Graph-to-Graph Energy Minimization for Video Object Segmentation

Yuezun Li[1*], Longyin Wen[2*], Ming-Ching Chang[1] and Siwei Lyu[1]

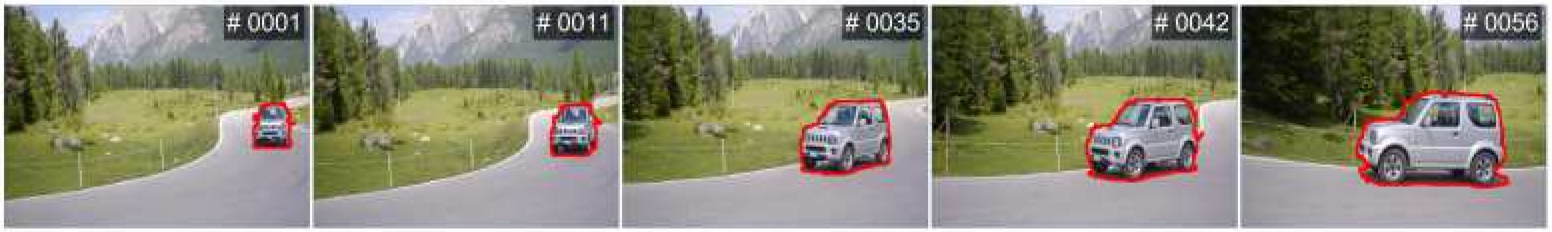
[1]University at Albany, State University of New York, USA

[2]JD Finance AI Lab, USA

UNIVERSITY AT ALBANY
State University of New York

FROM SEEING
TO UNDERSTANDING

Video Object Segmentation



- Semi-supervised video object segmentation

  Given the initial mask in the first frame.

- Unsupervised video object segmentation

  Given NO initial information

Unsupervised Video Object Segmentation

Overview

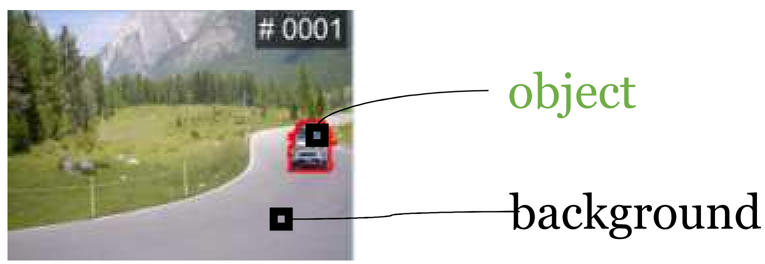
- Problem Formulation
- Optimization
- Experiments

Problem Formulation

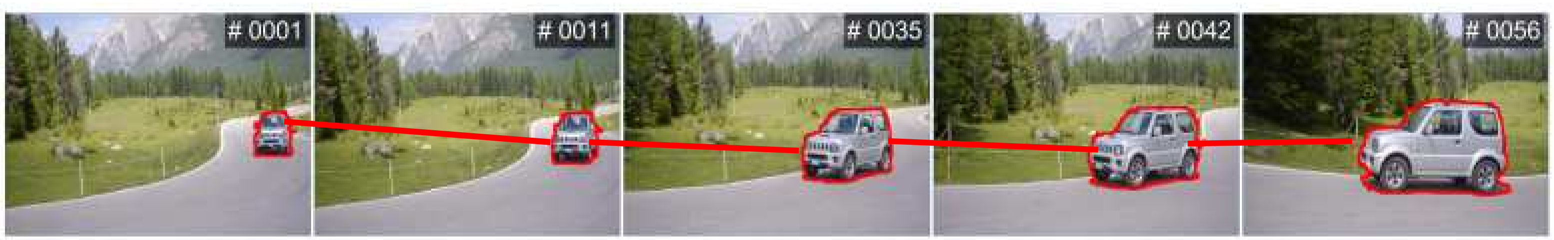
Given the video sequence, we aim to generate several space-time object tubes (regions across time).

We consider:

- In the **spatial domain**, each pixel / superpixel will be assigned label that which object it belongs to in the image.



- In the **temporal domain**, the regions shall be linked to construct space-time tubes.



Motivation

Previous works typically treat this problem as two separate steps (*labeling* and *linking*).

In contrast, we integrate these two steps into an unified framework to **mutually benefit each other**.

Our joint optimization framework consists of two terms:

- **Region-Energy** ($E_{\mathbf{R}}$) term for regions selection across the temporal domain
- **Superpixel-Energy** (E_{Φ}) term for regions refinement in the spatial domain

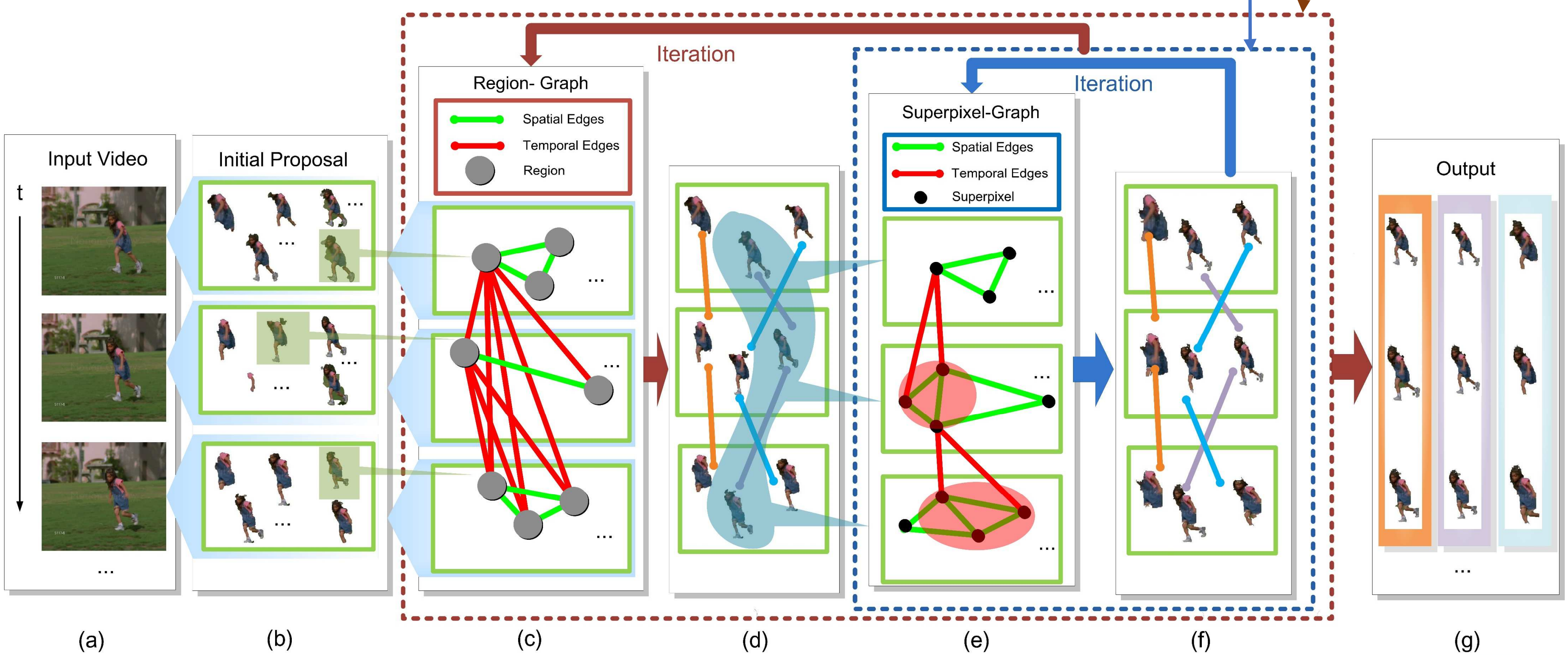
The overall label assignment problem is the optimization:

$$\min \{ E_{\mathbf{R}} + E_{\Phi} \}$$

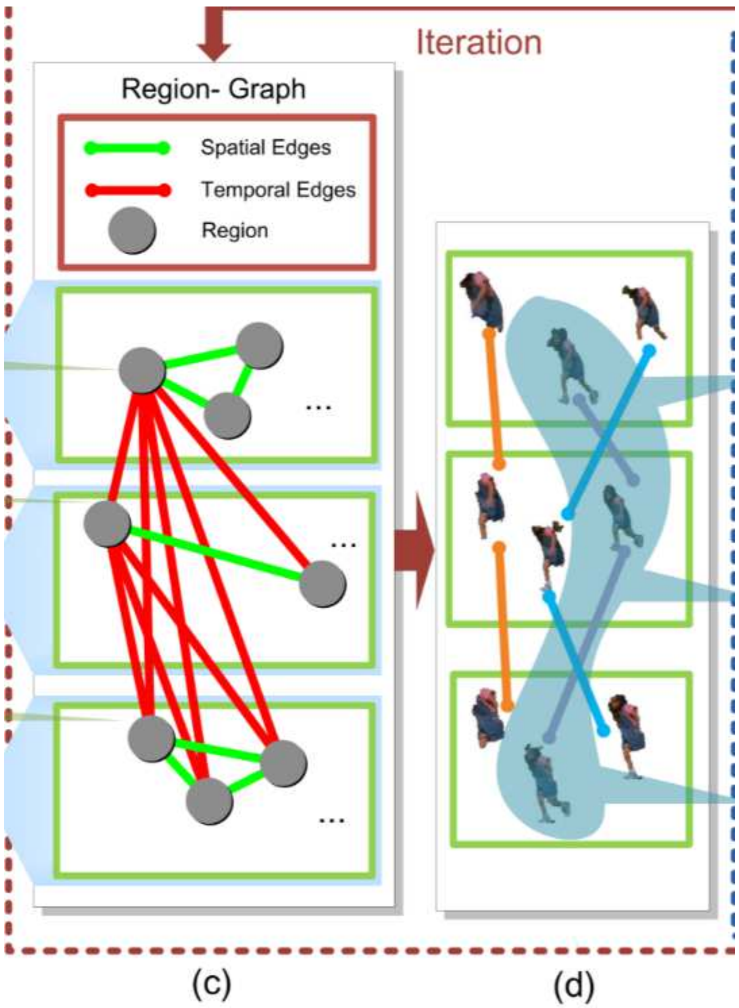
Method Overview

Each energy term is implemented by constructing different **Graph** model and convert minimization into Graph Clustering problem:

- **Region-Graph** for Region-Energy
- **Superpixel-Graph** for Superpixel-Energy



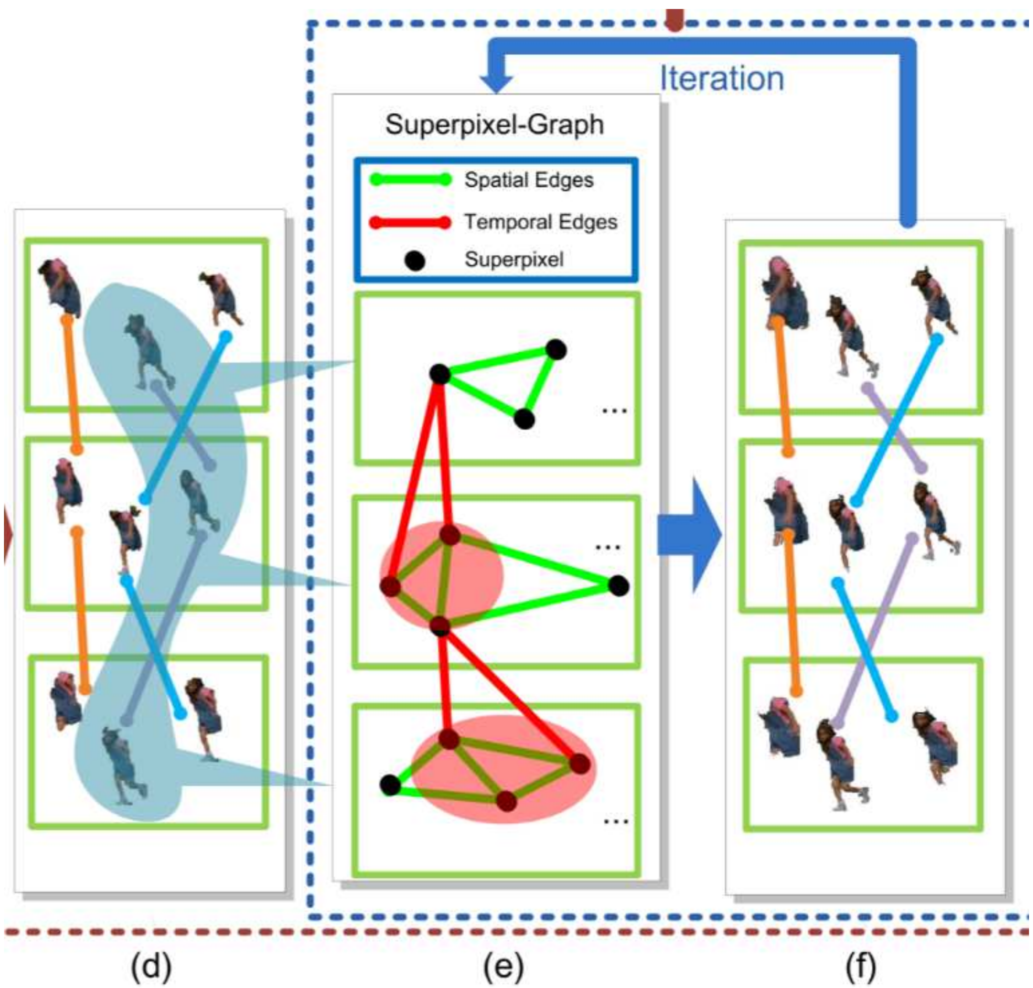
Region-Energy (Spatial-Temporal)



This energy is expected to be small if two regions in consecutive frames have similar features, including:

- *LAB color histogram*
- *Average Convolutional Neural Network (CNN) features*
- *Interaction-over-Union(IoU) between two regions*
- *Size similarity*

Superpixel-Energy (Spatial)



This energy is expected to be small if the **data cost** and **smoothness cost** of this graph are small

Data cost:

How well a superpixel can fit in its corresponding region

Color Guassian Mixture Model (GMM), CNN features, location cue.

Smoothness cost:

Two adjacent superpixels in a same region should have similar *RGB color and CNN features*

Y. Boykov etal. Fast approximate energy minimization via graph cuts. TPAMI, 2001

Unsupervised Video Object Segmentation

Overview

- Problem Formulation
- Optimization
- Experiments

Optimization

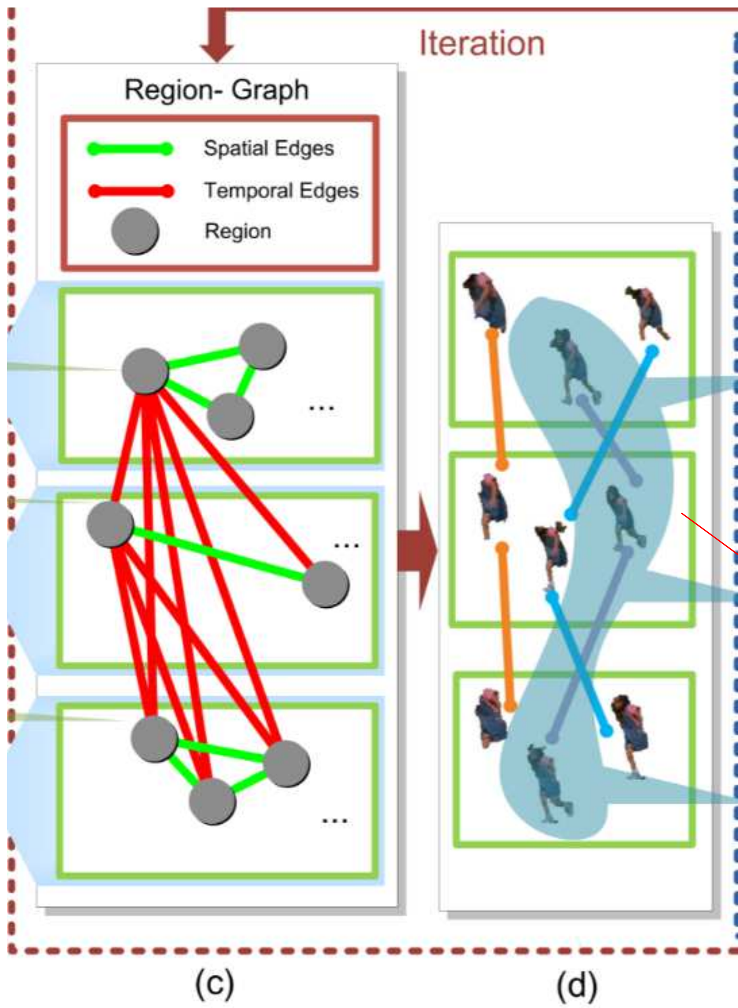
- We propose a **iterative algorithm** to minimize the energy.

iterate

1. Minimize the Region-Energy term to generate object tubes, which can be used as the initial guidance for Superpixel-Energy term.
2. Minimize the Superpixel-Energy term to refine the shape of regions in spatial and temporal domain.
3. If the overall energy is reduced, we employ the current results. Otherwise we keep the current results intact.

The iterative process will be terminated until convergence, *i.e.*, the overall energy is no longer reduced.

Minimizing S-T Region-Energy Term



We employ the greedy scheme used in [1] to cluster regions into tubes across the whole video.

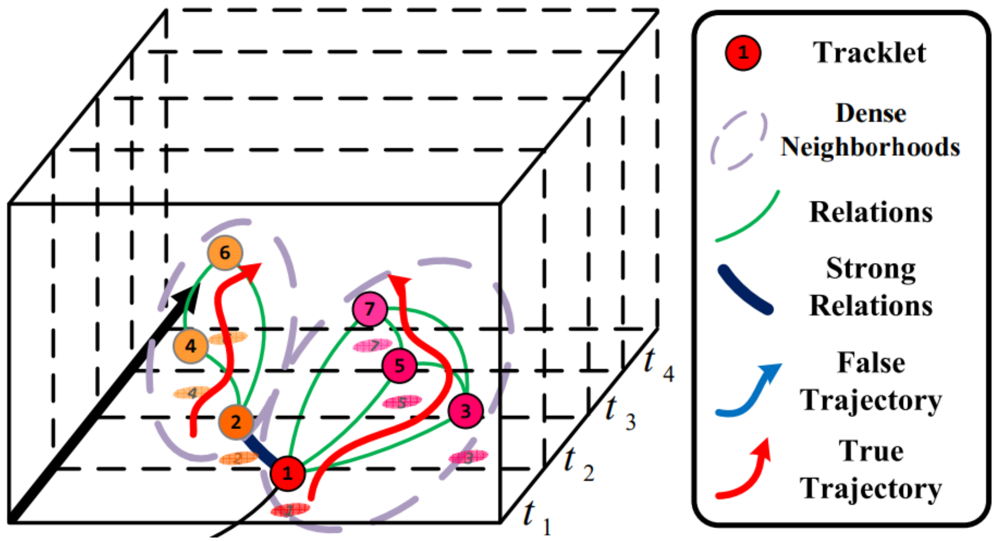
[1] L. Wen *et al*. Multiple target tracking based on undirected hierarchical relation hypergraph, CVPR 2014.

The shapes of regions does not change within this stage.

Only the selection of region across time is iteratively refined.

Undirected Hierarchical Relation Hypergraph

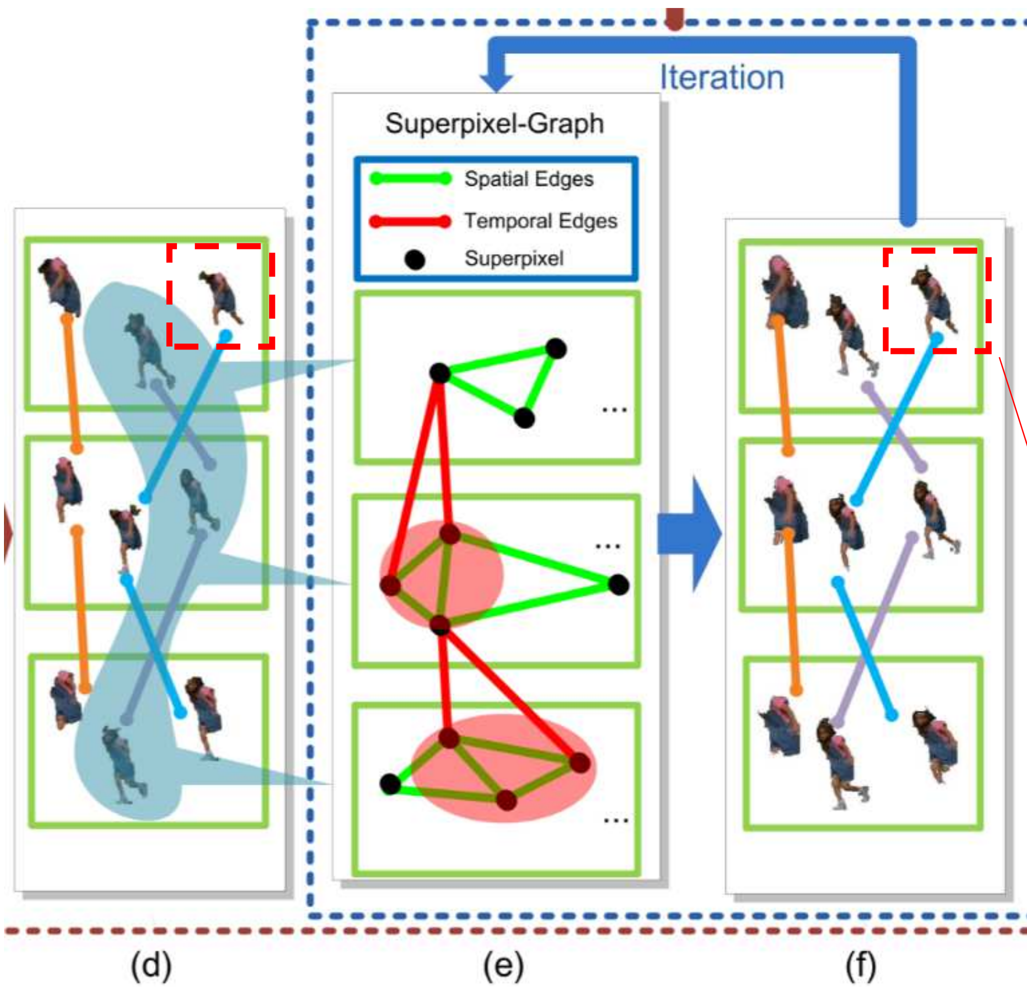
L. Wen *et al*. Multiple target tracking based on undirected hierarchical relation hypergraph. In CVPR, 2014



Tracking is cast as a **hierarchical dense neighborhoods searching problem** on a dynamically constructed **undirected affinity graph**.

Consider high-order spatial-temporal relationship between nodes in the hypergraph. This helps to track nearby similar targets that are spatially close.

Minimizing Superpixel-Energy Term



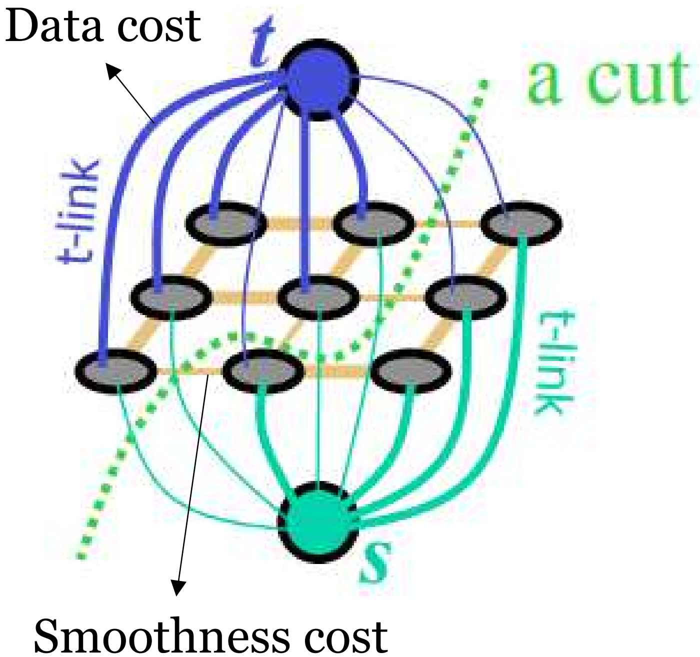
We employ the **alpha expansion** based **graph-cut** algorithm [2] for solving the minimization.

[2] Y. Boykov etal. Fast approximate energy minimization via graph cuts. PAMI, 2001.

The selection of regions across time does not change within this stage.

Only the shape of region is iteratively refined.

Energy Minimization via Graph Cuts



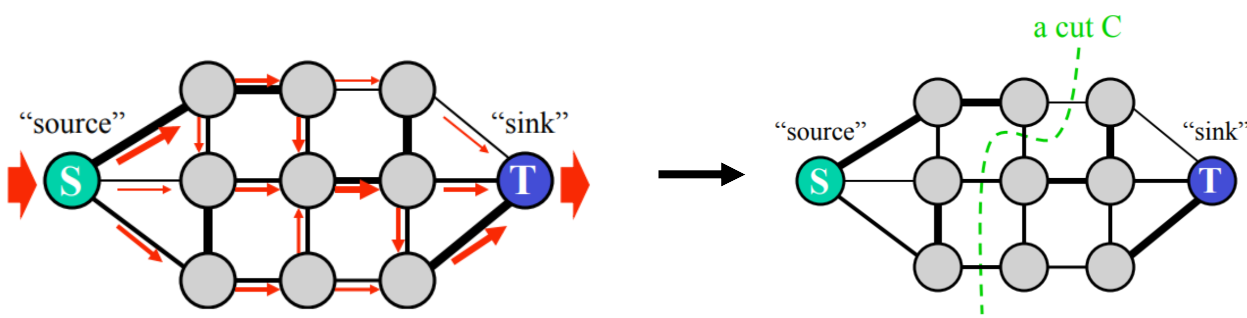
Max-flow/min-cut problem

Image pixels correspond to graph nodes.

- Nearby pixels (nodes) connected by an edge, the **n-link**
- Terminal **s** (with label *0*) connects to every image pixel via a **t –link.**
- Terminal **t** (with label *1*) connects to every image pixel via a **t-link**.

A cut separates **t** from **s:** Each pixel stays connected to either **t** or **s** (label 1 or 0).

http://www.cs.jhu.edu/~hager/teaching/cs461/Notes/2008/GraphCuts.pdf



Each graph-cut step generates the segmentation of one object class.

Unsupervised Video Object Segmentation

Overview

- Problem Formulation
- Optimization
- Experiments

UNIVERSITY AT ALBANY State University of New York

FROM SEEING TO UNDERSTANDING

Experiments

Experiments on SegTrack v2 dataset* with IoU metric

Category	Semi-Supervised				Unsupervised				
Methods	[37]	[16]	[4]	[34]	[20]	[18]	[12]	[39]	GEM
Mean per Object	71.8	67.4	35.6	74.1	65.9	45.3	51.8	69.1	**71.3**
Mean per Sequence	72.2	68.8	40.4	75.3	71.2	57.3	50.8	73.9	**75.0**
Avg.# of Proposals	N/A	N/A	N/A	N/A	60.0	10.6	336.6	121.9	339.0

* F. Li, *et al.* Video segmentation by tracking many figure-ground segments. ICCV 2013.

Experiments on DAVIS dataset** with IoU metric

Measure		Semi-Supervised					Unsupervised								
		MSK [16]	JMP [8]	FCP [26]	BVS [22]	OFL [34]	NLC [7]	CVOS [30]	TRC [11]	KEY [18]	SAL [36]	FST [24]	CUT [15]	LMP [31]	GEM
$\mathcal{J}$	Mean ↑	0.797	0.607	0.631	0.665	0.711	0.641	0.514	0.501	0.569	0.426	0.575	0.552	**0.697**	0.696
	Recall ↑	0.931	0.693	0.778	0.764	0.800	0.731	0.581	0.560	0.671	0.386	0.652	0.575	**0.892**	0.867
	Decay ↓	0.089	0.372	0.031	0.260	0.227	0.086	0.127	0.050	0.075	0.084	0.044	**0.022**	0.056	0.058
$\mathcal{F}$	Mean ↑	0.754	0.586	0.546	0.656	0.679	0.593	0.490	0.478	0.503	0.383	0.536	0.552	**0.663**	0.596
	Recall ↑	0.871	0.656	0.604	0.774	0.780	0.658	0.578	0.519	0.534	0.264	0.579	0.610	**0.783**	0.662
	Decay ↓	0.090	0.373	0.039	0.236	0.240	0.086	0.138	0.066	0.079	0.072	0.065	**0.034**	0.067	0.077
$\mathcal{T}$	Mean ↓	0.218	0.131	0.285	0.316	0.221	0.356	0.243	0.327	**0.190**	0.600	0.276	0.277	0.686	0.246

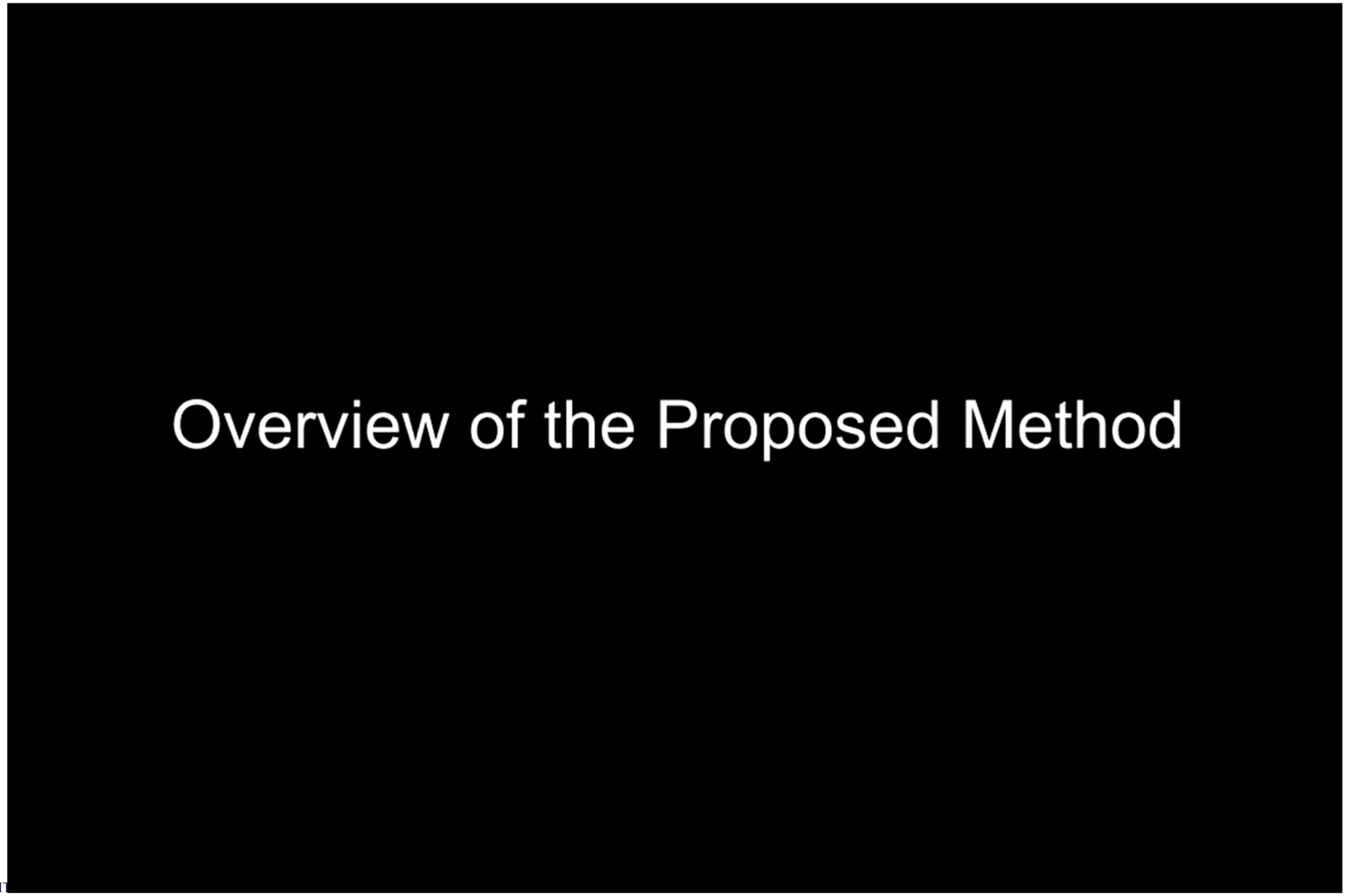
** F. Perazzi, *et al.* A benchmark dataset and evaluation methodology for video object segmentation, CVPR 2016.

Results – Visual Snapshots

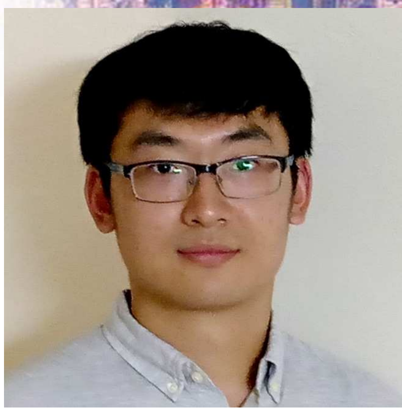
UNIVERSITY AT ALBANY
State University of New York

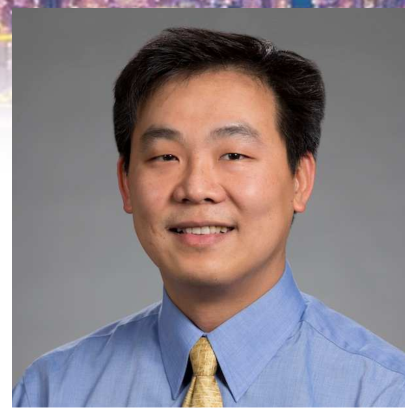
FROM SEEING TO UNDERSTANDING

Results – Video Demo



Thank You





Yuezun Li
SUNY Albany

Longyin Wen
JD Research

Ming-Ching Chang
SUNY Albany

Siwei Lyu
SUNY Albany

UNIVERSITY AT ALBANY
State University of New York

FROM SEEING TO UNDERSTANDING