Spatiotemporal Graph Neural Network based Mask Reconstruction for Video Object Segmentation

Daizong Liu[1], Shuangjie Xu[2], Xiao-Yang Liu[3], Zichuan Xu[4], Wei Wei[1], Pan Zhou[1*]

[1]Huazhong University of Science and Technology
[2] DEEPROUTE.AI
[3]Columbia University
[4]Dalian University of Technology

{dzliu, weiw, panzhou}@hust.edu.cn, shuangjiexu@deeproute.ai, xl2427@columbia.edu, z.xu@dlut.edu.cn

Abstract

This paper addresses the task of segmenting class-agnostic objects in semi-supervised setting. Although previous detection based methods achieve relatively good performance, these approaches extract the best proposal by a greedy strategy, which may lose the local patch details outside the chosen candidate. In this paper, we propose a novel spatiotemporal graph neural network (STG-Net) to reconstruct more accurate masks for video object segmentation, which captures the local contexts by utilizing all proposals. In the spatial graph, we treat object proposals of a frame as nodes and represent their correlations with an edge weight strategy for mask context aggregation. To capture temporal information from previous frames, we use a memory network to refine the mask of current frame by retrieving historic masks in a temporal graph. The joint use of both local patch details and temporal relationships allow us to better address the challenges such as object occlusion and missing. Without online learning and fine-tuning, our STG-Net achieves state-of-the-art performance on four large benchmarks (DAVIS, YouTube-VOS, SegTrack-v2, and YouTube-Objects), demonstrating the effectiveness of the proposed approach.

Introduction

Video object segmentation (VOS) in semi-supervised setting aims to segment the class-agnostic objects or instances from the background according to the annotation in the first frame, which has been widely applied to video editing, automatic driving, etc. Tremendous progress (Johnander et al. 2019; Caelles et al. 2017; Wug Oh et al. 2018; Wang et al. 2019a) has been made with deep learning methods in recent years, most of which directly embed the whole frame image or propagate the previous mask into current frame. However, it is still challenging due to the background noise, object missing, or severe occlusion in real world scenarios.

To address such challenges, detection based schemes (Li et al. 2017; Luiten, Voigtlaender, and Leibe 2018) are proposed, which restore missing objects or re-establish objects with bounding box proposals. These proposals of target objects are either generated individually in each frame by detectors like Mask R-CNN (He et al. 2017), or further merged
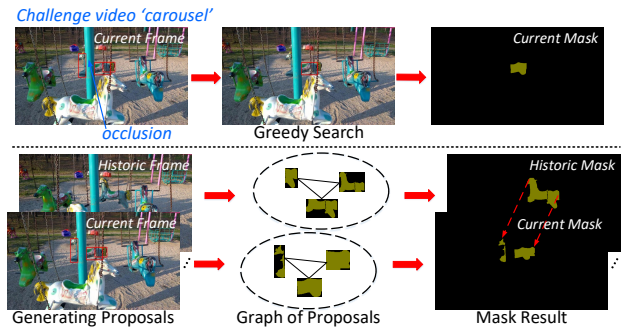

Figure 1: Different from previous detection based methods that generally utilize a greedy strategy to choose the best proposal for segmentation, our spatiotemporal graph considers all proposals of each frame in the spatial domain and utilizes historic masks for refinement in the temporal domain, which provides better mask details.

with a few adjacent neighboring frames (Luiten, Voigtlaender, and Leibe 2018). Although they are effective in object missing and occlusion scenarios, these approaches rely on a greedy search scheme that extracts the best proposal in a frame, as shown in the upper part of Figure 1, resulting in a strong dependence not only on the proposal quality, but also on a reliable Re-ID network (Li et al. 2017) for proposal selection. Since the local patch details may be contained in all proposals scattered across the frame, instead of choosing the best proposal by a greedy search scheme, we argue that one should leverage the rich contexts of all proposals to reconstruct a more accurate mask.

Recently, graph neural network (GNN) (Scarselli et al. 2008; Gilmer et al. 2017; Cheng et al. 2020b,a; Liu et al. 2020) is recognized as a promising approach in sequential information processing. It takes advantage of aggregating the information with node-wise correlation establishment. Previous GNN based methods (Wang et al. 2019b; Haller, Florea, and Leordeanu 2019; Bao, Wu, and Liu 2018) in the VOS task either represent image frames as nodes to explore their temporal correlations, or represent pixels as nodes that may lose instance-level relations. They may also fail to recover the spatial mask details in local contexts. To better capture the local patch details for more accurate mask reconstruction, different from them, we take masks of all detection

based proposals in the same frame as nodes to construct a spatial graph, and correlate current frame result to the previous frame masks to build a temporal graph. Such a joint graph neural network is capable of not only aggregating the scattered instance-level mask details of current frame (in the spatial domain), but also capturing the temporal correlations with historic masks (in the temporal domain).

In this paper, to better capture the local patch details from spatial information of the current frame and capture motion clues with temporal information from the previous frames, we propose a novel spatiotemporal graph neural network (STG-Net) for video object segmentation. Specifically, we construct a fully-connected spatial graph on the mask proposals of current frame to establish intra-object proposals relationship, and propose a temporally-connected graph to link the historic masks, as shown in the bottom part of Figure 1. In a spatial graph, we develop an edge weight strategy to represent the correlation between two instance-level mask proposals by considering their feature similarity. After spatial graph updating, we design a score function based on motion estimation and mask propagation to choose the best reconstructed mask from all nodes for each object in current frame. Then a temporal graph is developed to correlate the chosen mask with historic masks of previous frame for mask refinement, which can also be regarded as a reconstruction process. Therefore, the mask is reconstructed in both spatial and temporal domains to produce a more accurate segmentation that recovers detailed contexts of objects.

The contributions of our work are summarized as follows:

- We propose a novel VOS method named STG-Net based on a spatiotemporal graph to recover the local patch details in an instance level. With the cooperation of spatial and temporal graph networks, STG-Net has sufficient capacity to aggregate detailed mask contexts for more accurate mask reconstruction. To the best of our knowledge, it is the first time to take advantage of both spatial and temporal correlations with GNN in VOS task.

- Instead of searching the best proposal in a greedy manner, our spatial graph network takes all object proposals into consideration with an edge weight strategy, which is measured by the feature similarity of a proposal pair. It helps to aggregate mask details from scattered locations. A score function is then employed to choose the reconstructed mask from spatial graph by considering both motion estimation and mask propagation.

- We develop a temporal graph based on the chosen masks from spatial graph along the time dimension. For each node, we utilize a memory network to retrieve mask contexts from the historic masks, and use them to refine the current mask with a temporal graph network.

Experimental results show that the proposed STG-Net achieves state-of-the-art performance on DAVIS, YouTube-VOS, SegTrack-v2, and YouTube-Objects datasets without online learning on the annotation of the first frame. The superior visual results show better mask details than others, which demonstrates the effectiveness of our method in handling the challenging occluded and missing objects.

Related Works

Semi-Supervised Video Object Segmentation. Semi-supervised video object segmentation can be roughly classified into three categories: matching-based, propagation-based, and detection-based methods. Matching-based methods (Caelles et al. 2017; Voigtlaender and Leibe 2017; Zeng et al. 2019) generally utilize the given mask in the first frame to extract appearance information for objects of interest, which is then used to find similar objects in succeeding frames. Some works (Caelles et al. 2017; Voigtlaender and Leibe 2017) trained a parent network on still images and then fine-tuned the pre-trained work with one-shotonline learning. Embedding approaches (Chen et al. 2018b; Hu, Huang, and Schwing 2018) mapped pixels to group the pixels of same object, and there are methods (Voigtlaender et al. 2019; Wang et al. 2019c) extend from them for multiple object segmentation with correlation operation. Propagation-based methods (Wug Oh et al. 2018; Johnander et al. 2019; Xu et al. 2019) utilize temporal information to refine masks propagated from preceding frames. The above two methods mainly depend on the robustness of the feature extractor to segment the foreground object in the whole image where much background noise may be induced. Different from them, detection-based methods (Li et al. 2017; Wang et al. 2019a; Luiten, Voigtlaender, and Leibe 2018) first detect the best bounding box of each object in a frame, then crop out the target and input it into a segmentation model. Although it can decrease background noises and improve the performance of segmentation, it relies on the quality of the generated bounding box of each object. To avoid losing local details, our method aggregates the mask contexts of all proposals in the same frame to automatically reconstruct a more accurate mask.

Graph Neural Networks. Graph neural network (GNN) (Scarselli et al. 2008) is an extension for recursive neural networks and random walk based models for graph structured data. (Gilmer et al. 2017; Li, Han, and Wu 2018) further adapted GNN to sequential outputs with a learnable message passing module. Generally, for each node v in a graph, the updating process includes two steps: message aggregation and hidden state update. For the updating of l-th iteration, the node v first aggregates messages from its neighbors $\mathcal{N}(v)$ into a single message $\boldsymbol{m}^v$ and then updates the hidden state h^v itself with $\boldsymbol{m}^v$, it is according to:

$$\boldsymbol{m}^v = F(h^u, u \in \mathcal{N}(v)), \quad (h^v)' = U(h^v, \boldsymbol{m}^v), \quad (1)$$

where $F(\cdot), U(\cdot)$ are the functions to update the message and hidden state. Different from (Haller, Florea, and Leordeanu 2019; Wang et al. 2019b) that take a naive GNN to the VOS task and treat frames as nodes for temporal contexts exploring, our spatial graph is constructed with object proposals of each frame with a tailored edge weight strategy.

Memory Networks. Memory networks (Vaswani et al. 2017; Sukhbaatar et al. 2015) have external memory where information can be further written and read. Given an input, the information is separately embedded into key and value feature maps, where the key feature maps are used to address relevant memories whose corresponding value feature maps are returned. Different from the embedding process
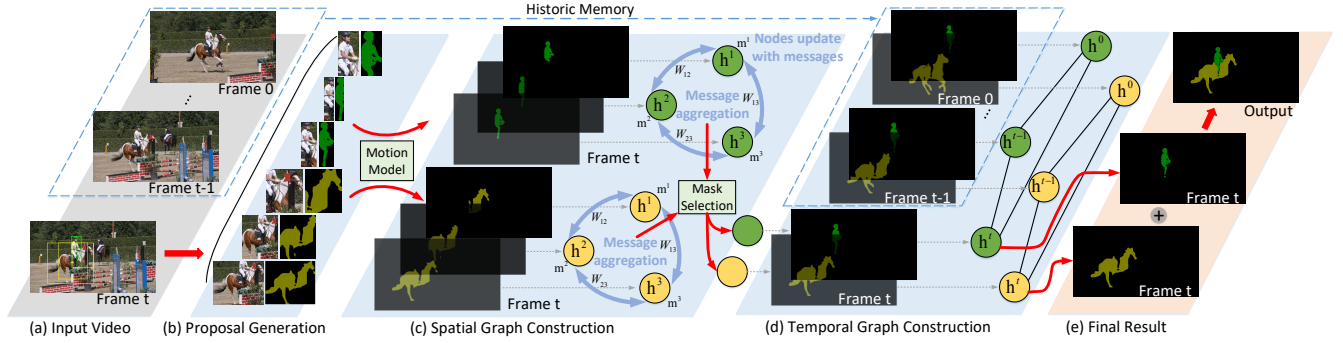
Figure 2: The framework of our proposed scheme. Given a video frame t, we first obtain bbox proposals by an offline detection model and generate corresponding masks by an offline segmentation model. Then we classify each proposal with our motion model to construct intra-object spatial graphs, where each node is a mask proposal. During the spatial graph updating, each node aggregates mask contexts from its neighbors and updates itself to reconstruct more accurate mask. We further design a mask selection function to choose the best reconstructed mask from all nodes in graph. At last, we construct a temporal graph to retrieve historic masks of previous frames for refinement, and obtain the final result of frame t.

(Oh et al. 2019; Lu et al. 2020) for value features maps, we take the mask result of each frame as value, to retrieve historic masks for refining the current mask in temporal graph.

Spatiotemporal Graph Neural Network

In this section, we present our method STG-Net, as illustrated in Figure 2. Given a video frame, we first utilize a class-agnostic detection model to generate object bounding box (bbox) proposals and a class-agnostic segmentation model to segment corresponding masks, which is performed offline. Then, we design a motion model to classify the proposals, and take the proposals of the same object as nodes to construct a spatial graph. After spatial graph updating, we choose the best reconstructed mask from all nodes, and capture the temporal information by using a temporal graph that utilizes historic masks of previous frames for refinement. We pre-process the proposal generation offline, thus our whole framework can be end-to-end training.

Proposal Generation

We first generate possible object bbox proposals of a video frame by using an offline detection method Mask R-CNN (He et al. 2017), where each bbox proposal contains different local patch details of objects. As VOS task only considers foreground objects, we change the number of categories in Mask R-CNN into only one class to make the bbox proposals class-agnostic. Specifically, we extract object bboxes with pre-defined thresholds of detection confidence and non-maximum suppression to keep the most possible bboxes. Given a video frame $\boldsymbol{I} \in \mathbb{R}^{3 \times H \times W}$, we denote the extracted bbox proposal as $b^v = (x^v_{\min}, y^v_{\min}, x^v_{\max}, y^v_{\max})$, where v is the v-th proposal of detection results in current frame. To further segment its corresponding mask $\boldsymbol{M}^v$, we employ Deeplabv3+ (Chen et al. 2018a) network with ResNet101 (He et al. 2016) backbone offline. Since objects tend to move smoothly through space in time, we use optical flow generated by FlowNet2.0 (Ilg et al. 2017) as a direct source of information, to estimate a rough mask $\boldsymbol{Q}$ for current frame by a warp operation (Khoreva et al. 2017). And

we take $\boldsymbol{Q}$ as an additional input besides the cropped RGB image $\boldsymbol{I}^v$ which is based on bbox b^v, to guide the segmentation module to produce more accurate mask.

Spatial Graph Construction

Different proposals may contain different local patch details, we construct spatial graphs in Figure 2(c) to aggregate the mask contexts of all proposals to enrich the segmentation information. To handle the class-agnostic mask proposals $\boldsymbol{M}^v$, we first introduce a motion model to prepare for proposal-wise classification, then our spatial graph can be built with intra-object proposals for contexts sharing. We construct our spatial graph of each object as follows.

Preparation. Before constructing the spatial graph of object o, we first group the o-class proposals from all proposals in current frame by classification. As object smoothly moves across the video, we can utilize its mask results in previous frames and generate corresponding bboxes $\{b\}$ and center points $\{c\}$ as motion history. Based on such history memory, we can predict a bbox p as objective probability location in current frame based on the prior knowledge of previous frame bboxes, then take the closest proposal b^v into object o class. Since each bbox mainly depends on the characters of size s_t and center point c_t where s_t is composed of the height and weight of bbox, t denotes the time-step, we develop a motion model to estimate the center point of p based on the previous n steps movements by:

$$c_t = c_{t-1} + \frac{1}{n} \sum_{r=t-n}^{t-1} (c_r - c_{r-1}), \qquad (2)$$

where the second term means the average velocity. The bbox size s_t of p can also be estimated as $s_t = \frac{1}{n}\sum_{r=t-n}^{t-1} s_r$, since most object sizes change smoothly in video sequence. Therefore, the predicted objective probability bbox p of object o in current frame t is composed of (c_t, s_t). Given a bbox proposal b^v, we first calculate the intersection over union (IoU) scores between the area of b^v and the area of

p for all object. Then we rank the IoU scores to find the object class of the highest one to classify the bbox b^v. If the top ranked score refers to the object o, we add the bbox b^v into the object o-class. After getting the proposal index $v, v = 1, ..., N$ of object o-class, we past corresponding mask into a void image to rebuild back a full mask $M^v \in \mathbb{R}^{H \times W}$.

Graph construction. To recover the local patch details in intra-object proposals, we construct a fully-connected spatial graph of object o with masks $\{M^v\}$ as node, to propagate beneficial information for mask reconstruction. Each node aggregates the mask contexts from its neighbors to reconstruct the mask itself. Generally, the correlation between different nodes are not always the same, as the proposals which have closer position and similar representation tends to be more relevant. To selectively propagate the mask information more between the relevant nodes while reduce the noise between less relevant ones, we define an edge weight W_{vu} on the edge between node v and u, which can be formalized as follows:

$$W_{vu} = \begin{cases} \alpha \cos(X^v, X^u) + \beta \mathrm{IoU}(b^v, b^u), & v \neq u \\ 0, & v = u \end{cases},$$
(3)

where the first item is the cosine similarity to measure the correlation between features X^v and X^u extracted by a learnable CNN for the proposals b^v and b^u, respectively. α and β control the ratio between the feature similarity score and the IoU score, and we set W_{vv} to 0 to avoid self-enhancing. This weight strategy also helps to reduce the influence on the edge between the wrong classified proposal from motion model and the correct one as their similarity score will be much smaller.

Graph updating. For node v in graph, the updating process contains two steps: 1). Mask message aggregation: To attach the mask information from other nodes u, node v first aggregates the mask message m^v from its neighbors by a weighted summation with the edge weight W_{vu}; 2). Node mask update: Then node v updates the state h^v itself with the aggregated messages m^v to reconstruct more contextual segmentation result. In details, we first initial the state h^v of node v with the mask proposal M^v, and defined its neighbor sets as $\mathcal{N}(v)$. During the graph updating, it first aggregates messages m^v from neighbors $\mathcal{N}(v)$ by function $F(\cdot)$ as follow:

$$m^v = F(h^u) = \sum_u W_{vu} h^u, u \in \mathcal{N}(v).$$
(4)

Then node v reconstructs the mask of its state with the aggregated mask message $m^v \in \mathbb{R}^{H \times W}$ by:

$$(h^v)' = U(h^v, m^v) = \frac{(1 - W_{vv})h^v + m^v}{1 + \sum_u W_{vu}}, u \in \mathcal{N}(v),$$
(5)

where $U(\cdot)$ is the function to update the mask state, and $(1 + \sum_u W_{vu})$ is used to normalize the mask result. Specially, we iterative the graph updating process for several steps. To avoid over-smoothing problem (Li, Han, and Wu 2018) in graph nodes, we only conduct less than three iterations and keep the edge weight W_{vu} unchanged during the

graph iterative updating. At last, for node v, we get the binary reconstructed mask $\widehat{M}^v = (h^v > thr)$ by a threshold thr, and it recovers the local patch details from intra-object proposals.

Temporal Graph Construction

After getting masks $\{\widehat{M}^v\}$ in spatial graph, we design a score function to choose the best mask of object o from all nodes. Then, we refine it using the temporal information of the same instance by a temporal graph in Figure 2(d).

Mask selection. We define the score function as follows:

$$\mathcal{S}(\widehat{M}^v | (p, Q)) = \lambda_1 \mathrm{IoU}(\mathcal{B}(\widehat{M}^v), p) + \lambda_2 \frac{\widehat{M}^v \cap Q}{\widehat{M}^v \cup Q}, \quad (6)$$

where $\mathcal{B}(\cdot)$ is the function to extract the bbox of the reconstructed mask $\widehat{M}^v$, and Q is the warped mask from previous frame with optical flow. Our mask selection score contains two parts: 1) bbox IoU score between the bbox of current segmented object and predicted probability bbox p from motion model, which stands for the measurement of object motion estimation; 2) the intersection area over union area score between the current estimated mask and the warped mask, which represents the performance of mask propagation. λ_1 and λ_2 are the parameters to control the ratio of these two scores. We choose the best mask index v with $\max_v \mathcal{S}(\widehat{M}^v | (p, Q))$, and denote the chosen mask as current frame segmentation result $\widehat{M} \in \mathbb{R}^{H \times W}$ of object o.

Graph construction. Inspired by differential memory networks (Vaswani et al. 2017; Sukhbaatar et al. 2015), we utilize the chosen mask of object o in current frame and the previous frame mask results as nodes, to construct a temporal graph. This graph is designed to refine the mask of the current frame by aggregating historic masks as memories and is evolved by adding new nodes over time. At frame t, there are $t + 1$ nodes with mask $\widehat{M}^r, r \leq t$ including the first frame 0. We first crop the image I based on $\mathcal{B}(\widehat{M}^r)$, and resize it to $I^r \in \mathbb{R}^{3 \times H_1 \times W_1}$. Then, we extract its embedding features by a learnable CNN as the key feature maps $K^r \in \mathbb{R}^{C \times H_1 \times W_1}$. We take the cropped $\widehat{M}^r$ as the corresponding value map. The similarity between key feature maps of the current and previous frames are computed to determine when-and-where to retrieve relevant previous value maps from. Therefore, every pixel of each previous frame value map can be utilized to construct a new self-predicted mask for current frame based on such similarities. These constructed masks are taken as neighborhood messages m^t for current frame node t to update its mask h^t.

The main differences of our memory network compared to STM (Oh et al. 2019) are: For structure, our temporal graph takes bbox based image/mask as key/value with only one network. Instead, STM takes whole image as input, and utilizes an encoder-decoder structure to embed image into two feature spaces named as key/value. For goal, we aim to propagate previous masks on edges for current mask refinement. And STM is to predict mask for each input.

Graph updating. In details, node t first initials its state h^t with mask $\widehat{M}^t$, and then retrieves the mask memories from
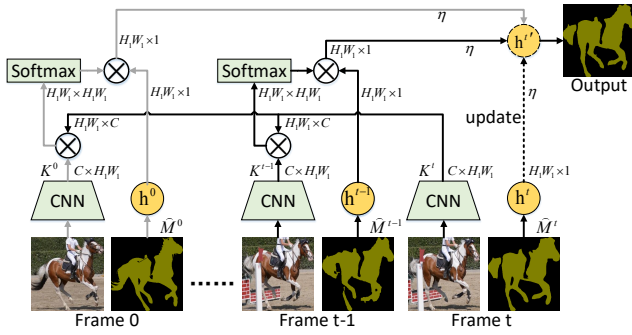
Figure 3: Illustration of temporal graph updating process at frame t, where '$\otimes$' denotes the matrix multiplication.

previous t frames including the first frame 0 as:

$$(\boldsymbol{m}^t)_i = F(h^r) = \sum_{r=0}^{t-1} \frac{\sum_j \exp((\boldsymbol{K}^t)_i \odot (\boldsymbol{K}^r)_j)(h^r)_j}{\sum_j \exp((\boldsymbol{K}^t)_i \odot (\boldsymbol{K}^r)_j)}, \tag{7}$$

where i, j are the location index and $(\boldsymbol{K}^t)_i \in \mathbb{R}^{C \times 1}$, $\odot$ denotes the dot-product. Therefore, the refinement process of frame t by using the mask memories can be seen as a process of reconstruction, which is formulated by:

$$(h^t)' = U(h^t, \boldsymbol{m}^t) = \eta(h^t + \boldsymbol{m}^t), \tag{8}$$

where $\eta = 1/(t+1)$ is used to normalize the mask result. Details of temporal graph updating at frame t are shown in Figure 3. Note that our temporal graph only updates within one iteration, and the refined output mask $\widehat{\boldsymbol{M}} \in \mathbb{R}^{H \times W}$ of current frame is also obtained by $\widehat{\boldsymbol{M}} = (h^t > thr)$.

Network Structure and Loss Function

For the feature extractor in spatial graph, we use a ResNet101 backbone of which the weights are initialized from the released model of RVOS (Ventura et al. 2019), and obtain the proposal features from its last layer. As for memory key maps, we use the stage-4 feature map of a ResNet50 which is finetuned online. During the training, as for the reconstructed mask $\widehat{\boldsymbol{M}}^v$ in spatial graph, we utilize the ground truth pair $(b, \mathbf{M})$ to choose the best mask instead of $(p, \boldsymbol{Q})$. To make our model end-to-end trainable with the mask selection operation, we develop an additional distance loss to backpropagate the gradient for all reconstructed masks. Our total loss function $\mathcal{L}$ is formulated as:

$$\mathcal{L}(\widehat{\boldsymbol{M}}) = \gamma \sum_v |\mathcal{S}(\widehat{\boldsymbol{M}}^v|(b,\mathbf{M})) - \mathcal{S}(\widehat{\boldsymbol{M}}^v|(p,\boldsymbol{Q}))|$$
$$- \mathbf{M}\log(\sigma(\widehat{\boldsymbol{M}})) - (1-\mathbf{M})\log(1-\sigma(\widehat{\boldsymbol{M}})), \quad (9)$$

where $\sigma(\cdot)$ denotes the sigmoid function, and γ is the hyper-parameter to balance the two losses.

Experiments

Datasets

DAVIS. DAVIS2016 and DAVIS2017 (Pont-Tuset et al. 2017) are widely used to evaluate VOS methods, where the

former one focuses on object-level segmentation and the latter one is challenging in multiple objects which correspond to different targets. It consists of 60 videos in training set and 30 videos in the validation set. It also provides extra test-dev data with 30 challenging videos, which contains some similar objects in the same videos and object occlusion or missing in the continues frames. And there are three evaluation metrics: region similarity $\mathcal{J}$, the intersection over union of the estimated segmentation and the ground truth mask; contour accuracy $\mathcal{F}$: F-measure between the contour-based precision and recall; and global mean value $\mathcal{G}$: average score of $\mathcal{J}$ and $\mathcal{F}$.

YouTube-VOS. YouTube-VOS (Xu et al. 2018) is the latest large-scale dataset which consists of 4453 videos annotated with multiple objects. The validation set contains 474 videos including 91 object categories, and it has separate measures for 65 of seen and 26 of unseen object categories. Like DAVIS dataset, we adopt $\mathcal{J}$, $\mathcal{F}$ and $\mathcal{G}$ for evaluation.

Others. The SegTrack-v2 dataset (Li et al. 2013) consists of 14 test video sequences with 24 objects. YouTube-Objects (Prest et al. 2012) comprises 126 video sequences which belong to 10 object categories. Following the protocol, we use metric $\mathcal{J}$ to measure the segmentation performance.

Implementation Details

To adapt the Mask R-CNN network to generate class-agnostic foreground object bboxs offline, we first train it on COCO dataset with the pre-trained weights on ImageNet. Then we finetune it on DAVIS2017 and YouTube-VOS respectively. In testing phase, we set detection confidence as 0.05 and non-maxminum suppression as 0.6. To feed the bbox proposals to segmentation module Deeplabv3+, we crop the bbox of the four channel input by using the spatial information of the annotation with a margin ratio 0.15. Then we resize the cropped data into 512×512, jitter the image color. Similar to the training process of Mask R-CNN, we first pre-train Deeplabv3+ on COCO dataset, and then train it on DAVIS and YouTube-VOS with learning rate 1e-5 for 100 epochs respectively.

To train our spatiotemporal graph together with the two feature extractors, we set Adam optimizer with learning rate 0.1 which reduces by power of 0.9 for every 10 epochs. We adopt $n = 10$ to store history in our motion mechanism. The balanced ratio α and β in Eq.(3) are set to 0.7 and 0.3 for DAVIS and YouTube-VOS, 0.1 and 0.9 for SegTrack-v2 and YouTube-Objects. And we set the parameters λ_1 and λ_2 in Eq.(6) to 0.4 and 0.6 respectively for their relative importance. The γ in Eq.(9) is set to 100. The thr is set to 0.2. All experiments are implemented on a single NVIDIA 1080Ti GPU. Our codes and trained model will be available online.

Experimental Results

DAVIS. For experiments on DAVIS, we only train our STG-Net on DAVIS 2017 training set. We compare with a wide range of recent competitors both on the DAVIS2016 and DAVIS2017 datasets. Compared with approaches without online learning in Table 1, our method achieves the state-of-the-art performance and outperforms others over most evaluation criteria. Especially on DAVIS2017 test-dev set

Method	OL	DAVIS2017 test-dev			DAVIS2017 val			DAVIS2016 val		
		$\mathcal{J}_\mathcal{M}$	$\mathcal{F}_\mathcal{M}$	$\mathcal{G}_\mathcal{M}$	$\mathcal{J}_\mathcal{M}$	$\mathcal{F}_\mathcal{M}$	$\mathcal{G}_\mathcal{M}$	$\mathcal{J}_\mathcal{M}$	$\mathcal{F}_\mathcal{M}$	$\mathcal{G}_\mathcal{M}$
OSMN (Yang et al. 2018)	✗	37.7	44.9	41.3	52.5	57.1	54.8	74.0	72.9	73.5
SiamMask (Wang et al. 2019a)	✗	40.6	45.8	43.2	54.3	58.5	56.4	71.7	67.8	69.8
FAVOS (Cheng et al. 2018)	✗	42.9	44.2	43.6	54.6	61.8	58.2	82.4	79.5	81.0
RVOS (Ventura et al. 2019)	✗	47.9	52.6	50.3	57.5	63.6	60.6	-	-	-
AGAME (Johnander et al. 2019)	✗	49.2	55.3	52.3	68.5	73.6	71.0	81.5	82.2	81.9
RGMP (Wug Oh et al. 2018)	✗	51.3	54.4	52.8	64.8	68.6	66.7	81.5	82.0	81.8
AGSS (Lin, Qi, and Jia 2019)	✗	51.5	57.1	54.3	63.4	69.8	66.6	-	-	-
FEEL (Voigtlaender et al. 2019)	✗	51.2	57.5	54.4	65.9	72.3	69.1	80.3	83.1	81.7
RANet (Wang et al. 2019c)	✗	53.4	57.2	55.3	63.2	68.2	65.7	**85.5**	85.4	85.5
AGSS* (Lin, Qi, and Jia 2019)	✗	54.8	59.7	57.2	64.9	69.9	67.4	-	-	-
FEEL* (Voigtlaender et al. 2019)	✗	55.1	60.4	57.8	69.1	74.0	71.5	81.1	82.2	81.7
Ours	✗	**59.7**	**66.5**	**63.1**	**71.5**	**77.9**	**74.7**	85.4	**86.0**	**85.7**

Table 1: Quantitative comparison of state-of-the-art methods on DAVIS2016 validation, DAVIS2017 validation and test-dev sets. $\mathcal{M}$ denotes the mean value. "OL" indicates online learning with the annotation of the first frame. * indicates the use of YouTube-VOS for pre-training.

Method	OL	$\mathcal{J}_\mathcal{S}$	$\mathcal{J}_\mathcal{U}$	$\mathcal{F}_\mathcal{S}$	$\mathcal{F}_\mathcal{U}$	$\mathcal{G}_\mathcal{M}$	FPS
OSMN	✗	60.0	40.6	60.1	44.0	51.2	8.0
DMM	✗	58.3	41.6	60.7	46.3	51.7	12
SiamMask	✗	60.2	45.1	58.2	47.7	52.8	**55**
RGMP	✗	59.5	-	45.2	-	53.8	7
OnAVOS	✔	60.1	46.6	62.7	51.4	55.2	0.1
RVOS	✗	63.6	45.5	67.2	51.0	56.8	24
S2S	✗	66.7	48.2	65.5	50.3	57.7	6
DMM	✔	60.3	50.6	63.5	57.4	58.0	-
OSVOS	✔	59.8	54.2	60.5	60.7	58.8	0.1
S2S	✔	71.0	55.5	70.0	61.2	64.4	0.1
AGAME	✗	66.9	61.2	-	-	66.0	-
AGSS	✗	71.3	65.5	**75.2**	73.1	71.3	12
Ours	✗	**72.7**	**69.1**	**75.2**	**74.9**	**73.0**	6

Table 2: Quantitative comparison of state-of-the-art methods on YouTube-VOS validation set. '$\mathcal{S}$' and '$\mathcal{U}$' denote the seen and unseen categories. Specially, DMM (Zeng et al. 2019), OnAVOS (Voigtlaender and Leibe 2017), S2S (Xu et al. 2018) and OSVOS (Caelles et al. 2017) contain online learning.

Method	OL	SegTrack-v2	Method	OL	YouTube-Objects
OSVOS	✔	65.4	OSMN	✗	69.0
RGMP	✗	71.1	FEEL	✗	78.9
DMM	✗	76.7	OnAVOS	✔	80.5
Ours	✗	**79.5**	**Ours**	✗	**84.1**

Table 3: Quantitative comparison of state-of-the-art methods on SegTrack-v2 and YouTube-Objects datasets.

ing that our method is more efficient.

Others. We test our STG-Net (trained on DAVIS2017) on SegTrack-V2 and YouTube-Objects datasets directly. Our method also achieves the state-of-the-art performance without online learning as shown in Table 3.

Qualitative Visualization and Analysis

Figure 4 shows qualitative examples of our results, where we choose challenging videos from DAVIS2017, DAVIS2016, and YouTube-VOS datasets. Our method is robust to occlusions and complex motions. We also show the visual comparison in Figure 5 where other detection based method loses the local patch details and matching based method wrongly segment the background visual similar object. Compared to them, our STG-Net reconstructs the mask result and provides better details. The reason is that our model recovers the local patch details in spatial domain by aggregating the contexts from all intra-object proposals. The temporal graph also helps to refine the mask by retrieving the historic mask memory. More qualitative results can be found in our supplementary.

Ablation Study

We conduct thorough ablation study to analyze the effectiveness of different components and hyperparameters of STG-Net on DAVIS2017 test-dev set. Detailed results are shown in Table 4. We start by a baseline model, which directly choose the best mask from the proposals with λ_1, λ_2 and achieves global mean value 53.2.

Effectiveness of the motion model. The baseline model computes the selection score in Eq. (6) using the previous

that contains much occlusion or object missing scenarios, our method achieves global mean value $\mathcal{G}_\mathcal{M}$ of 63.1, which gains a great margin of improvement than others. Compared to FEEL, we outperform them by 4.6%, 6.1% and 5.3% on three metrics, respectively. Since DAVIS2016 only contains single object without challenge scenario, we perform the similar result to others but still gain improvement of 0.2.

YouTube-VOS. For experiments on YouTube-VOS, we train our model on YouTube-VOS, and show the comparison with previous start-of-the-art approaches in Table 2. Our method achieves a new state-of-the-art of global mean value $\mathcal{G}_\mathcal{M}$ 73.0 in terms of overall scores. Compared to the SOTA model AGSS, we outperform it by 1.7%. Beside, we obtain a good trade off between the performance and running time. Compared with approaches without using online learning (like DMM, AGSS), our method achieves better performance on evaluation metrics. Compared with the approaches using online learning (like OSVOS, S2S), our method achieves faster FPS than these model, demonstrat-
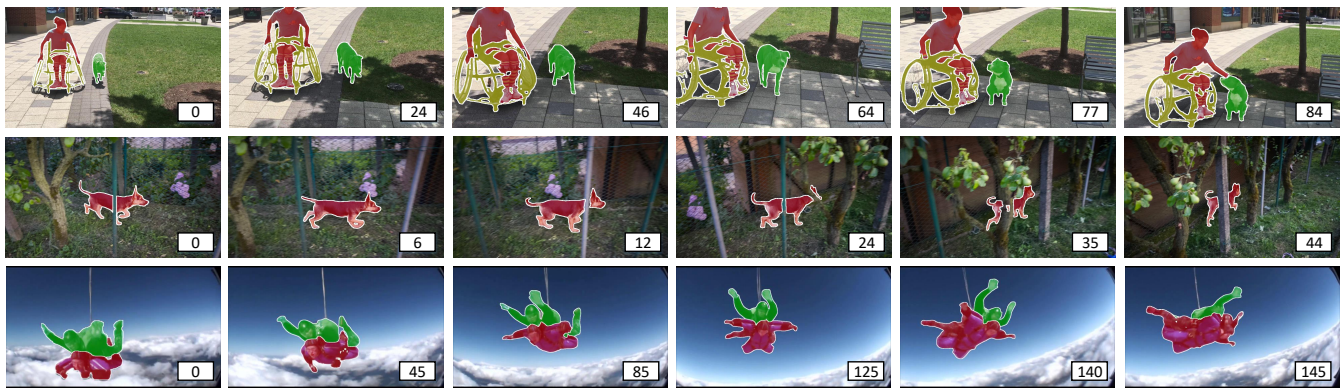
Figure 4: Qualitative results of our method, where frames are sampled at the important moments (*e.g.* multi-view or occlusion) for each video. From top to bottom, the sequences are "girl-dog" in DAVIS2017, "libby" in DAVIS2016 and "0daaddc9da" in YouTube-VOS, respectively.

Figure 5: Visual comparison with detection based (Luiten, Voigtlaender, and Leibe 2018) and matching based (Voigtlaender et al. 2019) methods.

bbox instead of the predicted p. Considering the motion history in the previous frames, our motion model predicts a coarse position p in current frame which is more accurate than the previous bbox. From the table, we find that it has the improvement of performance with 0.9.

Effectiveness of the spatial graph. Our spatial graph reconstructs the mask by utilizing the local patch details from all proposals in current frame. It helps to deal with the occlusion and object missing problems. As shown in the table, our spatial graph construction makes the maximum improvement of 4.3 which recovers the local patch details among intra-object proposals. And the defined hyperparameters of α, which controls the operation of the edge weighting, has another improvement of 0.8.

Effectiveness of the temporal graph. Our temporal graph helps to refine the current frame mask result by using the previous frame masks, which can provides the temporal contextual information for better boundary results. As shown in table, the temporal graph takes mask memory for refinement and has the second maximum improvement of 2.1. The hyperparameters λ_1 in Eq. (6), which controls the operation of mask choosing, has another improvement of 0.6.

How to choose the number of graph layer. We further investigate the performances on different iteration number l of graph updating process. Although more graph layer will bring better performance, too much layers will result in oversmoothing problem. We find that the graph with 2-steps updating achieves the best result of 63.1.

Comparison on different training strategy. We also do the experiments on different training strategy to investigate the benefits. Results show that online leaning with the annotations of the first frame brings the improvement of 1.2. Fine-

Settings						OL	FT	+ytb	$\mathcal{G}_\mathcal{M}$
Mo	SG	α	l	λ_1	TG				
✗	✗	-	-	0.5	✗	✗	✗	✗	53.2
✔	✗	-	-	0.5	✗	✗	✗	✗	54.1
✔	✔	0.5	1	0.5	✗	✗	✗	✗	58.4
✔	✔	0.7	1	0.5	✗	✗	✗	✗	59.2
✔	✔	0.7	1	0.4	✗	✗	✗	✗	59.8
✔	✔	0.7	1	0.4	✔	✗	✗	✗	61.9
✔	✔	0.7	2	0.4	✔	✗	✗	✗	<u>63.1</u>
✔	✔	0.7	3	0.4	✔	✗	✗	✗	62.3
✔	✔	0.7	2	0.4	✔	✔	✗	✗	64.3
✔	✔	0.7	2	0.4	✔	✔	✔	✗	66.5
✔	✔	0.7	2	0.4	✔	✔	✔	✔	**69.4**

Table 4: Ablation study evaluated on the DAVIS2017 test-dev set. 'Mo' means the motion model, 'SG' and 'TG' mean the spatial and temporal graph. α controls edge weights where $\alpha = 1 - \beta$, and λ_1 controls the mask selection score where $\lambda_1 = 1 - \lambda_2$. l represents the spatial graph iteration number. 'OL' indicates online learning and 'FT' indicates fine-tuning with lucid (Khoreva et al. 2017) augmentation. '+ytb' denotes pre-training on YouTube-VOS dataset.

tuning on the augmented train-set of DAVIS2017 makes the improvement of 2.2. And we add YouTube-VOS dataset for pre-training, it has another improvement of 2.9.

Conclusion

In this paper, we propose a spatiotemporal graph neural network (STG-Net) for video object segmentation. By the co-operation of spatial and temporal graph networks, STG-Net has sufficient capacity to reconstruct more detailed masks. Contrasted to the previous detection based approaches utilizing greed search strategy only in the current frame, the abundant use of both local patch details in spatial graph and time dimension relationships in temporal graph make STG-Net able to obtain a superior representation for the class-agnostic objects segmentation. Extensive experiments demonstrate that our method is robust to challenging scenarios thanks to our spatiotemporal graph, and outperforms state-of-the-art methods on all four benchmarks, even compared to online learning methods.

References

Bao, L.; Wu, B.; and Liu, W. 2018. CNN in MRF: Video object segmentation via inference in a CNN-based higher-order spatio-temporal MRF. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5977–5986.

Caelles, S.; Maninis, K.-K.; Pont-Tuset, J.; Leal-Taixé, L.; Cremers, D.; and Van Gool, L. 2017. One-shot video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 221–230.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018a. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 801–818.

Chen, Y.; Pont-Tuset, J.; Montes, A.; and Van Gool, L. 2018b. Blazingly fast video object segmentation with pixel-wise metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1189–1198.

Cheng, D.; Wang, X.; Zhang, Y.; and Zhang, L. 2020a. Graph Neural Network for Fraud Detection via Spatial-temporal Attention. *IEEE Transactions on Knowledge and Data Engineering* .

Cheng, D.; Xiang, S.; Shang, C.; Zhang, Y.; Yang, F.; and Zhang, L. 2020b. Spatio-Temporal Attention-Based Neural Network for Credit Card Fraud Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 362–369.

Cheng, J.; Tsai, Y.-H.; Hung, W.-C.; Wang, S.; and Yang, M.-H. 2018. Fast and accurate online video object segmentation via tracking parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7415–7424.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *International Conference on Machine Learning (ICML)*, 1263–1272.

Haller, E.; Florea, A. M.; and Leordeanu, M. 2019. Space-time Graph Optimization for Video Object Segmentation. *arXiv preprint arXiv:1907.03326* .

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hu, Y.-T.; Huang, J.-B.; and Schwing, A. G. 2018. Video-match: Matching based video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 54–70.

Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; and Brox, T. 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2462–2470.

Johnander, J.; Danelljan, M.; Brissman, E.; Khan, F. S.; and Felsberg, M. 2019. A generative appearance model for end-to-end video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8953–8962.

Khoreva, A.; Benenson, R.; Ilg, E.; Brox, T.; and Schiele, B. 2017. Lucid data dreaming for multiple object tracking. *International Journal of Computer Vision (IJCV)* .

Li, F.; Kim, T.; Humayun, A.; Tsai, D.; and Rehg, J. M. 2013. Video segmentation by tracking many figure-ground segments. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2192–2199.

Li, Q.; Han, Z.; and Wu, X.-M. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Li, X.; Qi, Y.; Wang, Z.; Chen, K.; Liu, Z.; Shi, J.; Luo, P.; Tang, X.; and Loy, C. C. 2017. Video object segmentation with re-identification. In *CVPR Workshop*.

Lin, H.; Qi, X.; and Jia, J. 2019. AGSS-VOS: Attention Guided Single-Shot Video Object Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3949–3957.

Liu, D.; Qu, X.; Liu, X.; Dong, J.; Zhou, P.; and Xu, Z. 2020. Jointly Cross- and Self-Modal Graph Attention Network for Query-Based Moment Localization. In *Proceedings of the 28th ACM International Conference on Multimedia (MM'20)*, 4070–4078.

Lu, X.; Wang, W.; Danelljan, M.; Zhou, T.; Shen, J.; and Van Gool, L. 2020. Video object segmentation with episodic graph memory networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Luiten, J.; Voigtlaender, P.; and Leibe, B. 2018. PReMVOS: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision (ACCV)*, 565–580.

Oh, S. W.; Lee, J.-Y.; Xu, N.; and Kim, S. J. 2019. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 9226–9235.

Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; and Van Gool, L. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675* .

Prest, A.; Leistner, C.; Civera, J.; Schmid, C.; and Ferrari, V. 2012. Learning object class detectors from weakly annotated video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3282–3289.

Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; and Monfardini, G. 2008. The graph neural network model. *IEEE Transactions on Neural Networks* 20(1): 61–80.

Sukhbaatar, S.; Weston, J.; Fergus, R.; et al. 2015. End-to-end memory networks. In *NIPS*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*.

Ventura, C.; Bellver, M.; Girbau, A.; Salvador, A.; Marques, F.; and Giro-i Nieto, X. 2019. Rvos: End-to-end recurrent network for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5277–5286.

Voigtlaender, P.; Chai, Y.; Schroff, F.; Adam, H.; Leibe, B.; and Chen, L.-C. 2019. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 9481–9490.

Voigtlaender, P.; and Leibe, B. 2017. Online adaptation of convolutional neural networks for video object segmentation. In *BMVC*.

Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; and Torr, P. H. 2019a. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1328–1338.

Wang, W.; Lu, X.; Shen, J.; C, D.; and S, L. 2019b. Zeroshot video object segmentation via attentive graph neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9236–9245.

Wang, Z.; Xu, J.; Liu, L.; Zhu, F.; and Shao, L. 2019c. Ranet: Ranking attention network for fast video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3978–3987.

Wug Oh, S.; Lee, J.-Y.; S, K.; and JK, S. 2018. Fast video object segmentation by reference-guided mask propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7376–7385.

Xu, N.; Yang, L.; Fan, Y.; Yang, J.; Yue, D.; Liang, Y.; Price, B.; Cohen, S.; and Huang, T. 2018. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 585–601.

Xu, S.; Liu, D.; Bao, L.; Liu, W.; and Zhou, P. 2019. MHP-VOS: Multiple hypotheses propagation for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 314–323.

Yang, L.; Wang, Y.; Xiong, X.; Yang, J.; and K, A. 2018. Efficient video object segmentation via network modulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6499–6507.

Zeng, X.; Liao, R.; Gu, L.; Xiong, Y.; Fidler, S.; and Urtasun, R. 2019. DMM-Net: Differentiable Mask-Matching Network for Video Object Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3929–3938.