Unsupervised Video Object Segmentation using Motion Saliency-Guided Spatio-Temporal Propagation

Yuan-Ting Hu[1] Jia-Bin Huang[2] Alexander G. Schwing[1]

[1] University of Illinois Urbana-Champaign [2] Virginia Tech

1. Introduction

Problem

- Segmenting the foreground objects in a video sequence
- No manual annotation is available (unsupervised)

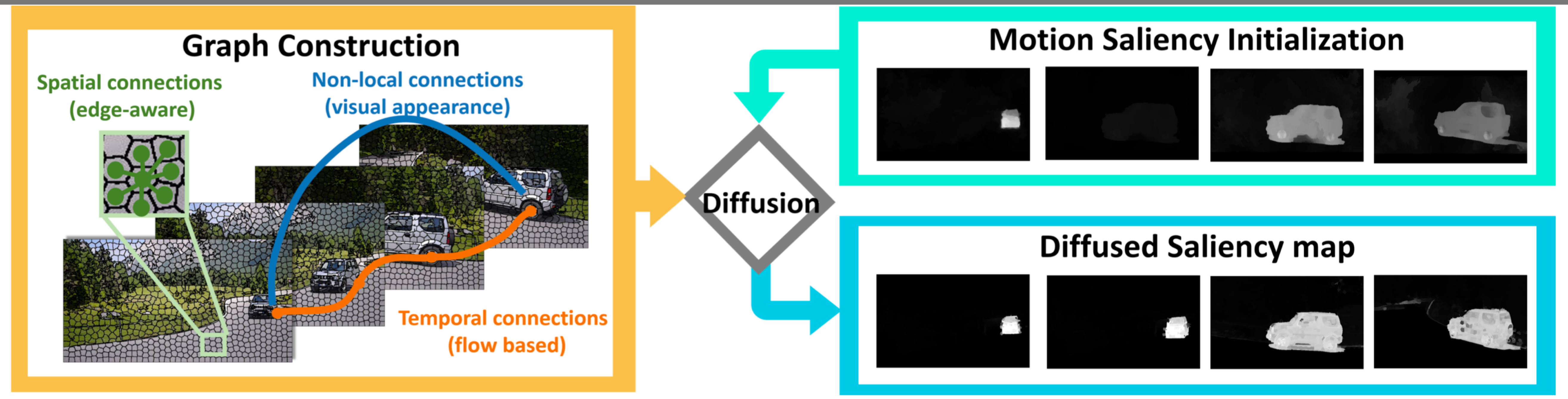
Challenges

- occlusion, deformation, dynamic background

Contributions

- A novel graph construction method (Sec. 3)
- A novel saliency estimation technique (Sec. 4)
- State-of-the-art performance (outperforming deep learning based methods) in the unsupervised setting (Sec. 5)

2. Overview



Problem definition

- Input: a video sequence $\{x_1, x_2, \ldots, x_F\}$
- Goal: predict FG/BG segmentation $\{y_1, y_2, \ldots, y_F\}$

A diffusion-based approach

- Extract superpixels for each frame independently
- Diffuse the saliency estimate to denoise the rough initial prediction

$$v_t \leftarrow G v_{t-1}$$

v_t: the FG estimate at time t

G: the adjacency matrix of the spatio-temporal graph

3. Graph Construction

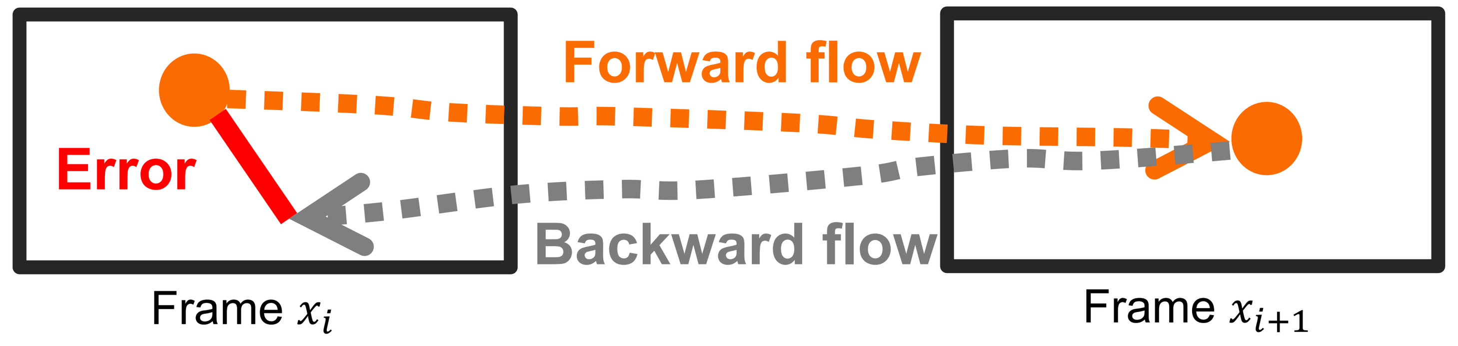
Spatio-temporal graph $G = (V, E)$

Long range non-local connections

- Search k nearest neighbors within adjacent f frames
- Weights = Visual similarity between superpixels
- Features: HOG + color histogram + (x,y) position

Intra-frame flow-based temporal connections

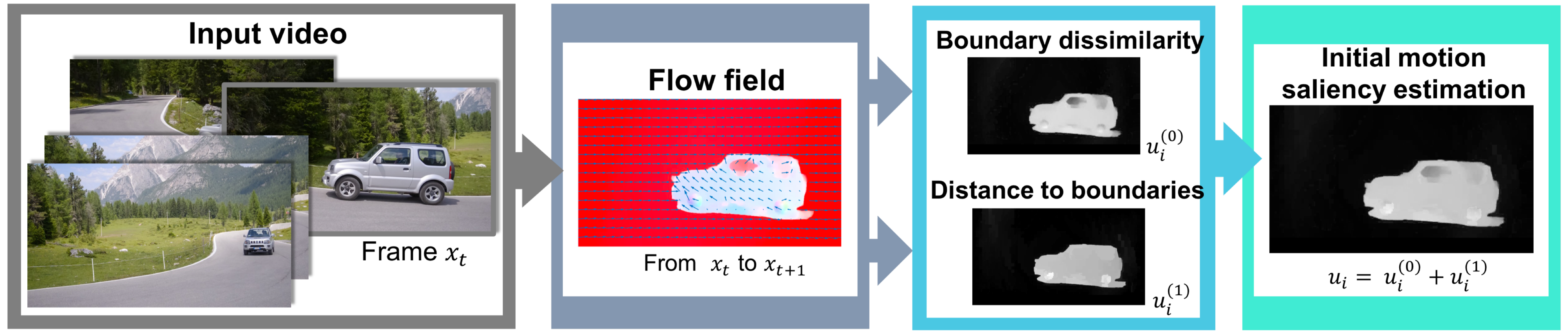
- Superpixels with consistent flow vectors



Inter-frame edge-aware spatial connections

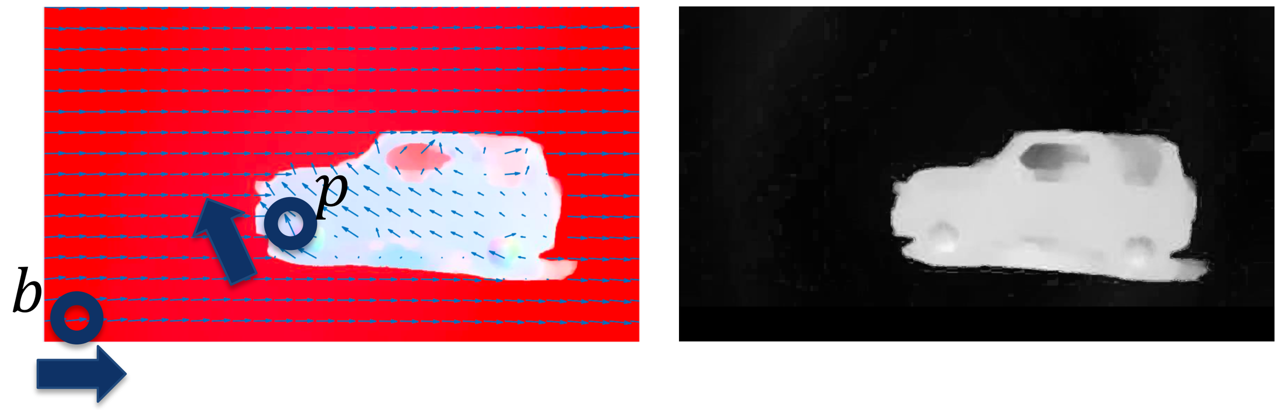
- Connect neighboring superpixels and avoid crossing strong edges

4. Motion Saliency Estimation



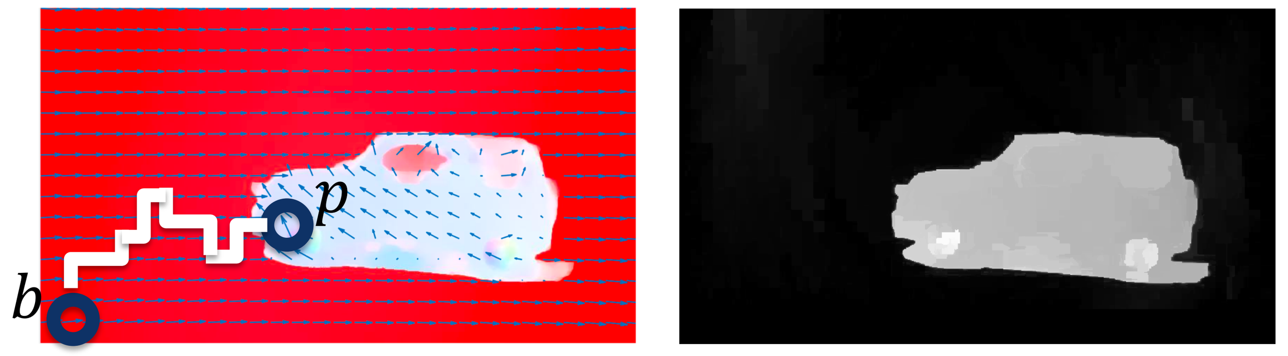
Boundary Dissimilarity

- Compute the **flow difference** between a pixel p and the boundary pixels



Distance to Boundaries

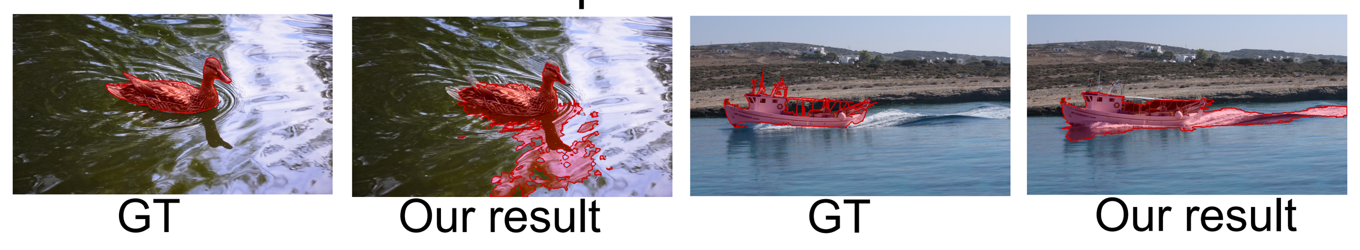
- Compute the smallest **barrier distance** between a pixel p to the boundary pixels



5. Experimental Results

Qualitative Results of Our Method

- Failure Cases: complex motion



Quantitative Results

- DAVIS dataset

		NLC	MSG	KEY	FST	FSG	LMP	ARP	**OURS-U**
Deep features		-	-	-	-	✓	✓	-	-
$\mathcal{J}$	Mean $\mathcal{M}$ ↑	0.641	0.543	0.569	0.575	0.716	0.697	0.763	**0.776**
	Recall $\mathcal{O}$ ↑	0.731	0.636	0.671	0.652	0.877	0.829	**0.892**	0.886
	Decay $\mathcal{D}$ ↓	0.086	0.028	0.075	0.044	**0.017**	0.056	0.036	0.044
$\mathcal{F}$	Mean $\mathcal{M}$ ↑	0.593	0.525	0.503	0.536	0.658	0.663	0.711	**0.750**
	Recall $\mathcal{O}$ ↑	0.658	0.613	0.534	0.579	0.790	0.783	0.828	**0.869**
	Decay $\mathcal{D}$ ↓	0.086	0.057	0.079	0.065	0.043	0.067	0.073	**0.042**
$\mathcal{T}$	Mean $\mathcal{M}$ ↓	0.356	0.250	**0.190**	0.276	0.286	0.689	0.352	0.243

- Segtrack v2 dataset

Sequence	KEY	FST	NLC	FSG	Ours
Average IoU	0.573	0.527	0.672	0.614	**0.701**

- FBMS dataset

	NLC	POR	POS	FST	ARP	**OURS**
Average IoU	0.445	0.473	0.542	0.555	0.598	**0.608**

Initialization Quality

- Intersection over union (IoU) of the initialization on the DAVIS

DAVIS	NLC	FST	FSG	LMP	Ours
Training?	-	-	✓	✓	-
Initial saliency	0.402	0.456	**0.602**	0.569	0.575

Ablation Study

Connections: Inter-frame	Connections: Intra-frame	Connections: Long range	FDiff	IoU (%)
-	-	-	-	57.52
✓	-	-	-	62.75
-	✓	-	-	62.13
-	-	✓	-	72.38
✓	✓	-	-	65.01
✓	-	✓	-	72.70
-	✓	✓	-	74.13
✓	✓	✓	-	74.34
✓	✓	✓	✓	**77.56**