Rethinking Graph Neural Networks for Unsupervised Video Object Segmentation

Daizong Liu
dzliu@stu.pku.edu.cn
Wei Hu*
forhuwei@pku.edu.cn

Wangxuan Institute of Computer Technology,
Peking University,
Beijing, China

Abstract

This paper addresses the task of video object segmentation in an unsupervised manner. Prevailing solutions can be grouped into two categories: 1) two-stream approaches combine both local motion and appearance information, which heavily rely on the quality of optical flow and are not robust to occluded or static objects; 2) appearance matching approaches utilize Siamese networks to learn the relation between two frames (generally the first frame and the current frame), which lack robustness to the appearance variation in long videos. Although recent attentive graph neural networks tackle the above two limitations in an appearance matching manner by matching multiple frames at the same time, the performance is inferior to the counterparts thus far. In this paper, we argue that the performance of such attentive graph model is severely underestimated by current limited designs, including both the node design and the global graph matching. To this end, we develop a novel attentive graph-based model: **R**egion-wise **G**lobal-graph with **B**oundary-aware **L**ocal-learning (**RGBL**). Regarding the node design of the global graph network, instead of taking the whole image as a frame-wise node, RGBL predicts the foreground region in each frame and takes the corresponding regional features as the nodal input to filter out the background noise, which incidentally mitigates the noisy visual similarity among frames. Regarding the global graph matching, RGBL learns more local saliency in individual frames, which incorporates the boundary information to emphasize on the features along the foreground boundary for mask refinement in each frame. Extensive experiments on three challenging benchmarks show that our RGBL surpasses the state-of-the-arts with a large margin.

1 Introduction

Unsupervised video object segmentation (UVOS) aims to segment the most prominent and distinct objects in a video sequence without any prior knowledge of the foreground objects. Due to the lack of human intervention, this task faces significant challenges in effectively tackling visual similarity, occlusions, and appearance variation. Early non-deep-learning methods typically address this task by using handcrafted features, such as objectness [45], motion boundary [31], saliency [39], and trajectories [29], without using any training data. Recently, benefiting from the establishment of large datasets [32], more research efforts have

* indicates the corresponding author. This work was supported by National Natural Science Foundation of China (61972009).

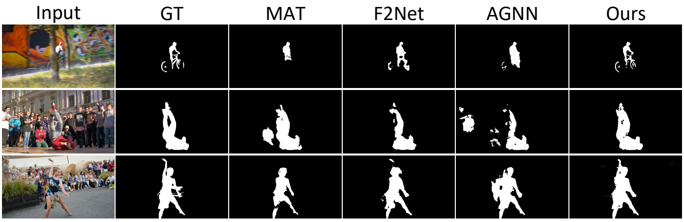


Figure 1: Results on DAVIS 2016. Compared to existing methods MAT (two-stream networks), F2Net (siamese appearance matching) and AGNN (graph-based appearance matching), our method is more robust to object occlusion, visual similarity, appearance changing.

been devoted to solving this unsupervised task in deep learning frameworks. They generally learn more powerful object representations from large-scale training data, and adapt the models to test videos without any annotations.

Existing deep learning frameworks can be grouped into two categories: 1) *two-stream networks* [3, 9, 12, 15, 34, 46]. This class of methods fuse both appearance and motion information via optical flows, which may fail to correctly infer the foreground when the object is occluded or nearly static [35]. 2) *appearance matching networks* [2, 6, 14, 24, 28, 44]. These methods explore the correlation between two frames by simply learning similarities between their pixel-wise embeddings without motion contexts. Since they all adopt Siamese networks to match the current frame only with the first frame, they cannot handle the appearance changing problem in long videos. Figure 1 shows that the above two types of networks are less effective to track and distinguish the foreground objects well.

To avoid the above drawbacks, the attentive graph neural network (AGNN) [40] was proposed to address UVOS in a pure appearance matching manner without using optical flows like two-stream networks. Compared to the other appearance matching networks, AGNN does not deploy Siamese networks to learn the limited relation between two frames. Instead, it handles multiple frames at the same time, which alleviates the problem of appearance drift in long videos. Specifically, the attentive graph neural network mainly consists of three components: a node-wise feature extraction module takes the whole frame as input to learn corresponding appearance representations; the global graph module explores the pixel-wise relations among multiple frames; the readout module decodes the updated node-wise features to predict the segmentation result for each frame. Although this GNN-based paradigm attempts to eliminate the shortcomings in previous frameworks, the performance is still inferior to that of other models thus far. We argue that the main reasons come from the limited architecture in the following two aspects:

Node design: 1) Each video contains complex and diverse scenes, *e.g.*, each frame may contain visually similar objects in the background. Thus, distinguishing the foreground and background objects is crucial to track the target object well. However, the existing node design of the global graph takes the whole frame as input and results in mismatching to similar objects in the background regions. 2) The target object generally appears only in a small region of each frame. Therefore, instead of matching features of the whole frame as in AGNN, matching the regions that only contain target objects is able to reduce useless computation and produce more fine-grained results.

Global graph matching: 1) To determine the foreground object, there are two essential properties: distinguishable in an individual frame (locally salient), and frequently appearing throughout the video sequence (globally consistent). However, the global graph-based network only focuses on finding the most frequently appearing object among frames, but fails to explore the salient information in individual frames. 2) Since frame-wise global matching endeavors more to locate a possible area of the target object but suffers from the ambiguity of boundary pixels, it is important to capture more local details in individual frames for refining the mask boundary of the foreground object.

To this end, we propose a novel framework called **R**egion-wise **G**lobal-graph with **B**oundary-aware **L**ocal-learning (**RGBL**), with delicate node design and local graph refinement for local-global representation learning, by rethinking and addressing the limitations of the existing AGNN model. **For the node design**, RGBL extracts *regional* features of each frame as the nodal input so as to filter out the background noise. In particular, we first develop a foreground localization branch to detect the region of the most salient object in each frame, and then obtain corresponding regional features by deploying a regional attention to the features of the entire frame. In this manner, our model not only locates possible regions of the target object better, but also alleviates mismatching to similar objects in the background. **For the global graph matching**, RGBL learns *boundary-aware* local saliency in each frame. Specifically, we first extract the object boundary by developing a boundary prediction approach on the extracted regional features, and then introduce a graph-based boundary attention to emphasize on the local features of boundary pixels during the pixel-wise feature matching, thus enforcing accurate segmentation along the object boundary.

We demonstrate the effectiveness of RGBL on three challenging UVOS benchmarks: DAVIS2016 [32], Youtube-Objects [33] and FBMS [30]. Experimental results show that our RGBL model achieves the state-of-the-art performance over all benchmarks and metrics.

2 Related Work

UVOS is a video based task [17, 18, 20, 21, 22, 25, 26, 27]. Traditional methods require no training data and typically utilize handcrafted features [5, 8, 19, 23, 31, 37, 43] for segmentation. Recently, benefiting from the establishment of large datasets [32], many approaches have been proposed to solve this task in deep learning frameworks to improve the performance further. Tokmakov *et al.* [35] proposed a purely optical flow based network that discards appearance modelling and casts segmentation as foreground motion prediction, which is not advantageous to static objects. To address this challenge, two-stream networks are introduced to fuse both appearance and motion information [3, 9, 12, 15, 34, 46]. However, these methods severely rely on the motion contexts in optical flow, thus suffering from the high computation complexity and the deterioration in the quality of the optical flows. Targeting this issue, recent approaches [2, 6, 14, 28, 44] tackle video object segmentation by simply learning similarities between pixel-wise embeddings without motion contexts. However, a major drawback of these approaches is that they all utilize a Siamese network to learn the correlation between two frames, which is thus not robust to the appearance drift in long videos. Therefore, we extend the AGNN [40] discussed in the Introduction to exploit a unified graph attention network, which handles multiple frames at the same time for capturing rich and inherent correlation within videos.

3 Method

We propose a novel local-global graph-based model RGBL for UVOS as illustrated in Figure 2, which mainly consists of three modules: **Regional feature extraction:** Given a video sequence, instead of encoding the whole image, we propose the foreground localization to

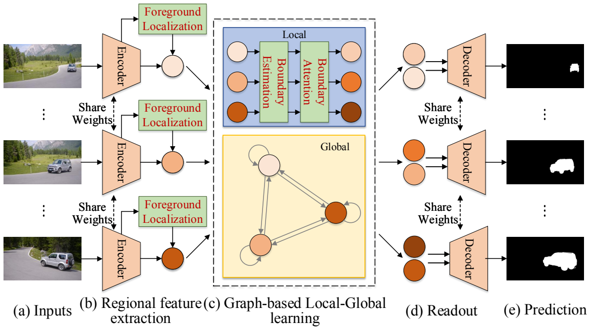


Figure 2: The overall architecture of our proposed RGBL model.

coarsely locate the region of the salient object in each frame and extract corresponding regional features to filter out the background noise. **Graph-based local-global learning:** This is designed to jointly capture the global correlations among frames and the local contexts within individual frames. In particular, the local learning module introduces boundary attention to aggregate contexts with pixel-to-pixel similarities along the boundary. Further, the global learning module takes regional features as nodal features and matches the appearance of these regions. **Readout module:** At last, we deploy a readout module to fuse both local and global features of each frame for more accurate segmentation.

3.1 Regional Feature Extraction

Given a video sequence $\mathcal{I} = \{\boldsymbol{I}_t\}_{t=1}^{T}$ of T frames, we leverage DeepLabV3 [1] as the main backbone, which consists of five convolution blocks from ResNet50 [7] and an atrous spatial pyramid pooling (ASPP) module, to extract effective frame-wise features. We denote the extracted embeddings of $\{\boldsymbol{I}_t\}_{t=1}^{T}$ as $\{\boldsymbol{F}_t\}_{t=1}^{T}$, where $\boldsymbol{F}_t \in \mathbb{R}^{W\times H\times C}$. We first introduce a foreground localization branch to locate the region of the primary object to filter out the background noise. Then, the regional features are extracted by the Hadamard product of the regional attention maps and the extracted frame-level embeddings.

Foreground localization. Considering the center point of an object can be taken as the spatial prior [24, 42, 47], we attempt to firstly locate the center point of the most salient object and then generate the Gauss map from the point to its surrounding for covering the whole object. Particularly, we transform the point localization task into a Gauss-based heatmap prediction task [36]. We propose a foreground localization branch as shown in Figure 3 (a). Specifically, this branch directly predicts the center point of the salient object without resorting to any motion information, and introduces a two-level coarse-to-fine supervision strategy. At each down-sampling path, we follow the down-sampling strategy of ResNet50 backbone to embed the image into different scales. At each up-sampling path, we upsample the embeded features with bilinear interpolation and further employ a convolutional layer. We also propose a cross-level feature aggregation strategy to propagate multi-scale features from coarse-level to fine-level for information strengthening. As shown in Figure 3 (a), let us denote the learned features in two-level down- and up-sampling modules of frame $\boldsymbol{I}_t$ as $\{\boldsymbol{D}_t^{i,j}\}_{j=1}^{4}, \{\boldsymbol{U}_t^{i,j}\}_{j=1}^{4}, i = \{1,2\}$, where $i=1$ refers to the coarse-level and $i=2$ refers to the fine-level, and j refers to the layer of the network. The cross-level feature aggregation is:

$$(\boldsymbol{D}_t^{2,j})' = \mathrm{Conv2d}(\boldsymbol{D}_t^{1,j}) + \mathrm{Conv2d}(\boldsymbol{U}_t^{1,j}), \quad \boldsymbol{D}_t^{2,j} = (\boldsymbol{D}_t^{2,j})' + \mathrm{Conv2d}(\boldsymbol{D}_t^{2,j-1}), \tag{1}$$

where $\mathrm{Conv2d}(\cdot)$ denotes the 2D convolutional layer. At last, we directly predict the coarse-

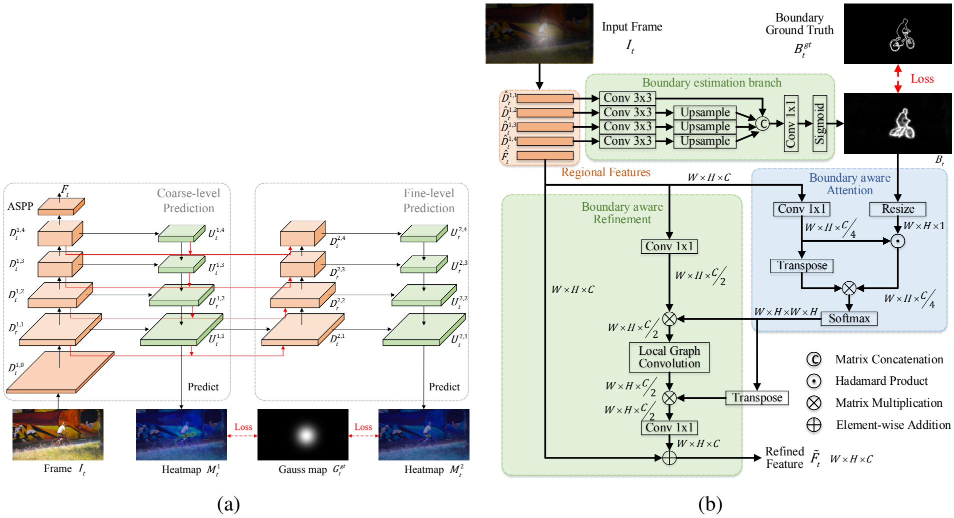


Figure 3: (a) Foreground localization module. (b) Local context learning module.

and fine-level heatmap $\boldsymbol{M}_t^1, \boldsymbol{M}_t^2$ on $\boldsymbol{U}_t^{1,1}, \boldsymbol{U}_t^{2,1}$ by applying a 3×3 convolutional layer with ReLU, followed by another 1×1 convolutional layer and a sigmoid function. We choose the best center $\boldsymbol{p}_t = (x_t, y_t) \in \mathbb{R}^2$ from the fine-level heatmap $\boldsymbol{M}_t^2$ with the maximum score, and encode it into Gauss map $\boldsymbol{G}_t \in \mathbb{R}^{W \times H}$ [24, 36] to cover the foreground region.

Regional features. We deploy a regional attention map $\boldsymbol{A}_t \in \mathbb{R}^{W \times H}$ to the embedding of the whole frame $\boldsymbol{F}_t \in \mathbb{R}^{W \times H \times C}$. $\boldsymbol{A}_t$ is obtained by:

$$\boldsymbol{A}_t(x,y) = \begin{cases} 1, & \text{if } \boldsymbol{G}_t(x,y) > 0 \\ 0, & \text{else} \end{cases}. \tag{2}$$

where $(x,y) \in \mathbb{R}^{W \times H}$, and the regional feature $\widehat{\boldsymbol{F}}_t$ is generated by $\widehat{\boldsymbol{F}}_t = \boldsymbol{A}_t \odot \boldsymbol{F}_t$.

3.2 Graph-based Local-Global Learning

Our RGBL considers both local and global contexts to jointly learn the local saliency and global consistency. For the local context learning, RGBL first estimates the boundary of the foreground object in each individual frame, and then employs boundary-aware attention to emphasize on features of boundary pixels during the pixel-wise graph convolution for boundary refinement. For the global context learning, RGBL takes the salient region of each frame as a node to build a global graph for matching.

Local context learning. We first estimate the boundary information. After obtaining the generated Gauss map $\boldsymbol{G}_t$ and the down-sampling features $\{\boldsymbol{D}_t^{1,j}\}_{j=1}^4$ of each individual frame $\boldsymbol{I}_t$, we acquire their regional features $\{\widehat{\boldsymbol{D}}_t^{1,j}\}_{j=1}^4$ as operated in Eq. (2). As shown in Figure 3 (b), we feed these features into a boundary estimation branch with a multi-scale context aggregation strategy to output the estimated boundary $\boldsymbol{B}_t$.

Secondly, given the regional feature $\widehat{\boldsymbol{F}}_t$ and the predicted boundary map $\boldsymbol{B}_t$, we introduce a boundary aware attention module to emphasize on the features of boundary pixels:

$$\boldsymbol{P}_t = \text{Softmax}((\text{Conv2d}(\widehat{\boldsymbol{F}}_t) \odot \boldsymbol{B}_t)(\text{Conv2d}(\widehat{\boldsymbol{F}}_t))^\top), \tag{3}$$

where Conv2d$(\cdot)$ is utilized to reduce the feature dimension. The Hadamard product $\odot$ essentially assigns a weight to the feature of each pixel, with larger weights to features of boundary pixels. The matrix multiplication captures the feature similarity between boundary pixels and all pixels of the frame, and the softmax function is for normalization. Eq. (3) leads to the attention map $\boldsymbol{P}_t \in \mathbb{R}^{(W\times H)\times(W\times H)}$. More details are presented in Figure 3 (b). With the acquired attention map $\boldsymbol{P}_t$, we aggregate pixels with similar features as the boundary pixels to bridge the connection between all pixels and the boundary pixels by:

$$(\widehat{\boldsymbol{F}}_t)' = \boldsymbol{P}_t(\mathrm{Conv2d}(\widehat{\boldsymbol{F}}_t)). \tag{4}$$

Next, to explore the connectivity between the boundary pixels, we propose a boundary-aware local graph to propagate information across these pixels to learn their higher-level semantic relations. Each pixel is taken as a node, and we fully connect all boundary pixels to build the local graph. Specifically, we utilize the intra-attention mechanism [38, 41] to calculate the response at one pixel position by attending to the other positions, and employ a single-layer graph convolution network [10] to reason their correlations as:

$$\boldsymbol{E}_t = \mathrm{softmax}(((\widehat{\boldsymbol{F}}_t)'\boldsymbol{W}_1)((\widehat{\boldsymbol{F}}_t)'\boldsymbol{W}_2)^\top), \quad (\widehat{\boldsymbol{F}}_t)'' = \mathrm{ReLU}[(\boldsymbol{I}-\boldsymbol{E}_t)(\widehat{\boldsymbol{F}}_t)'\boldsymbol{W}_3], \tag{5}$$

where $\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_3$ are learnable weights.

At last, we refine the local pixel features of the foreground based on the updated boundary-aware features, and take the sum of the refined and the original features as the output:

$$\widetilde{\boldsymbol{F}}_t = \mathrm{Conv2d}(\boldsymbol{P}_t^\top(\widehat{\boldsymbol{F}}_t)'') + \widehat{\boldsymbol{F}}_t. \tag{6}$$

Global context learning. We build a global graph among multiple video frames to capture the global correlations. Different from the AGNN [40] where the feature of each frame is taken as the signal on a node, we take the regional features $\widehat{\boldsymbol{F}}_t$ as the signal over a node so as to filter out the background noise and alleviate the mismatching problem. We initialize the node features as $\boldsymbol{H}_t^0 = \widehat{\boldsymbol{F}}_t$, and denote the final updated features as $\boldsymbol{H}_t^K$, where K is the number of the global graph network layers.

3.3 Readout Module

Having captured both the local boundary details and global object consistency, we concatenate the local refined features $\widetilde{\boldsymbol{F}}_t$ and the global graph node feature $\boldsymbol{H}_t^K$, and feed the combined feature into a readout module for segmenting the final mask result $\boldsymbol{R}_t$ (as shown in Figure 2). To preserve the spatial information, our readout module is composed of three convolution layers and a sigmoid function.

3.4 Training Loss

Given a video sequence, for each frame $\boldsymbol{I}_t$, our RGBL predicts the heatmaps $\boldsymbol{M}_t^1, \boldsymbol{M}_t^2$ of center points at both the coarse- and fine-level, an estimated boundary $\boldsymbol{B}_t$ and a mask result $\boldsymbol{R}_t$. For the point heatmap prediction, we apply an element-wise focal loss [16] on each predicted heatmap $\boldsymbol{M}_t^i$ and the ground-truth Gauss map $\boldsymbol{G}_t^{gt}$ as:

$$\mathcal{L}_f^i = \sum_{(x,y)} \begin{cases} (1-M^i_{t,(x,y)})^\alpha \log(M^i_{t,(x,y)}), \text{ if } G^{gt}_{t,(x,y)} = 1 \\ (1-G^{gt}_{t,(x,y)})^\beta (M^i_{t,(x,y)})^\alpha \log(1-M^i_{t,(x,y)}), \text{ o.w.} \end{cases} \tag{7}$$

where $i \in \{1,2\}$, $M^i_{t,(x,y)}$ is the score at location (x,y) in the predicted heatmap $\boldsymbol{M}_t^i$, and we set α as 2 and β as 4 following the default setting in [11]. For the boundary estimation, we

Table 1: Quantitative results of UVOS methods on the DAVIS2016 validation set.

	Method	FSEG	LVO	ARP	PDB	LSMO	MoA	EpO	AGS	AGNN	COS	AGNN*	AnDiff	MAT	F2Net	**RGBL**
$\mathcal{J}$&$\mathcal{F}$	Mean↑	68.0	74.0	73.4	75.9	77.1	77.3	78.1	78.6	79.9	80.0	80.5	81.1	81.5	83.7	**85.6**
	Mean↑	70.7	75.9	76.2	77.2	78.2	77.2	80.6	79.7	80.7	80.5	81.3	81.7	82.4	83.1	**85.2**
$\mathcal{J}$	Recall↑	83.5	89.1	91.1	90.1	89.1	87.8	95.2	91.1	94.0	93.1	93.1	90.9	94.5	95.7	**96.8**
	Decay↓	1.5	**0.0**	7.0	0.9	4.1	5.0	2.2	1.9	**0.0**	4.4	4.4	2.2	5.5	**0.0**	**0.0**
	Mean↑	65.3	72.1	70.6	74.5	75.9	77.4	75.5	77.4	79.1	79.5	79.7	80.5	80.7	84.4	**86.1**
$\mathcal{F}$	Recall↑	73.8	83.4	83.5	84.4	84.7	84.4	87.9	85.8	90.5	89.5	88.5	85.1	90.2	92.3	**93.9**
	Decay↓	1.8	1.3	7.9	**-0.2**	3.5	3.3	2.4	1.6	0.0	5.0	5.1	0.6	4.5	0.8	0.1
$\mathcal{T}$	Mean↓	32.8	26.5	39.3	29.1	21.2	27.9	19.3	26.7	33.7	**18.4**	33.7	21.4	21.6	20.9	28.8

treat it as pixel-wise binary classification and employ the binary cross-entropy loss on the predicted boundary $\boldsymbol{B}_t$ and ground truth $\boldsymbol{B}_t^{gt}$ as:

$$\mathcal{L}_b = -\sum_{(x,y)} B^{gt}_{t,(x,y)} \log(B_{t,(x,y)}) + (1 - B^{gt}_{t,(x,y)}) \log(1 - B_{t,(x,y)}), \tag{8}$$

where $B_{t,(x,y)}$ denotes the location (x,y) in the boundary $\boldsymbol{B}_t$. For the segmentation result, we deploy the binary cross-entropy loss on the predicted mask $\boldsymbol{R}_t$ and the ground truth $\boldsymbol{R}_t^{gt}$ as:

$$\mathcal{L}_s = -\sum_{(x,y)} R^{gt}_{t,(x,y)} \log(R_{t,(x,y)}) + (1 - R^{gt}_{t,(x,y)}) \log(1 - R_{t,(x,y)}). \tag{9}$$

Finally, our RGBL is trained end-to-end from scratch using the multi-task loss $\mathcal{L} = \lambda_1 \mathcal{L}_s + \lambda_2 \mathcal{L}_b + \frac{\lambda_3}{2} \sum_{i=1}^{2} \mathcal{L}_f^i$ where λ_1, λ_2, λ_3 are parameters to strike a balance among the three terms.

4 Experiments

4.1 Datasets and Metrics

DAVIS2016. This dataset is a challenging video object segmentation dataset [32] which consists of 50 videos in total (30 for training and 20 for validation) with pixel-wise annotations for every frame. Three evaluation criteria are used following [32]: region similarity $\mathcal{J}$, boundary accuracy $\mathcal{F}$, and time stability $\mathcal{T}$.
Youtube-Objects. It is a large dataset [33] of 126 web videos with 10 object categories and more than 20,000 frames. Following the common protocol, we use the region similarity $\mathcal{J}$ to measure the segmentation performance.
FBMS. This dataset [30] is comprised of 59 video sequences (29 training videos and 30 test videos). As in previous works, we use region similarity $\mathcal{J}$ as the metric.

4.2 Training Details

To obtain multiple video frames as input, we leverage a random sampling strategy to train our RGBL model. Specifically, we split each training video with a total of T frames into T' segments ($T' < T$) and randomly select one frame from each segment. Then we feed the T' sampled frames into a batch and train the model. Such a sampling strategy provides robustness to variations, and the diversity among the samples enables our model to better capture the underlying relationships and improve its generalizability. The size of each RGB frame is $473 \times 473 \times 3$, the input frame number is $T' = 7$ and the number of global graph network layers is $K = 3$. We set $\lambda_1 = \lambda_2 = \lambda_3 = 1$. The entire network is trained using the SGD optimizer with an initial learning rate of 2.5×10^{-4}. We set the batchsize as 16. All the experiments are conducted using 4 V100 GPUs on a server. The overall training time is about 12 hours, and it takes about 0.14s with one image in a forward pass.

Table 2: Fair comparison with RTNet [34] on different backbone models on the DAVIS2016.

Method	Backbone	$\mathcal{J}$&$\mathcal{F}$	$\mathcal{J}$		$\mathcal{F}$	
		Mean↑	Mean↑	Recall↑	Mean↑	Recall↑
RGBL	ResNet50	**85.6**	**85.2**	**96.8**	**86.1**	**93.9**
RTNet	ResNet34	84.1	84.8	95.8	83.5	93.1
RGBL	ResNet34	**85.3**	**85.0**	**96.4**	**85.7**	**93.6**
RTNet	ResNet101	85.1	85.6	96.1	84.7	93.8
RGBL	ResNet101	**86.3**	**86.1**	**97.3**	**86.6**	**94.5**

4.3 Quantitative Performance

Evaluation on DAVIS2016. We compare our RGBL with the top performing UVOS methods in the public leaderboard on the DAVIS2016 dataset, as shown in Table 1. Our RGBL outperforms all the reported methods over most metrics. Compared with the state-of-the-art method F2Net [24], our model achieves improvement of 1.9 in terms of $\mathcal{J}$&$\mathcal{F}$ Mean. Specifically, we obtain gains of 2.1 and 1.7 on $\mathcal{J}$ Mean and $\mathcal{F}$ Mean, respectively. Compared to appearance matching methods COS [28] and AnDiff [44], our RGBL handles multiple frames at the same time which learns more robust global consistency, thus outperforming them both over $\mathcal{J}$ Mean and $\mathcal{F}$ Mean by a large margin. Compared to methods like MAT [46] and EPO [4] which utilize both appearance information and motion cues, our model outperforms them by only utilizing appearance information. Compared to the original graph-based method AGNN [40] which often fails to distinguish visually similar backgrounds and lacks capturing local context, our novel node design with regional features and the boundary-aware local context learning lead to significant improvements. We further implement our RGBL with different backbone models (*i.e.*, ResNet34 and ResNet101) to fairly compare with the RTNet [34]. As shown in Table 2, we achieve better performance than RTNet.

Evaluation on Youtube-Objects. Table 3 lists the results of all compared methods for different categories on the Youtube-Objects dataset. Our approach brings improvement of 1.8 on Mean $\mathcal{J}$ than the state-of-the-art method F2Net by a large margin. It is worth noting that we outperform all compared methods on almost all categories. There are three main reasons: First, for optical guided methods MAT, FSEG [9] and LVO [35], sequences in the Airplane and Boat categories contain objects that have large appearance variation or move slowly, resulting in inaccurate estimation of optical flow. Compared to them, our appearance matching based framework handles these scenarios well. Second, compared to Siamese network based appearance matching methods F2Net and COS, our graph-based model matches multiple frames at the same time, which learns more robust global consistency. Third, compared to the graph-based method AGNN, our foreground localization and local graph alleviate the mismatching problem and refine the object boundary in individual frames.

Evaluation on FBMS. For completeness, we also evaluate our method on FBMS dataset. As shown in Table 5, our RGBL produces the best result over the evaluation metric Mean $\mathcal{J}$, which outperforms the state-of-the-art by 1.2. Since lots of foreground objects in FBMS share similar appearance with the background, our foreground localization branch helps to filter out the visually similar background for better segmentation.

4.4 Ablation Study

We conduct ablation studies on the DAVIS2016 dataset as shown in Table 4, where the baseline model is the original attentive graph neural network AGNN.

Table 3: Quantitative results on Youtube-Objects.

Method	FSEG	LVO	AGNN	COS	AMC	AGNN*	MAT	F2Net	RGBL
Airplane	81.7	86.2	81.1	81.1	78.9	86.0	72.9	85.8	**87.0**
Bird	63.8	81.0	75.9	75.7	80.9	75.7	77.5	82.8	**84.3**
Boat	72.3	68.5	70.7	71.3	67.4	68.7	66.9	81.9	**83.2**
Car	74.9	69.3	78.1	77.6	82.0	**82.4**	79.0	81.4	81.4
Cat	68.4	58.8	67.9	66.5	69.0	65.9	**73.7**	70.2	72.8
Cow	68.0	68.5	69.7	69.8	69.6	70.5	67.4	71.0	**73.2**
Dog	69.4	61.7	**77.4**	76.8	75.8	77.1	75.9	75.8	76.5
Horse	60.4	53.9	67.3	67.4	63.0	72.2	63.2	75.4	**77.1**
Motorbike	62.7	60.8	68.3	67.7	63.4	63.8	62.6	71.8	**73.6**
Train	62.2	**66.3**	47.8	46.8	57.8	47.8	51.0	59.6	64.9
Mean$\mathcal{J}\uparrow$	68.4	67.5	70.8	70.5	71.1	71.4	69.0	75.6	**77.4**

Table 4: Overall ablation studies.

Network Variant	Mean $\mathcal{J}\uparrow$	$\triangle\mathcal{J}$	Mean $\mathcal{F}\uparrow$	$\triangle\mathcal{F}$
Baseline (AGNN)	80.7	-4.5	79.1	-7.0
Node design				
Baseline + CCP	81.9	-3.3	82.0	-4.1
Baseline + CFCP (FL)	83.0	-2.2	83.5	-2.6
Local graph network				
Baseline + FL + BE	83.6	-1.6	84.3	-1.8
Baseline + FL + BE&BA	**85.2**	-	**86.1**	-
Graph layer = 1	**85.2**	-	**86.1**	-
Graph layer = 2	84.5	-0.7	85.7	-0.4
Other Variations				
Input frames $T' = 3$	83.2	- 2.0	83.7	-2.4
Input frames $T' = 5$	84.3	-0.9	85.0	-1.1
Input frames $T' = 7$	**85.2**	-	**86.1**	-
Input frames $T' = 9$	**85.2**	-	**86.1**	-
Global graph $K = 1$	83.9	-1.3	84.4	-1.7
Global graph $K = 3$	**85.2**	-	**86.1**	-
Global graph $K = 5$	84.7	-0.5	85.8	-0.3

Table 5: Quantitative results on FBMS.

Method	APR	MSTP	FSEG	IET	PDB	COS	MAT	AMC	F2Net	RGBL
Mean$\mathcal{J}\uparrow$	59.8	60.8	68.4	71.9	74.0	75.6	76.1	76.5	77.5	**78.7**

Table 6: Complexity comparison on DAVIS2016. “Speed" denotes the average time to segment one image in a forward pass.

Metric	AGNN	MAT	F2Net	**Ours**
$\mathcal{J}\&\mathcal{F}$	79.9	81.5	83.7	**85.6**
Speed (s)	0.12	**0.05**	0.10	0.14
Model Size (M)	**315**	506	432	468

Studies on the node design. We first investigate the effectiveness of our node design, which relies on the coarse-to-fine center prediction (CFCP) in the foreground localization (FL) branch for filtering out the background features in each frame. As shown in Table 4, compare to the baseline model, CFCP (FL) brings the improvement of 2.3 on $\mathcal{J}$ and 4.4 on $\mathcal{F}$, which indicates the effectiveness of our regional node features for filtering out the visually similar objects in the background. Compared to general coarse-level center prediction (CCP), our coarse-to-fine strategy performs better, demonstrating its effectiveness.

Studies on the local graph network. We also investigate the importance of our local graph network. Specifically, we develop a boundary estimation (BE) module to enforce the backbone model to extract the crucial semantic features. We also devise a boundary attention (BA) module to refine the local contexts in each frame. As shown in Table 4, both boundary estimation (BE) and boundary attention (BA) modules contribute a lot to the final performance. We observe that a single-layer graph is enough, more layers will result in oversmoothing [13].

Studies on other variations. To evaluate the impact of the number of input frames T', we report the performance with different T'. Our model achieves the best result with $T' = 7$. For the number of global graph layers K, the performance converges at $K = 3$.

Studies on model complexity. We compare the model complexity in Table 6. Our speed is comparable to AGNN with a reasonable model size.

4.5 Qualitative Results

Visualization on the foreground localization. To investigate the performance of our proposed foreground localization, we provide some visualization results on the generated heatmaps of center points. As shown in Figure 5 (a), there are four challenging sequences in which the surroundings have similar appearance to the foreground object ("breakdance" and "dog") or there exists large appearance drift ("parkour" and "scooter-black"). Without using any motion history, our foreground localization branch achieves better performance than F2Net by only extracting individual semantic features. This demonstrates that our coarse-to-fine strategy effectively captures the salient information to locate the target object.

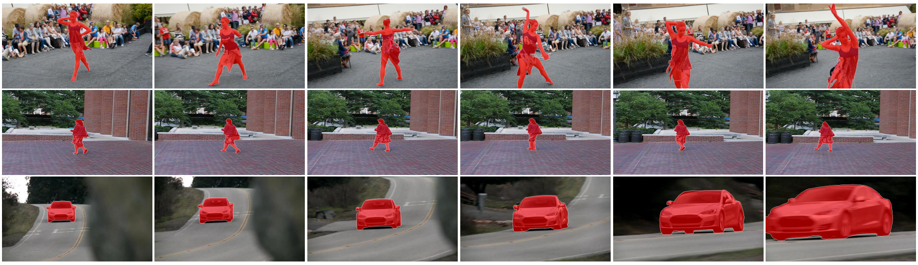
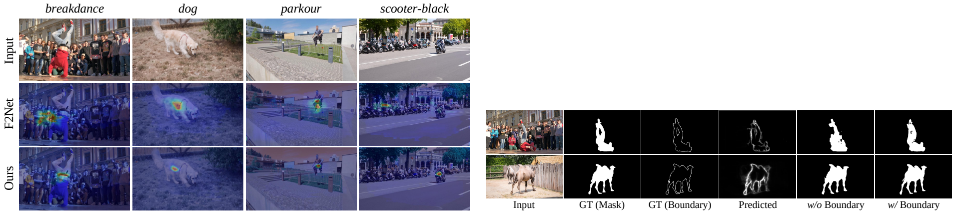


Figure 4: Qualitative results on DAVIS2016, FBMS and Youtube-Objects datasets.



(a) (b)

Figure 5: (a) Visualization of center point heatmaps on the first frames in videos. (b) Visualization of our boundary estimation and mask prediction. The model with local learning (w/ Boundary) obtains better boundary details of the foreground object.

Visualization on the boundary estimation. To evaluate the performance of our proposed boundary-aware local context learning, we demonstrate the visual results on two sequences in Figure 5 (b). We see that our boundary estimation module is able to estimate the object boundary well based on the regional features. Further, as shown in Figure 5 (b), without boundary learning, the model (w/o Boundary) lacks local contexts within individual frames for exploring the local saliency, while learning the boundary information (w/ Boundary) leads to much better prediction of the mask with accurate contours.

Visualization on the mask results. Figure 4 depicts sample results for representative sequences from the three datasets. The *dance-twirl* sequence from DAVIS-16 contains many challenging factors, such as object deformation, motion blur and background visual similarity. We see that our method is robust to these challenges and delineates the target with accurate boundaries. The effectiveness is further validated in *people1* from FBMS and *car0009* from Youtube-Objects, in which the target suffers from large-scale variations.

5 Conclusion

In this paper, we thoroughly analyze the existing attentive graph neural network based method for UVOS, especially the drawbacks of the the current AGNN model. Based on the analysis, we propose a novel local-global graph-based model RGBL for unsupervised video object segmentation (UVOS), which addresses major limitations in existing graph neural network based methods. On the one hand, the RGBL model localizes the center point of the salient object with a coarse-to-fine strategy and extracts the corresponding regional features in each frame to filter out the background noise. On the other hand, RGBL not only takes regional features as signals on nodes to build a global graph, but also emphasizes on the crucial boundary information in individual frames by learning boundary-aware contexts. Extensive experiments on three datasets demonstrate the superiority of our RGBL model. Future works include developing the RGBL model for the multi-object UVOS task.

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[2] Yuhua Chen, Jordi Pont-Tuset, Alberto Montes, and Luc Van Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1189–1198, 2018.

[3] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 686–695, 2017.

[4] Muhammad Faisal, Ijaz Akhter, Mohsen Ali, and Richard Hartley. Exploiting geometric constraints on dense trajectories for motion saliency. In *WACV*, 2019.

[5] Alon Faktor and Michal Irani. Video segmentation by non-local consensus voting. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 2, page 8, 2014.

[6] Alireza Fathi, Zbigniew Wojna, Vivek Rathod, Peng Wang, Hyun Oh Song, Sergio Guadarrama, and Kevin P Murphy. Semantic instance segmentation via deep metric learning. *arXiv preprint arXiv:1703.10277*, 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[8] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 786–802, 2018.

[9] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2126, 2017.

[10] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[11] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[12] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3243–3252, 2018.

[13] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[14] Siyang Li, Bryan Seybold, Alexey Vorobyov, Alireza Fathi, Qin Huang, and C-C Jay Kuo. Instance embedding transfer to unsupervised video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6526–6535, 2018.

[15] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C-C Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 207–223, 2018.

[16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[17] Daizong Liu and Wei Hu. Skimming, locating, then perusing: A human-like framework for natural language video localization. In *ACM Multimedia*, 2022.

[18] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *ACM Multimedia*, 2020.

[19] Daizong Liu, Dongdong Yu, Minghui Dong, Lei Ma, Jie Shao, Jian Wang, Changhu Wang, and Pan Zhou. An effective multi-level backbone for video object segmentation. In *CVPR Workshops*, 2020.

[20] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. Adaptive proposal generation network for temporal sentence localization in videos. In *EMNLP*, pages 9292–9301, 2021.

[21] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[22] Daizong Liu, Xiaoye Qu, and Pan Zhou. Progressively guide to attend: An iterative alignment framework for temporal sentence grounding. In *EMNLP*, 2021.

[23] Daizong Liu, Shuangjie Xu, Xiao-Yang Liu, Zichuan Xu, Wei Wei, and Pan Zhou. Spatiotemporal graph neural network based mask reconstruction for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2100–2108, 2021.

[24] Daizong Liu, Dongdong Yu, Changhu Wang, and Pan Zhou. F2net: Learning to focus on the foreground for unsupervised video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[25] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu Xu, and Pan Zhou. Memory-guided semantic learning network for temporal sentence grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[26] Daizong Liu, Xiaoye Qu, and Wei Hu. Reducing the vision and language bias for temporal sentence grounding. In *ACM Multimedia*, 2022.

[27] Daizong Liu, Xiaoye Qu, Yinzhen Wang, Xing Di, Kai Zou, Yu Cheng, Zichuan Xu, and Pan Zhou. Unsupervised temporal video grounding with deep semantic clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[28] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3623–3632, 2019.

[29] Peter Ochs and Thomas Brox. Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1583–1590, 2011.

[30] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *TPAMI*, 36(6):1187–1200, 2013.

[31] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1777–1784, 2013.

[32] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 724–732, 2016.

[33] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning object class detectors from weakly annotated video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3282–3289, 2012.

[34] Sucheng Ren, Wenxi Liu, Yongtuo Liu, Haoxin Chen, Guoqiang Han, and Shengfeng He. Reciprocal transformations for unsupervised video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15455–15464, 2021.

[35] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3386–3394, 2017.

[36] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, pages 1799–1807, 2014.

[37] Yi-Hsuan Tsai, Ming-Hsuan Yang, and Michael J Black. Video segmentation via object flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3899–3908, 2016.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008, 2017.

[39] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3395–3402, 2015.

[40] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9236–9245, 2019.

[41] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018.

[42] Yuqing Wang, Zhaoliang Xu, Hao Shen, Baoshan Cheng, and Lirong Yang. Centermask: single shot instance segmentation with point representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9313–9321, 2020.

[43] Shuangjie Xu, Daizong Liu, Linchao Bao, Wei Liu, and Pan Zhou. Mhp-vos: Multiple hypotheses propagation for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 314–323, 2019.

[44] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 931–940, 2019.

[45] Dong Zhang, Omar Javed, and Mubarak Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 628–635, 2013.

[46] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 2, page 3, 2020.

[47] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.