

Multi-grained Temporal Prototype Learning for Few-shot Video Object Segmentation

Nian Liu¹ Kepan Nan² Wangbo Zhao³ Yuanwei Liu² Xiwen Yao^{2†}

Salman Khan^{1,4} Hisham Cholakkal¹ Rao Muhammad Anwer¹ Junwei Han² Fahad Shahbaz Khan^{1,5}

¹ Mohamed bin Zayed University of Artificial Intelligence ² Northwestern Polytechnical University

³ National University of Singapore

⁴ Australian National University ⁵ CVL, Linköping University

Abstract

Few-Shot Video Object Segmentation (FSVOS) aims to segment objects in a query video with the same category defined by a few annotated support images. However, this task was seldom explored. In this work, based on IPMT, a state-of-the-art few-shot image segmentation method that combines external support guidance information with adaptive query guidance cues, we propose to leverage multi-grained temporal guidance information for handling the temporal correlation nature of video data. We decompose the query video information into a clip prototype and a memory prototype for capturing local and long-term internal temporal guidance, respectively. Frame prototypes are further used for each frame independently to handle fine-grained adaptive guidance and enable bidirectional clip-frame prototype communication. To reduce the influence of noisy memory, we propose to leverage the structural similarity relation among different predicted regions and the support for selecting reliable memory frames. Furthermore, a new segmentation loss is also proposed to enhance the category discriminability of the learned prototypes. Experimental results demonstrate that our proposed video IPMT model significantly outperforms previous models on two benchmark datasets. Code is available at <https://github.com/nankepan/VIPMT>.

1. Introduction

To mitigate the data-hungry issue of modern deep semantic segmentation models [18, 2, 4], few-shot semantic segmentation emerges by only requiring a few support samples with annotated masks for segmenting the objects of the same class in new images. Many recent works [25, 32, 13, 16] have shown very promising results on image data using the meta-learning scheme. They simulate the inference process and partition the training set into numerous episodes, in each

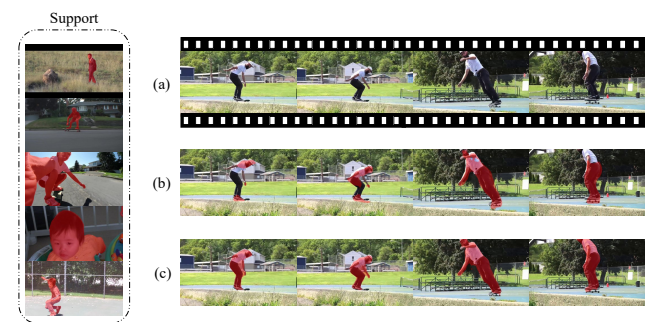


Figure 1. Given a few annotated support frames, FSVOS aims to segment the objects with the same category from a query video (a). Simply considering single-frame information leads to inconsecutive segmentation results (b), while our method generates accurate results by using multi-grained temporal prototypes (c).

of which, the model samples a few support images and learns to guide the segmentation on the query images.

Inspired by the classic Few-Shot Image Semantic Segmentation (FSISS) task, the Few-Shot Video Object Segmentation (FSVOS) task was introduced by [1, 24]. For FSISS, many works adopt the prototype-based methods, in which a prototype vector is extracted from the support to encode the category guidance information, and then a segmentation head learns to match the prototype with the feature at each query pixel for performing query segmentation. However, the intra-class diversity can cause the matching gap between the support-induced prototype guidance and the query features. The IPMT model [16] solved this problem by learning an intermediate prototype that integrates both *support-induced external category guidance knowledge* and *query-induced adaptive guidance information*. As for FSVOS, video data additionally show temporal correlation, from which we can also induce *internal temporal guidance*. If we ignore this prior knowledge and simply consider single-frame information, the learned prototypes may vary significantly among different frames, leading to inconsecutive segmentation results, as shown in Figure 1 (b).

[†]Corresponding author: yaowen517@gmail.com.

In this work, we extend the IPMT model [16] for video data by tackling the temporal correlation nature. Based on the intermediate prototype mechanism of IPMT, we propose to decompose the query video information into the clip-level, frame-level, and memory-level prototypes, consisting of a multi-grained temporal structure, which was *NEVER* explored by existing models [1, 24]. Among them, the clip prototype encodes local temporal object guidance information in each consecutive clip, while the memory prototype introduces long-term historical guidance cues. Combing them can effectively handle the temporal correlation problem. However, such a design may ignore fine-grained per-frame adaptive guidance information, hence may fail to handle large scene changes and object transformation. To this end, we further learn an adaptive frame prototype by independently encoding per-frame object features. Additionally, we enable bidirectional clip-frame prototype communication by making the clip-level and frame-level prototypes initialize each other, hence promoting intra-clip temporal correlation.

To better leverage the historical memory guidance, we also follow a Video Object Segmentation (VOS) method [17] and train an IoU regression network to select reliable memory frames for reducing the negative influence brought by noisy memory. However, different from [17], we explicitly consider the nature of the FSVOS task and propose to leverage the structural similarity relation among different predicted regions and the support information for predicting more accurate IoU scores. To make the learned prototype more category discriminative, we also propose a Cross-Category Discriminative Segmentation (CCDS) loss by leveraging negative batch samples. Extensive experimental results have verified the effectiveness of our proposed **Video IPMT** (VIPMT) model and showed its significant performance improvement over state-of-the-art models.

In conclusion, our contributions can be summarized as follows:

- For the very first time, we propose to learn multi-grained temporal prototypes for FSVOS, by extending the IPMT [16] model. Clip and memory prototypes are learned for internal temporal guidance. Frame prototypes are used for fine-grained adaptive guidance and also enable prototype communication.
- We propose to leverage the structural similarity relation among different predicted regions and the support for selecting reliable memory information. We also present a CCDS loss using the negative samples within each batch for promoting category discriminability of the learned prototypes.
- Experimental results have demonstrated the significant effectiveness of our proposed model, which improves state-of-the-art results by more than 4% and 3%, on two benchmark datasets, respectively.

2. Related Work

2.1. Few-Shot Image Semantic Segmentation

Most previous FSIS works adopted meta-learning-based methods [27], especially prototype-based models. Specifically, Dong and Xing [9] aggregated a prototype vector on the support images to encode the representative category knowledge first, and then evaluated its similarity with each query pixel in a matching network as the segmentation result. Following this idea, Yang *et al.* [29] constructed multiple prototypes from limited support images to represent diverse image regions. In [25], Tian *et al.* used high-level features to generate a prior mask as a supplement to the support prototype for refining the query feature. Liu *et al.* [15] leveraged non-target prototypes to eliminate distracting regions. BAM [13] used the prototypes of base classes to explicitly suppress corresponding regions. In IPMT [16], Liu *et al.* proposed to generate an intermediate prototype which encodes the category guidance information from both support and query. However, all these methods only focused on image data and the challenge of video data remains seldom explored.

2.2. Video Object Segmentation

Another closely related domain is VOS, especially the semi-supervised VOS, in which the mask label of an object in the first frame is given and the model is required to segment the same object in other frames. Some methods [22, 7, 28] tried to leverage optical flow to help segment target objects. Noticing the successful application of memory networks [8, 31] in computer vision, STM [21] proposed a memory mechanism to leverage information from previous frames for segmenting the current frame. Lu *et al.* [19] followed this idea and improve the memory mechanism with a graph model. For a comprehensive survey please refer to [33].

Different from semi-supervised VOS, we tackle the FSVOS task with two differences. First, semi-supervised VOS aims to segment *the same object* indicated in the annotation of *the first frame* while FSVOS requires to segment *the objects of the same class* with the annotated support set. Second, the support set in FSVOS can be composed of *any images or frames of any videos* that contain the target class. Hence, FSVOS is more generalized and faces much larger intra-class diversity between the support set and the query video. In this work, we propose multi-grained temporal prototype learning and bidirectional clip-frame prototype communication for FSVOS, which is different from existing VOS methods.

Our idea of using the IoU network to select memory frames with high segmentation quality is inspired by [17]. However, in [17], the authors directly regress the IoU score by only taking the image and the mask as the network input. In this work, we further consider the nature of FSVOS and propose to compute several structural similarity maps

which explicitly encode the quality assessment prior of the relations among the predicted foreground, background, and the support areas.

2.3. Few-shot Video Object Segmentation

For the FSVOS task, currently, only a few works have targeted this topic. Chen *et al.* [1] proposed the first FSVOS dataset and model. They proposed a Domain Agent Network (DAN) to alleviate the large computational cost of the many-to-many attention between the support images and the query video frames, which only considered clip-level temporal information. Siam *et al.* [24] proposed a temporal transductive inference model, which uses both global and local temporal constraints to obtain per-frame model weights and locally consistent predictions. Compared with them, we optimize multi-grained temporal prototypes while they optimized per-frame model weights. Furthermore, we explicitly use historical memory while they did not.

3. Preliminaries

3.1. Problem Definition

For FSVOS, a whole video object segmentation dataset \mathcal{D} with multiple object categories \mathcal{C} is divided into a training subset \mathcal{D}_{base} and a testing subset \mathcal{D}_{novel} , whose categories are \mathcal{C}_{base} and \mathcal{C}_{novel} , respectively, and satisfy $\mathcal{C}_{base} \cap \mathcal{C}_{novel} = \emptyset$ and $\mathcal{C}_{base} \cup \mathcal{C}_{novel} = \mathcal{C}$. Under the standard few-shot setting and the episodic training scheme, both \mathcal{D}_{base} and \mathcal{D}_{novel} are randomly partitioned into a lot of episodes. In each episode, a support set \mathcal{S} provides K frames \mathbf{I}^s with labeled segmentation masks \mathbf{O}^s of a specific object category (we follow [1, 24] to focus on one-way segmentation here), *i.e.*, $\mathcal{S} = \{(\mathbf{I}^{s_i}, \mathbf{O}^{s_i})\}_{i=1}^K$, as the meta knowledge of the target class. Then, given a query video $\mathcal{Q} = \{\mathbf{I}^{q_i}\}_{i=1}^N$ with N frames, an FSVOS model is used to predict the segmentation masks $\{\tilde{\mathbf{O}}^{q_i}\}_{i=1}^N$ of the same category for each frame. The model can be trained on the episodes sampled from \mathcal{D}_{base} with the given ground truth query masks $\{\mathbf{O}^{q_i}\}_{i=1}^N$, learning how to transfer the guidance knowledge from the support to the query. Finally, the FSVOS model is expected to make accurate predictions for unseen categories on \mathcal{D}_{novel} .

3.2. Review of IPMT

The IPMT model [16] adopts a transformer-based architecture for the FSIS task. Specifically, FSIS can be modeled by learning an intermediate prototype $\mathbf{G} \in \mathbb{R}^{1 \times C}$ and pixel-wise feature maps \mathbf{F}^s and \mathbf{F}^q for the support and query, respectively, and then conduct iterative mutual optimization between \mathbf{G} and \mathbf{F}^q within multiple IPMT layers. Under the K -shot setting, the support features $\mathbf{F}^s \in \mathbb{R}^{Khw \times C}$ and the flattened masks $\mathbf{O}^s \in \mathbb{R}^{Khw \times 1}$, where h, w, C are the height, width, and channel number, respectively.

In the l -th IPMT layer, given the previous query feature $\mathbf{F}_{l-1}^q \in \mathbb{R}^{hw \times C}$ and the flattened binarized mask prediction $\tilde{\mathbf{O}}_{l-1}^q \in \mathbb{R}^{hw \times 1}$, the intermediate prototype \mathbf{G} is updated by

$$\begin{aligned} \mathbf{G}_l &= \text{IPM}(\mathbf{G}_{l-1}, \mathbf{F}^s, \mathbf{F}_{l-1}^q, \mathbf{O}^s, \tilde{\mathbf{O}}_{l-1}^q) \\ &= \text{MLP}(\mathbf{G}_l^s + \mathbf{G}_l^q + \mathbf{G}_{l-1}), \end{aligned} \quad (1)$$

where MLP denotes a multi-layer perception. The \mathbf{G}_l^s can be seen as the *support prototype* which encodes the deterministic guidance knowledge from the support and the \mathbf{G}_l^q can be regarded as the *query prototype* that learns adaptive guidance information from the query. They can be obtained by

$$\mathbf{G}_l^s = \text{MAtt}(\mathbf{G}_{l-1}, \mathbf{F}^s, \mathbf{O}^s), \quad (2)$$

$$\mathbf{G}_l^q = \text{MAtt}(\mathbf{G}_{l-1}, \mathbf{F}_{l-1}^q, \tilde{\mathbf{O}}_{l-1}^q). \quad (3)$$

Here, MAtt means the masked attention operation [3]:

$$\text{MAtt}(\mathbf{G}, \mathbf{F}, \mathbf{O}) = \delta(f_Q(\mathbf{G})f_K(\mathbf{F})^\top + \Delta)f_V(\mathbf{F}), \quad (4)$$

where δ means the softmax normalization, $f_Q(\cdot), f_K(\cdot), f_V(\cdot)$ are three linear transformations following [26], and $\Delta = (1 - \mathbf{O}^\top) \cdot (-\infty)$ is used to modulate the attention matrix, making background attention weights become zeros after softmax.

Then, the updated \mathbf{G}_l is used to generate mask predictions for both support and query via a mask generation (MG) process:

$$\tilde{\mathbf{O}}_l^s = \text{MG}(\mathbf{G}_l, \mathbf{F}^s), \tilde{\mathbf{O}}_l^q = \text{MG}(\mathbf{G}_l, \mathbf{F}_{l-1}^q), \quad (5)$$

$$\text{MG}(\mathbf{G}, \mathbf{F}) = \text{Sigmoid}(f_G(\mathbf{G})\mathbf{F}^\top), \quad (6)$$

where $f_G(\cdot)$ is another linear transformation. At the same time, the generated prototype \mathbf{G}_l is also used to update the query feature maps via query activation (QA):

$$\mathbf{F}_l^q = \text{QA}(\mathbf{G}_l, \mathbf{F}_{l-1}^q) = \mathcal{F}_{actv}(\mathbf{G}_l \odot \mathbf{F}_{l-1}^q), \quad (7)$$

where \odot means concatenation and \mathcal{F}_{actv} is an activation network with two convolution layers.

In the original implementation, IPMT first feeds the input images into a froze backbone encoder (*e.g.* ResNet-50 [11]), obtaining multi-scale features $\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4\}$ from the last four convolution blocks, with the scale of 1/4, 1/8, 1/8, and 1/8, respectively. Then, \mathbf{X}_1 and \mathbf{X}_2 are fused to obtain the backbone features for both support and query. Next, a prototype activation (PA) module generates the support features \mathbf{F}^s , the initial query feature \mathbf{F}_0^q , and the initial query mask $\tilde{\mathbf{O}}_0^q$. The initial prototype \mathbf{G}_0 is randomly initialized at the beginning and then optimized during training. Afterward, by iteratively operating (1) to (7) in five layers, both prototype and features can be optimized step by step. For more details please refer to [16].

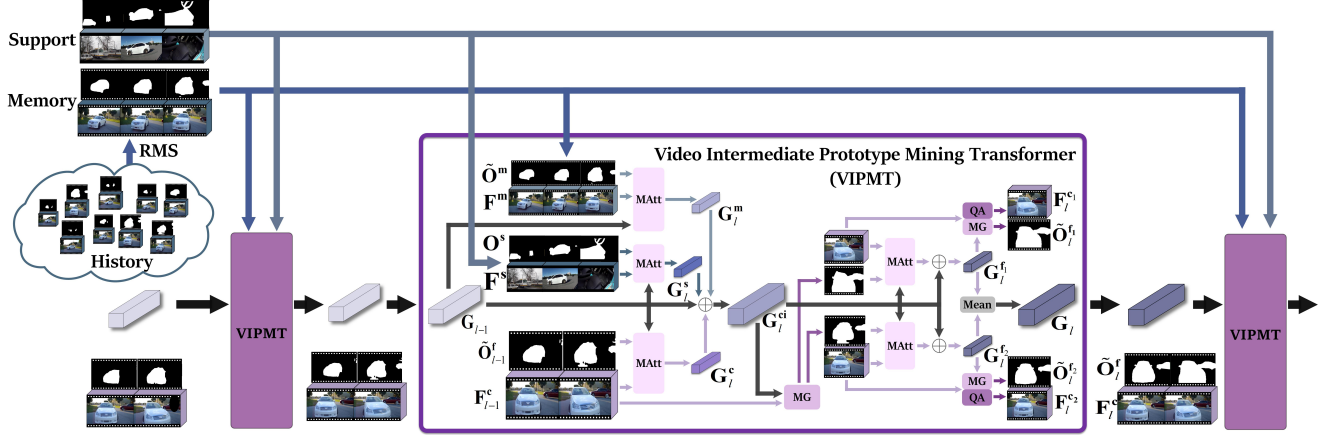


Figure 2. **Main architecture of the proposed VIPMT model for FSVOS.** “RMS” means our proposed reliable memory selection (Section 4.3), “MAtt” means masked attention (4), “MG” means mask generation (6), and “QA” means query activation (7). In each VIPMT iteration, the input intermediate prototype G_{l-1} first generates the support prototype G_l^s (2), the clip prototype G_l^c (8), and the memory prototype G_l^m (13), and combine them to obtain the clip-level intermediate prototype G_l^{ci} (14). Afterwards, a frame level prototype G_l^{fi} is obtained for each frame via (10) and then updates the frame mask \tilde{O}_l^{fi} and frame feature F_l^{fi} via (12). Finally, all G_l^{fi} are averaged to obtain G_l (11), which is input into the next iteration with \tilde{O}_l^f and F_l^c .

4. Video IPMT

As shown in Figure 2, we take the original IPMT model as the baseline and combine it with our proposed multi-grained prototype learning scheme for performing the FSVOS task. We elaborate on the design of our clip-level, frame-level, and memory-level prototypes first. Finally, we present the training loss, including the newly proposed CCDS loss.

4.1. Clip Prototype Learning

Since a video is composed of several consecutive frames, a straightforward way is to directly apply the IPMT model on each query frame. However, such a naive method does not consider any video temporal information and also lacks efficiency. Hence, we propose to take video clips as query units for both training and inference. Given a query video clip $\mathcal{C} = \{I^{c_i}\}_{i=1}^{T_c}$ with T_c frames, we perform the masked attention with clip-level query features to generate the clip prototype:

$$G_l^c = \text{MAtt}(G_{l-1}, F_{l-1}^c, \tilde{O}_{l-1}^f), \quad (8)$$

where $F_{l-1}^c \in \mathbb{R}^{T_c h w \times C}$, $\tilde{O}_{l-1}^f \in \mathbb{R}^{T_c h w \times 1}$ is the previous predicted masks and will be explained later. Then, we obtain the clip-level intermediate prototype:

$$G_l^{ci} = \text{MLP}(G_l^s + G_l^c + G_{l-1}), \quad (9)$$

which encodes local guidance information within the whole query clip. Next, we can use (6) to segment the target objects in the whole clip and obtain clip predictions $\tilde{O}_l^c \in \mathbb{R}^{T_c h w \times 1} = \text{MG}(G_l^{ci}, F_{l-1}^c)$, which preserves temporal coherence within the T_c frames.

4.2. Frame Prototype Learning

The clip-level intermediate prototype G_l^{ci} uses one prototype to represent the whole clip and encodes the clip consensus object information, however, may ignore frame-level fine-grained cues. This problem may cause a performance drop when the object appearance changes significantly within the clip. To mitigate this problem, we propose to further generate frame-level query prototypes $\{G_l^{fi}\}_{i=1}^{T_c}$ by updating G_l^{ci} with related fine-grained object cues in each frame:

$$G_l^{fi} = \text{MLP}(\text{MAtt}(G_l^{ci}, F_{l-1}^{c_i}, \tilde{O}_l^{c_i}) + G_l^{ci}). \quad (10)$$

In this process, we use G_l^{ci} as a good initialization to learn each G_l^{fi} and subsequently aggregate the frame-level query prototypes as the initialization for the next iteration:

$$G_l = \frac{1}{T_c} \sum_{i=1}^{T_c} G_l^{fi}. \quad (11)$$

As such, we enable *bidirectional* clip-frame prototype communication, which promotes the intra-clip temporal correlation. On the contrary, directly using the clip-level intermediate prototype G_l^{ci} as the initialization for the next iteration only enables *one-way* communication from the clip information to the frame-level.

At the same time, we use each frame prototype G_l^{fi} to generate the frame-level segmentation mask and update the query feature for each frame independently:

$$\begin{aligned} \tilde{O}_l^{fi} &= \text{MG}(G_l^{fi}, F_{l-1}^{c_i}), \\ F_l^{c_i} &= \text{QA}(G_l^{fi}, F_{l-1}^{c_i}). \end{aligned} \quad (12)$$

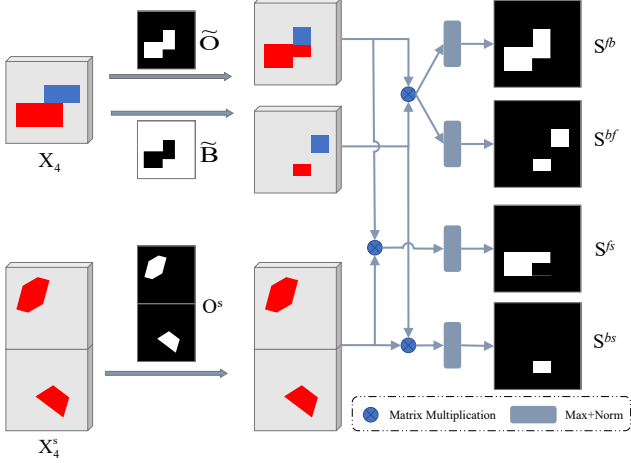


Figure 3. **Illustration of the generation process of the four structural similarity maps.** Red and blue regions indicate objects of two categories, respectively. Here we only show two support frames for concision.

Next, $\tilde{\mathbf{O}}_l^f = \{\tilde{\mathbf{O}}_l^{f_i}\}_{i=1}^{T_c}$, $\mathbf{F}_l^c = \{\mathbf{F}_l^{c_i}\}_{i=1}^{T_c}$, and \mathbf{G}_l are input to (8) again for the next iteration.

4.3. Memory Prototype Learning

Since a whole video usually contains numerous frames, only considering the information within a local clip is suboptimal as long-term historical temporal information is ignored. Many VOS works have also demonstrated that leveraging history memory is crucial for video data to enhance the temporal correlation. To this end, we propose to learn a memory prototype for providing historical guidance information. For each clip except the first one, we can use T_m previous frames as the memory set $\mathcal{M} = \{\mathbf{I}^{m_i}\}_{i=1}^{T_m}$, which also has predicted masks $\tilde{\mathbf{O}}^m \in \mathbb{R}^{T_m h w \times 1}$. Then, we use masked attention on memory features to generate the memory prototype:

$$\mathbf{G}_l^m = \text{MAtt}(\mathbf{G}_{l-1}, \mathbf{F}^m, \tilde{\mathbf{O}}^m), \quad (13)$$

where $\mathbf{F}^m \in \mathbb{R}^{T_m h w \times C}$ and $\tilde{\mathbf{O}}^m \in \mathbb{R}^{T_m h w \times 1}$. Then, the update of the clip-level intermediate prototype \mathbf{G}_l^{ci} in (9) can be rewrote as

$$\mathbf{G}_l^{ci} = \text{MLP}(\mathbf{G}_l^s + \mathbf{G}_l^c + \mathbf{G}_l^m + \mathbf{G}_{l-1}). \quad (14)$$

We do not consider using memory prototype at the frame level since \mathbf{G}_l^{ci} is designed to combine comprehensive guidance information at the clip-level, simultaneously from support, clip, and memory. In contrast, $\mathbf{G}_l^{f_i}$ encodes pure frame adaptive cues and uses \mathbf{G}_l^{ci} as the initialization.

Reliable Memory Selection. During the training stage, we can use the ground truth memory masks \mathbf{O}^m to replace the predicted masks for training an accurate model. However, during the inference stage, we can only use the predicted masks $\tilde{\mathbf{O}}^m$, which may be very noisy and hence heavily contaminate the learned prototype. To mitigate this problem,

we follow [17] and train an IoU regression network (IoUNet) to select reliable memory frames which have higher segmentation quality. Different from [17], we explicitly consider the nature of the FSVOS task and propose to leverage the structural similarity among the predicted foreground region, background region, and object region in support images for predicting more accurate IoU scores.

Specifically, as shown in Figure 3, given the backbone features of an memory frame $\mathbf{X}_4 \in \mathbb{R}^{h w \times C}$ and the support set $\mathbf{X}_4^s \in \mathbb{R}^{K h w \times C}$, the predicted memory mask $\tilde{\mathbf{O}} \in \mathbb{R}^{h w \times 1}$ (here we omit the superscript m for conciseness), and the ground truth support masks $\mathbf{O}^s \in \mathbb{R}^{K h w \times 1}$, we first obtain the background mask for the memory $\tilde{\mathbf{B}} = 1 - \tilde{\mathbf{O}}$. Then, we can evaluate the segmentation quality of the memory frame using a simple and intuitive idea: once it is well segmented, its foreground region $\tilde{\mathbf{O}}$ should be dissimilar to the background $\tilde{\mathbf{B}}$ and similar to the support foreground \mathbf{O}^s , while the background $\tilde{\mathbf{B}}$ should be dissimilar to both $\tilde{\mathbf{O}}$ and \mathbf{O}^s . To this end, we compute the structural similarity map of the foreground area of the memory w.r.t. the background area:

$$\mathbf{S}^{fb} = \text{Norm}(\text{Max}((\mathbf{X}_4 \odot \tilde{\mathbf{O}})(\mathbf{X}_4 \odot \tilde{\mathbf{B}})^T, 1)), \quad (15)$$

where \odot means element-wise multiplication, $\text{Max}(*, 1)$ means obtaining the maximum value along each row, and Norm denotes the min-max normalization to scale the data to the range of $[0, 1]$. Then, the output map $\mathbf{S}^{fb} \in \mathbb{R}^{h w \times 1}$ exactly represents how similar each foreground pixel is w.r.t. the background area, as shown in Figure 3.

Likewise, we can obtain a similarity map \mathbf{S}^{bf} of the background area w.r.t. the foreground, similarity map \mathbf{S}^{fs} of the foreground area of the memory w.r.t. the foreground of the support, and similarity map \mathbf{S}^{bs} of the background area of the memory w.r.t. the foreground of the support:

$$\begin{aligned} \mathbf{S}^{bf} &= \text{Norm}(\text{Max}((\mathbf{X}_4 \odot \tilde{\mathbf{B}})(\mathbf{X}_4 \odot \tilde{\mathbf{O}})^T, 1)), \\ \mathbf{S}^{fs} &= \text{Norm}(\text{Max}((\mathbf{X}_4 \odot \tilde{\mathbf{O}})(\mathbf{X}_4^s \odot \mathbf{O}^s)^T, 1)), \\ \mathbf{S}^{bs} &= \text{Norm}(\text{Max}((\mathbf{X}_4 \odot \tilde{\mathbf{B}})(\mathbf{X}_4^s \odot \mathbf{O}^s)^T, 1)). \end{aligned} \quad (16)$$

Once obtain the four structural similarity maps, we combine them with multi-scale memory features and the original predicted mask $\tilde{\mathbf{O}}$ to regress the IoU score of the memory frame. Concretely, we first downsample \mathbf{X}_1 to the 1/8 scale with stride convolution and then fuse it with $\mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4$ into a combined feature with 256 channels. Then, we concatenate the feature with the four structural similarity maps and $\tilde{\mathbf{O}}$ and input them into four convolution layers and two fully connected layers for predicting the IoU score.

During training, we design three ways to train the IoUNet. First, we can jointly train it with the FSVOS model, in which case real predicted masks can be used for training the IoUNet. Second, we can train it independently using synthesized video data. Specifically, we randomly add noises to the ground truth masks in the training set and use the synthesized

masks to mimic good and bad predictions. In the third way, we train the IoUNet using synthesized image data with the same noisy mask strategy. We compute the ground truth IoU score for each training mask and use a L_1 loss as supervision. During testing, we simply select reliable memory frames whose predicted IoU scores are larger than a threshold.

4.4. Training Loss

We follow IPMT [16] and use the final query feature \mathbf{F}_5^c to predict the final segmentation masks via two convolution layers, where Dice loss and IoU loss are both used for optimization. To ease the network training, in each iteration we use binary cross entropy (BCE) loss on the predicted support masks $\tilde{\mathbf{O}}_l^s$, clip masks $\tilde{\mathbf{O}}_l^c$, and frame masks $\tilde{\mathbf{O}}_l^f$. As for memory frames, we generate predictions for them at each iteration using $\text{MG}(\mathbf{G}_{l,b}^{\text{ci}}, \mathbf{F}^m)$, and then adopt the BCE loss computed with their ground truth masks as supervision.

Cross Category Discriminative Segmentation Loss. To make the learned clip-level intermediate prototypes \mathbf{G}^{ci} more category discriminative, we furthermore propose a CCDS loss and adopt it within each batch. Suppose our batchsize is set to $B > 1$, we select B video clips with different object categories to form each batch. Then, during training, we use the prototypes of each video to perform segmentation on other videos and let the predicted masks be all zeros. The idea behind is intuitive: if the prototypes are well optimized for a specific category, they should not activate any region on other videos with different categories. However, since a video may contain multiple object categories and we only have the ground truth of one class under the one-way few-shot learning setting, we propose to use the ground truth masks to filter out the regions for loss calculation:

$$L_{ccds} = \frac{1}{(B-1)B} \sum_{b=1}^B \sum_{j \neq b} \frac{1}{\sum \mathbf{O}_j} L_{b,j}, \quad (17)$$

$$L_{b,j} = \sum_{l=1}^5 \text{BCE}(\text{MG}(\mathbf{G}_{l,b}^{\text{ci}}, \mathbf{F}_j) \odot \mathbf{O}_j, \mathbf{Z}), \quad (18)$$

where $\mathbf{G}_{l,b}^{\text{ci}}$ is the clip-level intermediate prototype in the l -th iteration of the b -th video, \mathbf{O}_j and \mathbf{F}_j are the ground truth masks and the backbone features of the j -th video. \mathbf{Z} is an all zero matrix.

5. Experiments

5.1. Datasets and Evaluation Setting

Datasets. Following [1, 24], our experiments are conducted on the training set of YouTube-VIS [30] dataset, which has 40 categories and contains 2238 videos with 3774 instances. The dataset is divided into four folds, following the same slit with [1, 24]. In each fold, we use 30 categories for training and the rest 10 categories for testing. Following [1, 24], we

set our experiments on the 5-shot setting and randomly select five images from different videos of the same category as the support set. We also follow [24] and adopt the MiniVSPW [24] dataset to evaluate the generalizability of FSVOS methods. It includes 20 categories and provides longer video sequences than YouTube-VIS, since being more challenging.

Evaluation Setting. As the prior work [1] and VOS methods, we adopt the region similarity \mathcal{J} and contour similarity \mathcal{F} for performance evaluation. Apart from this, we also follow [24] and consider the video consistency metric VC_7 , which captures temporal prediction consistency among long-range adjacent frames over a temporal window of 7. We adopt the average score on four folds, *i.e.* mVC_7 , to evaluate the overall performance. We also follow [1, 24] and run the evaluation process five times and report the average results.

Furthermore, there are two different evaluation protocols proposed in [1] and [24] for sampling episodes during testing. *Protocol I* fixes the sampled support set for all query videos belonging to the same class in each run, while *protocol II* randomly samples a support set for every query video, which ensures a more stable performance evaluation. Hence, *in this paper we adopt protocol II for evaluation.*

Implementation Details. We adopt ResNet-50 [11] pre-trained on ImageNet [23] as our encoder backbone. Following IPMT [16], we freeze the parameters of the backbone during training and do not adopt online finetuning as [1, 24] did. The iteration step is set to 5 as IPMT. We set the clip length T_c and the memory length T_m to 5 without further tuning. The IoU threshold for selecting reliable memory frames is set to 0.8. When the reliable memory has more frames, we simply randomly select five frames from them. During training, we randomly select three clips from a video as a sample, where the first clip is trained without using memory, the second clip uses the first clip as memory, and the third clip randomly selects five frames from the first two as the memory. We also share the **MAtt** operation for support and memory since we found this leads to better performance.

We use Adam as our optimizer. The batchsize is set to $B = 4$ and the learning rate is set to $5e-4$. We train our model 100 epochs in total. All experiments are conducted on a NVIDIA Tesla A100 GPU. We adopt random horizontal flip, random crop and random resize to augment the training data. During training and testing, all video frames are downsampled to the resolution of (240,424) as the inputs.

5.2. Comparison with State-of-the-art Methods

In Table 1, we compare our VIPMT with four recently published FSIS methods and two FSVOS methods, on the YouTube-VIS [30] dataset. We find that our VIPMT largely improves all metrics, *i.e.* 4% \mathcal{J} mean score, near 6% \mathcal{F} mean score, and near 5% mVC_7 score, although previous

Methods	Name	\mathcal{J}					\mathcal{F}					mVC ₇
		Fold-1	Fold-2	Fold-3	Fold-4	Mean	Fold-1	Fold-2	Fold-3	Fold-4	Mean	
FSISS	NTRENet [15]	39.0	66.4	61.7	61.2	57.1	41.1	63.0	60.1	59.4	55.9	55.7
	SSP [10]	46.7	64.3	59.3	54.5	56.2	34.8	53.0	46.0	46.7	45.1	49.6
	VAT [12]	42.6	62.8	57.0	56.7	54.8	41.6	56.4	50.8	53.0	50.5	54.9
	IPMT [16]	43.8	65.8	61.0	60.7	57.8	42.5	59.5	57.8	55.9	53.9	57.7
FSVOS	DAN [1]	43.9	64.5	61.1	62.1	57.9	42.4	62.0	60.0	60.0	56.1	41.5
	TTI [24]	47.2	68.8	61.4	63.5	60.2	-	-	-	-	-	60.8
	VIPMT(Ours)	50.6	70.9	68.8	66.5	64.2	51.3	66.9	65.2	64.4	62.0	65.7

Table 1. Comparison with state-of-the-art methods on YouTube-VIS [30]*. Bold means the best performance.

Methods	Name	\mathcal{J}					\mathcal{F}					mVC ₇
		Fold-1	Fold-2	Fold-3	Fold-4	Mean	Fold-1	Fold-2	Fold-3	Fold-4	Mean	
FSVOS	TTI [24]	25.2	37.1	25.0	29.6	29.2	-	-	-	-	-	24.4
	VIPMT(Ours)	26.2	42.2	31.6	29.4	32.4	30.6	45.7	36.3	34.2	36.7	42.1

Table 2. Comparison with TTI [24] on MiniVSPW [24].

Datasets	Methods	\mathcal{J} -Mean	\mathcal{F} -Mean	mVC ₇
YouTube-VIS	IPMT+STCN [6]	58.9	56.5	61.7
	IPMT+XMem [5]	59.2	56.8	62.2
	IPMT+RDE-VOS [14]	59.3	58.0	63.3
	VIPMT(Ours)	64.2	62.0	65.7
MiniVSPW	IPMT+STCN [6]	27.7	31.1	39.3
	IPMT+XMem [5]	27.7	31.7	41.1
	IPMT+RDE-VOS [14]	28.8	33.5	40.1
	VIPMT(Ours)	32.4	36.7	42.1

Table 3. Comparison with state-of-the-art VOS methods.

works [1, 24] used on-line learning while we didn’t.

To verify the generalization ability of our method, we compare our method with TTI [24] on MiniVSPW. Table 2 shows our method obtains a significant improvement of more than 3% on \mathcal{J} and more than 17% on mVC₇, which verifies that our method works well in more challenging scenarios.

In Figure 4, we give some visual comparison cases. Compared with TTI [24], our VIPMT is less distracted by other objects and achieves more precise segmentation, even in highly occluded scenes (bottom row).

We also include three state-of-the-art semi-supervised VOS methods for comparison, *i.e.* STCN [6], XMem [5], and RDE-VOS [14]. Specifically, we use IPMT for segmenting the first frame and then use VOS methods for propagating the segmentation to the full video. Table 3 demonstrates that such an ad-hoc combination of FSIS and VOS methods can not achieve as good results as our model does.

5.3. Ablation Study

In this section, we report ablation study results on the YouTube-VIS dataset. For \mathcal{J} and \mathcal{F} we use their average on four folds, *i.e.* \mathcal{J} -Mean and \mathcal{F} -Mean.

Effectiveness of Each Component. In Table 4, we first

*Please note that here we re-tested the performance of DAN [1] using *protocol II*. Hence, its scores are different from the original paper. For TTI [24], since the authors did not provide the scores for \mathcal{F} , we leave them blank. We retrained all FSIS methods on YouTube-VIS for a fair comparison.

Name	Settings				\mathcal{J} -Mean	\mathcal{F} -Mean	mVC ₇
	Clip	Frame	Memory	CCDS			
Baseline					57.8	53.9	57.8
+C	✓				61.8	58.8	60.7
+C+F	✓	✓			62.5	60.3	62.1
+C+F+M	✓	✓	✓		63.5	61.9	62.8
VIPMT	✓	✓	✓	✓	64.2	62.0	65.7

Table 4. Ablation study on the effectiveness of each model component. Baseline means IPMT [16].

adopt IPMT [16] as the baseline, which processes each video frame independently. Then, we adopt our clip prototype learning, denoted as “+C”. Results show that “+C” surpasses the baseline by a large margin on all metrics, which verifies the effectiveness of using local temporal guidance with our clip prototype learning. After that, we further add the frame prototype learning to “+C”, resulting in “+C+F”. Benefiting from the frame-level fine-grained cues, “+C+F” achieves large performance improvement for \mathcal{F} -Mean and mVC₇. Moreover, we adopt “+C+F+M” to represent adding the memory prototype learning on “+C+F”. It largely improves the performance for both \mathcal{J} -Mean and \mathcal{F} -Mean, which shows the importance of long-term temporal information brought by the memory prototype learning. Based on this, we further add the CCDS loss to obtain our final model VIPMT. It also brings obvious improvements, especially on mVC₇, from 62.8 to 65.7. This proves the effectiveness of increasing category discrimination for FSVOS.

In Figure 5, we visualize the predicted masks from the different models mentioned above. We can find that progressive improvements can be brought by using multi-grained temporal prototypes and the CCDS loss.

Clip-frame Communication. As mentioned in Section 4.2, we use the mean of the frame prototypes ((11)) to initialize the intermediate prototype for the next iteration, which enables *bidirectional* clip-frame communication. On the contrary, directly using the clip-level intermediate prototype G_l^{ci} as the initialization only enables *one-way* communi-



Figure 4. **Visualization of comparison cases.** (a) Support sets. (b) Predicted masks of our VIMPT. (c) Predicted masks from TTI [24].

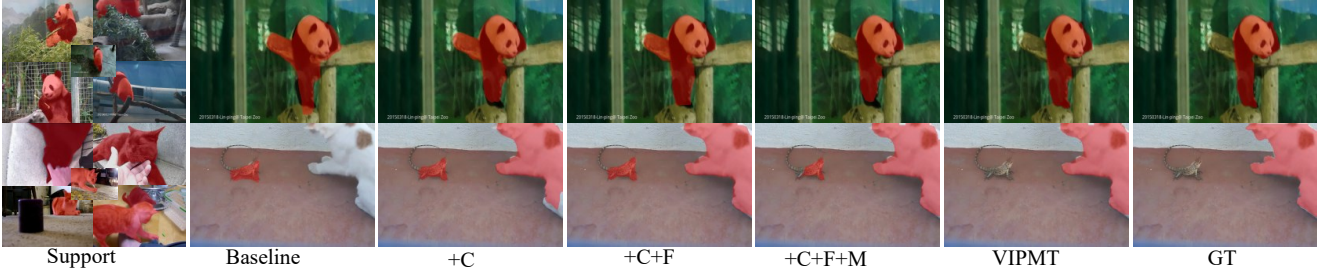


Figure 5. **Visualization of the comparison of different ablation study models.** We can see progressive improvements brought by the multi-grained temporal prototypes and the CCDS loss.

Clip-frame Communication	\mathcal{J} -Mean	\mathcal{F} -Mean	mVC ₇
One-way	61.4	59.1	60.1
Bidirectional	62.5	60.3	62.1

Table 5. **Comparison of two clip-frame communication schemes.**

tion. Here we compare the effectiveness of the two schemes. Results in Table 5 demonstrate that our design of using bidirectional clip-frame communication surpasses the one-way scheme by a large margin on all metrics.

Different Training Strategies of IoUNet. As mentioned in Section 4.3, we can use three ways to train the IoUNet, *i.e.* using real predicted masks (“Real”), synthesized video data (“Video”), and synthesized image data (“Image”), respectively. For “Image”, we use an FSIS dataset COCO-20ⁱ [20] and remove the classes overlapped with our video data for IoUNet training. We compare their performance in Table 6. We also compare the performance of using our proposed structural similarity maps or not (“w/ SSM” and “w/o SSM”). The results are reported under the IoU threshold of 0.5. The table demonstrates that using synthesized image data results in the best performance in both “w/ SSM” and “w/o SSM” settings and using the structural similarity maps always leads to better performance. Another benefit of using image data is that we can train a unified IoUNet for all folds instead of training an IoUNet for each fold. Hence, finally, we use synthesized image data for IoUNet training.

IoU Threshold. Finally, we investigate the best IoU threshold for memory selection. Table 7 indicates that 0.8 is the best threshold and using the proposed structural similarity maps always leads to better performance.

Data	w/o SSM			w/ SSM		
	\mathcal{J} -Mean	\mathcal{F} -Mean	mVC ₇	\mathcal{J} -Mean	\mathcal{F} -Mean	mVC ₇
Real	62.8	61.5	62.6	63.2	61.6	62.8
Video	62.9	61.6	62.6	63.2	61.6	62.8
Image	63.0	61.7	62.7	63.3	61.7	63.0

Table 6. **Different training strategies of IoUNet.** SSM means the proposed structural similarity maps.

Threshold	w/o SSM			w/ SSM		
	\mathcal{J} -Mean	\mathcal{F} -Mean	mVC ₇	\mathcal{J} -Mean	\mathcal{F} -Mean	mVC ₇
0.5	63.0	61.7	62.7	63.3	61.7	63.0
0.6	63.1	61.7	62.9	63.3	61.7	62.7
0.7	63.1	61.8	62.9	63.4	61.8	62.6
0.8	63.1	61.8	62.5	63.5	61.9	62.8
0.9	63.0	61.8	62.3	63.5	61.9	62.5

Table 7. **Experimental results of using different IoU thresholds.**

6. Conclusion

We explored learning multi-grained temporal prototypes to tackle the FSVOS task. Based on the IPMT model, the query video information was decomposed into a clip prototype, a memory prototype, and frame prototypes. We also improved an IoU regression method for selecting reliable memory for FSVOS by leveraging the task prior knowledge and proposed a new loss to enhance the category discriminability of the prototypes. The effectiveness of our proposed VIPMT has been verified on two benchmark datasets.

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China under Grants 62071388, 62136007, U20B2065 and 62036005, and the Fundamental Research Funds for the Central Universities under Grant D5000230057.

References

- [1] Haoxin Chen, Hanjie Wu, Nanxuan Zhao, Sucheng Ren, and Shengfeng He. Delving deep into many-to-many attention for few-shot video object segmentation. In *CVPR*, pages 14040–14049, 2021. [1](#), [2](#), [3](#), [6](#), [7](#)
- [2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2017. [1](#)
- [3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. [3](#)
- [4] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Pixel classification is not all you need for semantic segmentation. In *NeurIPS*, volume 34, pages 17864–17875, 2021. [1](#)
- [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. [7](#)
- [6] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *NeurIPS*, 34:11781–11794, 2021. [7](#)
- [7] 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](#)
- [8] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *CVPR*, pages 895–903, 2017. [2](#)
- [9] Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, volume 3, 2018. [2](#)
- [10] Qi Fan, Wenjie Pei, Yu-Wing Tai, and Chi-Keung Tang. Self-support few-shot semantic segmentation. In *ECCV*, 2022. [7](#)
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. [3](#), [6](#)
- [12] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *ECCV*, pages 108–126. Springer, 2022. [7](#)
- [13] Chunbo Lang, Gong Cheng, Binfei Tu, and Junwei Han. Learning what not to segment: A new perspective on few-shot segmentation. In *CVPR*, pages 8057–8067, 2022. [1](#), [2](#)
- [14] 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. [7](#)
- [15] Yuanwei Liu, Nian Liu, Qinglong Cao, Xiwen Yao, Junwei Han, and Ling Shao. Learning non-target knowledge for few-shot semantic segmentation. In *CVPR*, pages 11573–11582, 2022. [2](#), [7](#)
- [16] Yuanwei Liu, Nian Liu, Xiwen Yao, and Junwei Han. Intermediate prototype mining transformer for few-shot semantic segmentation. In *NeurIPS*, 2022. [1](#), [2](#), [3](#), [6](#), [7](#)
- [17] 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. [2](#), [5](#)
- [18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. [1](#)
- [19] Xiankai Lu, Wenguan Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. Video object segmentation with episodic graph memory networks. In *ECCV*, pages 661–679. Springer, 2020. [2](#)
- [20] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *CVPR*, pages 622–631, 2019. [8](#)
- [21] 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. [2](#)
- [22] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, pages 2663–2672, 2017. [2](#)
- [23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. [6](#)
- [24] Mennatullah Siam, Konstantinos G Derpanis, and Richard P Wildes. Temporal transductive inference for few-shot video object segmentation. *arXiv preprint arXiv:2203.14308*, 2022. [1](#), [2](#), [3](#), [6](#), [7](#), [8](#)
- [25] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE TPAMI*, 2020. [1](#), [2](#)
- [26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, pages 5998–6008, 2017. [3](#)
- [27] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, volume 29, 2016. [2](#)
- [28] Huaxin Xiao, Jiashi Feng, Guosheng Lin, Yu Liu, and Maojun Zhang. Monet: Deep motion exploitation for video object segmentation. In *CVPR*, pages 1140–1148, 2018. [2](#)
- [29] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *ECCV*, pages 763–778. Springer, 2020. [2](#)
- [30] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, pages 5188–5197, 2019. [6](#), [7](#)
- [31] Tianyu Yang and Antoni B Chan. Learning dynamic memory networks for object tracking. In *ECCV*, pages 152–167, 2018. [2](#)
- [32] Gengwei Zhang, Guoliang Kang, Yi Yang, and Yunchao Wei. Few-shot segmentation via cycle-consistent transformer. *NeurIPS*, 34:21984–21996, 2021. [1](#)

- [33] Tianfei Zhou, Fatih Porikli, David J Crandall, Luc Van Gool, and Wenguan Wang. A survey on deep learning technique for video segmentation. *IEEE TPAMI*, 2022. [2](#)