



#### **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.

- 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.

#### **MOTIVATION**

To this end, we propose a novel framework called Region-wise Global-graph with Boundaryaware Local-learning (RGBL), with delicate node design and local graph refinement for localglobal 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.

# Rethinking Graph Neural Networks for Unsupervised Video Object Segmentation Daizong Liu<sup>1</sup>, Wei Hu<sup>1</sup> <sup>1</sup>Wangxuan Institute of Computer Technology, Peking University



METHOD





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

# Visualization



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

# Quantitative comparison

	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
ć	<i>T</i> &.	FMean↑	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
	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
J		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
0	$\mathcal{T}$	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

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

Method	Backhone	$\mathcal{J}\&\mathcal{F}$		J	$\mathcal{F}$		
Method	Backbolle	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	

# Table 3: Ouantitative 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 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

Ablation study

Tabl	e
Network	Va

Baseline (. \_\_\_\_\_

Baseline + 0 Baseline + C 

Baseline + I Baseline + F Graph layer Graph layer

Input frames Input frames Input frames Input frames Global graph Global graph Global graph



### **Results**

## 4: Overall ablation studies.

				~.
riant	Mean $\mathcal{J}$ $\uparrow$	$\bigtriangleup \mathcal{J}$	Mean $\mathcal{F}\uparrow$	$ riangle \mathcal{F}$
GNN)	80.7	-4.5	79.1	-7.0
N	lode design			
CCP	81.9	-3.3	82.0	-4.1
CFCP (FL)	83.0	-2.2	83.5	-2.6
Local	graph netwo	ork		
L + BE	83.6	-1.6	84.3	-1.8
L + BE&BA	85.2	-	86.1	-
= 1	85.2	-	86.1	-
= 2	84.5	-0.7	85.7	-0.4
Oth	ner Variations	S		
T'=3	83.2	- 2.0	83.7	-2.4
T'=5	84.3	-0.9	85.0	-1.1
T' = 7	85.2	-	86.1	-
T'=9	85.2	-	86.1	-
K = 1	83.9	-1.3	84.4	-1.7
K = 3	85.2	-	86.1	-
n $K = 5$	84.7	-0.5	85.8	-0.3