.5** | **90.4** | **80.5** | **89.2** | **85.9** | **84.8** | **89.0** | **80.5** | **89.1** | **81.5** | **77.7** | **85.2** | **87.2** | **83.8** | **90.6** |
| STCN [7] | 83.0 | 81.9 | 86.5 | 77.9 | 85.7 | 82.7 | 81.1 | 85.4 | 78.2 | 85.9 | 76.1 | 72.7 | 79.6 | 85.4 | 82.2 | 88.6 |
| STCN w/ TMRN | **84.2** | **82.8** | **87.9** | **79.2** | **86.9** | **84.1** | **82.6** | **87.0** | **79.1** | **87.5** | **78.2** | **74.2** | **73.3** | **87.0** | **83.6** | **90.4** |

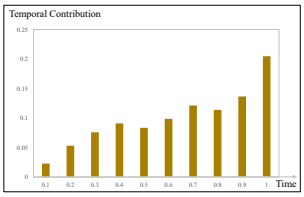
Table 2: Evaluation of the effectiveness of different components on DAVIS 2017 validation set. ST denotes the bald spatial alignment module coupled with temporal aggregation module to model the trajectory, that is, TMRN without SVD initialization (SVD) and agent-level correlation (Agent).

| Configuration | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ |
|---|---|---|---|
| Baseline | 85.4 | 82.2 | 88.6 |
| Baseline+ST | 86.1 | 83.0 | 89.2 |
| Baseline+ST+Agent | 86.5 | 83.6 | 89.4 |
| Baseline+ST+Agent+SVD (**TMRN**) | 87.0 | 84.2 | 89.8 |

Table 3: Different strategies for agent construction.

| | $\mathcal{J}\&\mathcal{F}$ | $\mathcal{J}$ | $\mathcal{F}$ |
|---|---|---|---|
| All | 85.9 | 82.9 | 88.9 |
| Rand | 86.6 | 84.1 | 89.1 |
| Top-$K$ | 86.6 | 83.9 | 89.3 |
| **SVD** | 87.0 | 84.2 | 89.8 |

Figure 5: Statistical distribution of memory frame contributions.




set contains 474 videos, including 65 training (seen) categories and 26 unseen ones. And the 2019 version further expands the number of videos to 507. As summarized in Table 1, Our TMRN all surpasses the corresponding baselines respectively (*e.g.*, 1.2%/1.4% in $\mathcal{G}$ for STCN on YouTube 18/19), which further confirms the effectiveness of our model to characterize the temporal correspondence and is more sensible in dealing with complex video scenes.

**Qualitative Comparison.** Figure 6 showcases qualitative comparison between STCN w/ TMRN and other competitive methods including STM [32], GSFM [23], and STCN [21]. We can observe that STM and STCN fail to predict target objects when multiple similar objects *human* have appeared. Benefiting from the inherent property that agent-level correlation in the spatial alignment module can alleviate false matches caused by direct pixel-level correlation, our method yields more precise segmentation. Besides, compared to the baseline STCN, we achieve better consistent segmentation results credited to modeling the temporal trajectory in a data-driven manner. Please refer to supplementary material for more qualitative results.

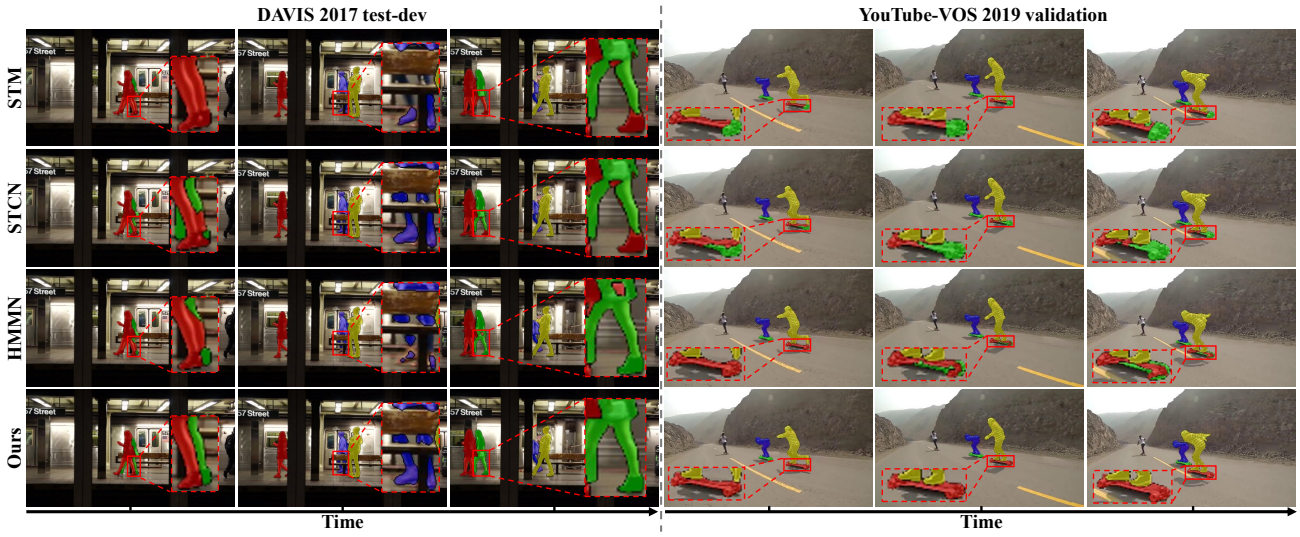



Figure 6: Qualitative comparison on DAVIS 2017 test-dev set. And we mark significant improvements using red boxes. Zoom for better view.

Table 4: Evaluation of the hyperparameters $K$.

| K | 64 | 96 | 128 | 160 | 192 |
|---|---|---|---|---|---|
| $\mathcal{J}\&\mathcal{F}$ | 86.1 | 86.5 | 87.0 | 86.9 | 86.4 |

### 4.3. Ablation Study and Analysis

To look deeper into our method, we perform a series of ablation studies on DAVIS 2017 validation following [19, 9] to analyze each component of TMRN, and the baseline method is STCN [7].

**Effectiveness of Trajectory Modeling.** From the comparison between the $1^{st}$ row and the $2^{nd}$ row of Table 2, We find that the introduction of trajectory modeling achieves clear performance gains (*i.e.*,0.7% in $\mathcal{J}\&\mathcal{F}$ ) even without customizing the spatial alignment module. We conclude that the performance gain comes from retrieving memory frame information along the object trajectory, which is in line with the nature of video.

**Effectiveness of Agent-level Correlation.** The addition of the agent-level correlation in spatial alignment module also contributes to a remarkable performance gain compared with the $3^{rd}$ in Table 2. The improvements can be mainly ascribed to the proposed agent-level correlation that can effectively rectify direct pairwise pixel-level correlation and ensure that true pixel-level correlations between query-memory frame enjoy higher weights.

**Effectiveness of SVD Initialization.** With the utilization of the SVD to construct agents, further improvements can be observed, *e.g.*, 0.5% $\mathcal{J}\&\mathcal{F}$. This proves that the singular value decomposition (SVD) can attain diverse and complementary agents benefiting from the inbuilt rapid decay properties of the singular value, and laying a good foundation for agent-level correlation ($3^{rd}$ *vs.* $4^{th}$ in Table 2).

**Analysis of Agent Construction.** To explore effectiveness of different strategies to construct agents for subsequent agent-level correlation, we conduct experiments in Table 3, where *All* denotes grabbing all pixels from the memory and query frame respectively, *Rand* refers to randomly sample $K$ pixel features, and *Top-K* means select top $K$ features conditioned on the cumulative correlation matrix along the memory and query dimension respectively. We can vividly observe that the inappropriate construction strategy will make the agents full of noise or incompleteness, leading to performance decay. While the strategy of SVD achieves the best results, which is in line with our design purpose, that is, representative agents can enjoy synergy with subsequent agent-level correlations.

**Analysis of Temporal Aggregation.** To vividly present the working mechanism of the temporal aggregation module, we visualize the contribution of each memory frame to the current frame for segmentation and normalize the time dimension, as illustrated in Figure 5. We can find an interesting fact that the segmentation of the current frame is more related to the adjacent memory frames at the statistical level, which is consistent with our intuition considering the inherent temporal smoothness of video.

**Hyperparameter Evaluations.** As shown in Table 4, we evaluate how $K$ affects our model learning. we can observe that the performance continues to grow until $K = 128$, We deem the main reason is too few agents cannot represent diverse semantic clues, while too many agents will lead to undesirable redundancy.

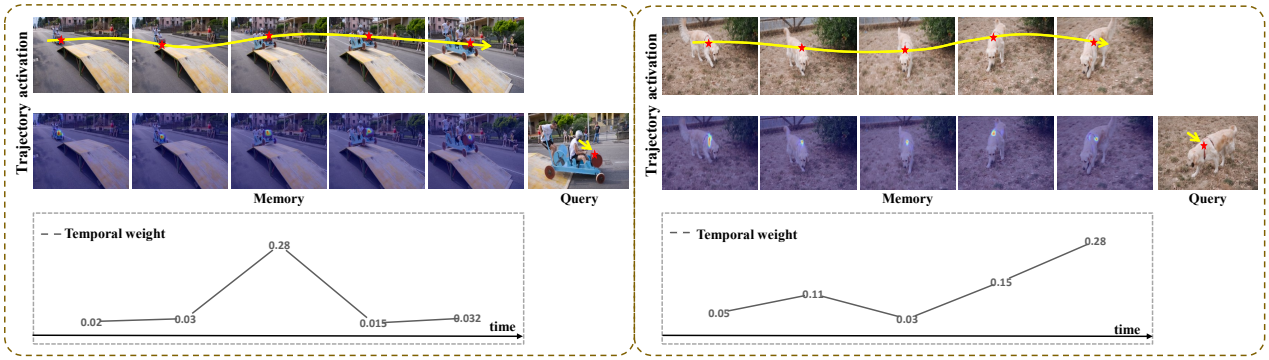



Figure 7: Visualization of trajectory. From top to bottom: (1) We see that the TMRN has the ability to retrieve memory frame information along the object trajectory (yellow arrow). (2) We visualize the spatial alignment module (*i.e.*, activation map of spatial location of trajectory). (3) We visualize the temporal aggregation module (*i.e.*, contribution of each memory frame).

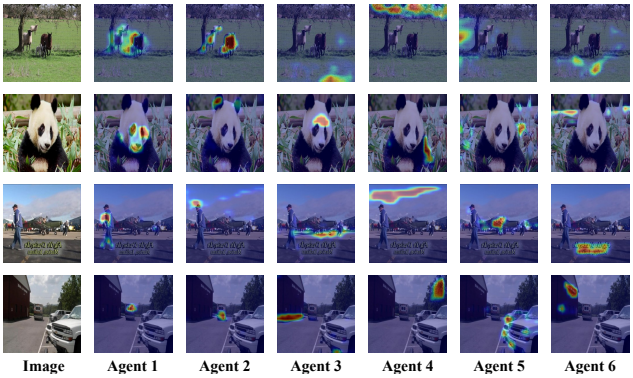



Figure 8: Visualization of target object-activated agents for better illustration. As we can see, the well-constructed agents resonate favorably with diverse semantic cues with a wide range of semantic contrast descriptive. Zoom for better view.


Figure 9: Visualization of direct pairwise pixel-level correspondence. The green and red arrows point to the top five memory pixels that match the query pixel with and without agent-level correlation, respectively. As we can see, the top five pixels corresponding to the query points tend to line in the corresponding location of memory frame thanks to the agent-level correlation. Zoom for better view.

### 4.4. Visualization and Analysis

**Visualization of Trajectory.** To qualitatively evaluate the effect of characterizing the trajectory, We visualize the spatial alignment module (*i.e.*, activation map of spatial location of trajectory) and the temporal aggregation module (*i.e.*, contribution of each memory frame) separately, summarized in Figure 7. We can find that the some well-matched trajectory segments occupies larger weight (*e.g.*, $3^{rd}$ row), while some trajectory segments with large pose differences caused by the movement of object *trolley* are assigned with smaller weights. This proves that our TMRN can seek the location of each memory frame and reason about inter-frame connections (*i.e.*, the contribution of each memory frame) in an adaptive manner.

**Visualization of Constructed Agents.** We visualize the target object-activated agents for better illustration in Figure 8. As we can see, the well-constructed agents resonate favorably with diverse semantic cues with a wide range of semantic contrast descriptive. This also confirms the effectiveness of singular value decomposition (SVD) which can obtain diverse and complementary agents benefiting from the inbuilt rapid decay properties of the singular value.

**Visualization of the Pixel-level Correspondence.** To vividly present the effect of agent-level correlation, we visualize differences in pixel correspondences according to whether agent-level correlation exits. As shown in Figure 9, with the utilization of agent-level correlation, the top five pixels corresponding to the query points tend to line in the corresponding location of memory frame. While these ones will contain large noises without the agent-level correlation to perform direct pairwise pixel-level matching. This is in line with the design idea, *i.e.*, ensuring that true pixel-level correlations between query-memory frame enjoy higher

weights in pursuit of spatially well-aligned memory features.

## 5. Conclusion

In this paper, we propose a a novel and coherent Trajectory Memory Retrieval Network (TMRN) that can be applied as a generic plugin, including a spatial alignment module and a temporal aggregation module to equip with the trajectory information in VOS. Besides, We customize the spatial alignment module by coupling SVD initialization with agent-level correlation for representative agents construction and rectifying false matches caused by direct pairwise pixel-level correlation, respectively. Extensive experimental results on challenging benchmarks show effectiveness.

## 6. Acknowledgments

This work was partially supported by the National Nature Science Foundation of China (Grant 62022078, 62021001), National Defense Basic Scientific Research Program (Grant JCKY2021130B016).

## References

[1] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *CVPR*, pages 221–230, 2017. 2

[2] Xi Chen, Zuoxin Li, Ye Yuan, Gang Yu, Jianxin Shen, and Donglian Qi. State-aware tracker for real-time video object segmentation. In *CVPR*, pages 9384–9393, 2020. 3

[3] Yuhua Chen, Jordi Pont-Tuset, Alberto Montes, and Luc Van Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In *CVPR*, pages 1189–1198, 2018. 3

[4] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: toward class-agnostic and very high-resolution segmentation via global and local refinement. In *CVPR*, pages 8890–8899, 2020. 5

[5] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, pages 640–658. Springer, 2022. 3, 5, 6

[6] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *CVPR*, pages 5559–5568, 2021. 3

[7] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *NIPS*, 2021. 1, 3, 5, 6, 7

[8] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *ICCV*, pages 686–695, 2017. 2

[9] Suhwan Cho, Heansung Lee, Minhyeok Lee, Chaewon Park, Sungjun Jang, Minjung Kim, and Sangyoun Lee. Tackling background distraction in video object segmentation. In *ECCV*, pages 446–462. Springer, 2022. 3, 6, 7

[10] Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. Capsulevos: Semi-supervised video object segmentation using capsule routing. In *ICCV*, pages 8480–8489, 2019. 2

[11] Zhihong Fu, Qingjie Liu, Zehua Fu, and Yunhong Wang. Stmtrack: Template-free visual tracking with space-time memory networks. In *CVPR*, pages 13774–13783, 2021. 3

[12] Wenbin Ge, Xiankai Lu, and Jianbing Shen. Video object segmentation using global and instance embedding learning. In *CVPR*, pages 16836–16845, 2021. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[14] Li Hu, Peng Zhang, Bang Zhang, Pan Pan, Yinghui Xu, and Rong Jin. Learning position and target consistency for memory-based video object segmentation. In *CVPR*, pages 4144–4154, 2021. 3

[15] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Videomatch: Matching based video object segmentation. In *ECCV*, pages 54–70, 2018. 3

[16] Xuhua Huang, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Fast video object segmentation with temporal aggregation network and dynamic template matching. In *CVPR*, pages 8879–8889, 2020. 3

[17] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *CVPR*, pages 2117–2126. IEEE, 2017. 3

[18] Meng Lan, Jing Zhang, Fengxiang He, and Lefei Zhang. Siamese network with interactive transformer for video object segmentation. In *AAAI*, volume 36, pages 1228–1236, 2022. 6

[19] Mingxing Li, Li Hu, Zhiwei Xiong, Bang Zhang, Pan Pan, and Dong Liu. Recurrent dynamic embedding for video object segmentation. In *CVPR*, pages 1332–1341, 2022. 3, 6, 7

[20] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *CVPR*, pages 2869–2878, 2020. 5

[21] Shuxian Liang, Xu Shen, Jianqiang Huang, and Xian-Sheng Hua. Video object segmentation with dynamic memory networks and adaptive object alignment. In *ICCV*, pages 8065–8074, 2021. 6

[22] Zhihui Lin, Tianyu Yang, Maomao Li, Ziyu Wang, Chun Yuan, Wenhao Jiang, and Wei Liu. Swem: Towards real-time video object segmentation with sequential weighted expectation-maximization. In *CVPR*, pages 1362–1372, 2022. 3, 6

[23] Yong Liu, Ran Yu, Jiahao Wang, Xinyuan Zhao, Yitong Wang, Yansong Tang, and Yujiu Yang. Global spectral filter memory network for video object segmentation. In *ECCV*, pages 648–665. Springer, 2022. 3, 6

[24] Yong Liu, Ran Yu, Fei Yin, Xinyuan Zhao, Wei Zhao, Weihao Xia, and Yujiu Yang. Learning quality-aware dynamic memory for video object segmentation. In *ECCV*, pages 468–486. Springer, 2022. 3, 6

[25] Xiankai Lu, Wenguan Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. Video object segmentation with episodic graph memory networks. In *Computer*

*Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 661–679. Springer, 2020. 3

[26] Naisong Luo, Yuwen Pan, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Camouflaged instance segmentation via explicit de-camouflaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17927, 2023. 2

[27] Huayu Mai, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Dualrel: Semi-supervised mitochondria segmentation from a prototype perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19617–19626, 2023. 2

[28] K-K Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. Video object segmentation without temporal information. *PAMI*, 41(6):1515–1530, 2018. 2

[29] Yunyao Mao, Ning Wang, Wengang Zhou, and Houqiang Li. Joint inductive and transductive learning for video object segmentation. In *ICCV*, pages 9670–9679, 2021. 6

[30] King Ngi Ngan and Hongliang Li. *Video segmentation and its applications*. Springer Science & Business Media, 2011. 1

[31] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *CVPR*, pages 7376–7385, 2018. 1, 3

[32] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, pages 9226–9235, 2019. 1, 3, 5, 6

[33] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Space-time memory networks for video object segmentation with user guidance. *PAMI*, (01):1–1, 2020. 3

[34] Kwanyong Park, Sanghyun Woo, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Per-clip video object segmentation. In *CVPR*, pages 1352–1361, 2022. 3, 6

[35] 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 *CVPR*, pages 724–732, 2016. 5, 6

[36] 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. 5, 6

[37] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *ECCV*, pages 629–645. Springer, 2020. 1, 3, 6

[38] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. Hierarchical memory matching network for video object segmentation. In *ICCV*, pages 12889–12898, 2021. 1, 6

[39] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *PAMI*, 38(4):717–729, 2015. 5

[40] Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, and Yongdong Zhang. Lesion-aware transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10938–10947, 2021. 2

[41] Rui Sun, Naisong Luo, Yuwen Pan, Huayu Mai, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Appearance prompt vision transformer for connectome reconstruction. IJCAI, 2023. 2

[42] Rui Sun, Huayu Mai, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Structure-decoupled adaptive part alignment network for domain adaptive mitochondria segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023. 2

[43] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364*, 2017. 2

[44] Haochen Wang, Xiaolong Jiang, Haibing Ren, Yao Hu, and Song Bai. Swiftnet: Real-time video object segmentation. In *CVPR*, pages 1296–1305, 2021. 3

[45] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pages 136–145, 2017. 5

[46] Yuan Wang, Rui Sun, and Tianzhu Zhang. Rethinking the correlation in few-shot segmentation: A buoys view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2023. 2

[47] Yuan Wang, Rui Sun, Zhe Zhang, and Tianzhu Zhang. Adaptive agent transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 36–52. Springer, 2022. 2

[48] Huaxin Xiao, Bingyi Kang, Yu Liu, Maojun Zhang, and Jiashi Feng. Online meta adaptation for fast video object segmentation. *PAMI*, 42(5):1205–1217, 2019. 2, 3

[49] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun. Efficient regional memory network for video object segmentation. In *CVPR*, pages 1286–1295, 2021. 3

[50] 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 *ECCV*, pages 585–601, 2018. 5, 6

[51] 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. 5

[52] Xiaohao Xu, Jinglu Wang, Xiao Li, and Yan Lu. Reliable propagation-correction modulation for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2946–2954, 2022. 6

[53] Xiaohao Xu, Jinglu Wang, Xiang Ming, and Yan Lu. Towards robust video object segmentation with adaptive object calibration. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2709–2718, 2022. 3, 6

[54] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *ECCV*, pages 332–348. Springer, 2020. 3, 6

[55] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *NIPS*, 2021. 3, 6

[56] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. *arXiv preprint arXiv:2210.09782*, 2022. 1, 3

[57] Ye Yu, Jialin Yuan, Gaurav Mittal, Li Fuxin, and Mei Chen. Batman: Bilateral attention transformer in motion-appearance neighboring space for video object segmentation. In *ECCV*, pages 612–629. Springer, 2022. 3

[58] Xiaohui Zeng, Renjie Liao, Li Gu, Yuwen Xiong, Sanja Fidler, and Raquel Urtasun. Dmm-net: Differentiable mask-matching network for video object segmentation. In *ICCV*, pages 3929–3938, 2019. 3

[59] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *ICCV*, pages 7234–7243, 2019. 5

[60] Yizhuo Zhang, Zhirong Wu, Houwen Peng, and Stephen Lin. A transductive approach for video object segmentation. In *CVPR*, pages 6949–6958, 2020. 3

[61] Ziyu Zhang, Sanja Fidler, and Raquel Urtasun. Instance-level segmentation for autonomous driving with deep densely connected mrfs. In *CVPR*, pages 669–677, 2016. 1