LEARNING TO SEGMENT MOVING OBJECTS IN VIDEOS – FRAGKIADAKI ET AL. 2015

Darshan Thaker

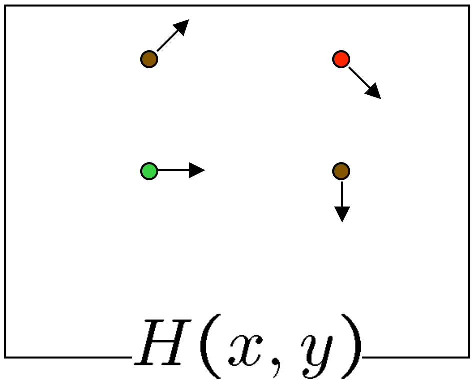
Oct 4, 2017

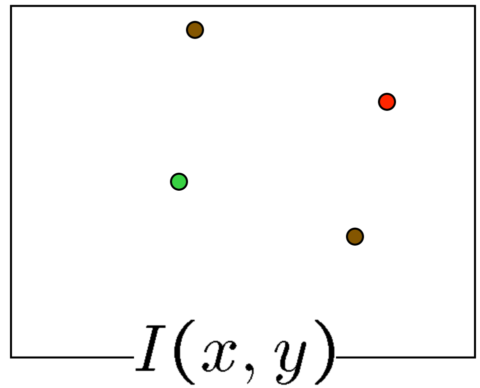
Problem Statement

- Moving object segmentation in videos
  - Applications: security tracking, pedestrian detection, etc.

GIF credit: https://giphy.com/search/football-is-back

Brief background on optical flow





- **Optical flow** problem: estimate pixel motion from image H to image I?
- Use *large displacement optical flow* approach [1]
  - Output can be interpreted as three channel image
- **Flow bleeding**: Optical flow misaligns with true object boundaries

[1]: T. Brox and J. Malik. Large displacement optical flow

Slide credit: Steve Seitz

Overview of Approach

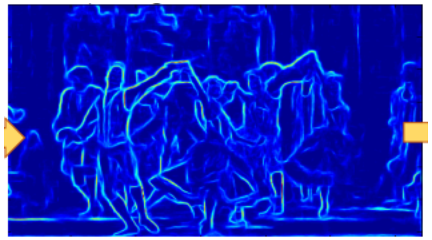
- Moving Object Proposals (MOPs)
- Moving Objectness Detector on optical flow + RGB channels
- Obtain dense point trajectories
  - Intersection of trajectories with MOPs yields foreground and background segmentation
- Propagate pixel labels to nearby frames using random walks
- Generate proposals by clustering superpixels across frames

Approach: Step 1

Ground Truth

Video Frame

Image boundaries



Note: this uses structured forest boundary detector

Image credit: Fragkiadaki et. al

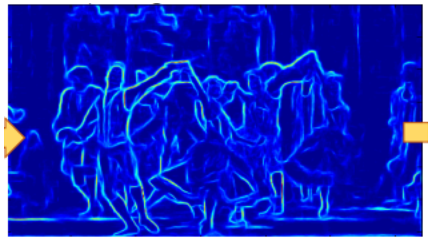
Approach: Step 1

Ground Truth

Video Frame

Note: this uses structured forest boundary detector

Image boundaries



Static Object Proposals

Image credit: Fragkiadaki et. al

Approach: Step 1

Ground Truth

Optical flow

Video Frame

Note: this uses structured forest boundary detector

Image boundaries

Static Object Proposals

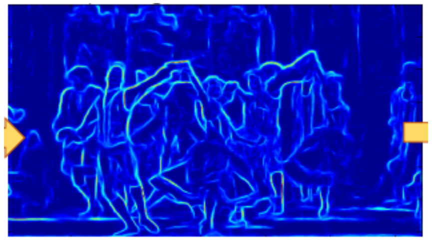


Image credit: Fragkiadaki et. al

Approach: Step 1

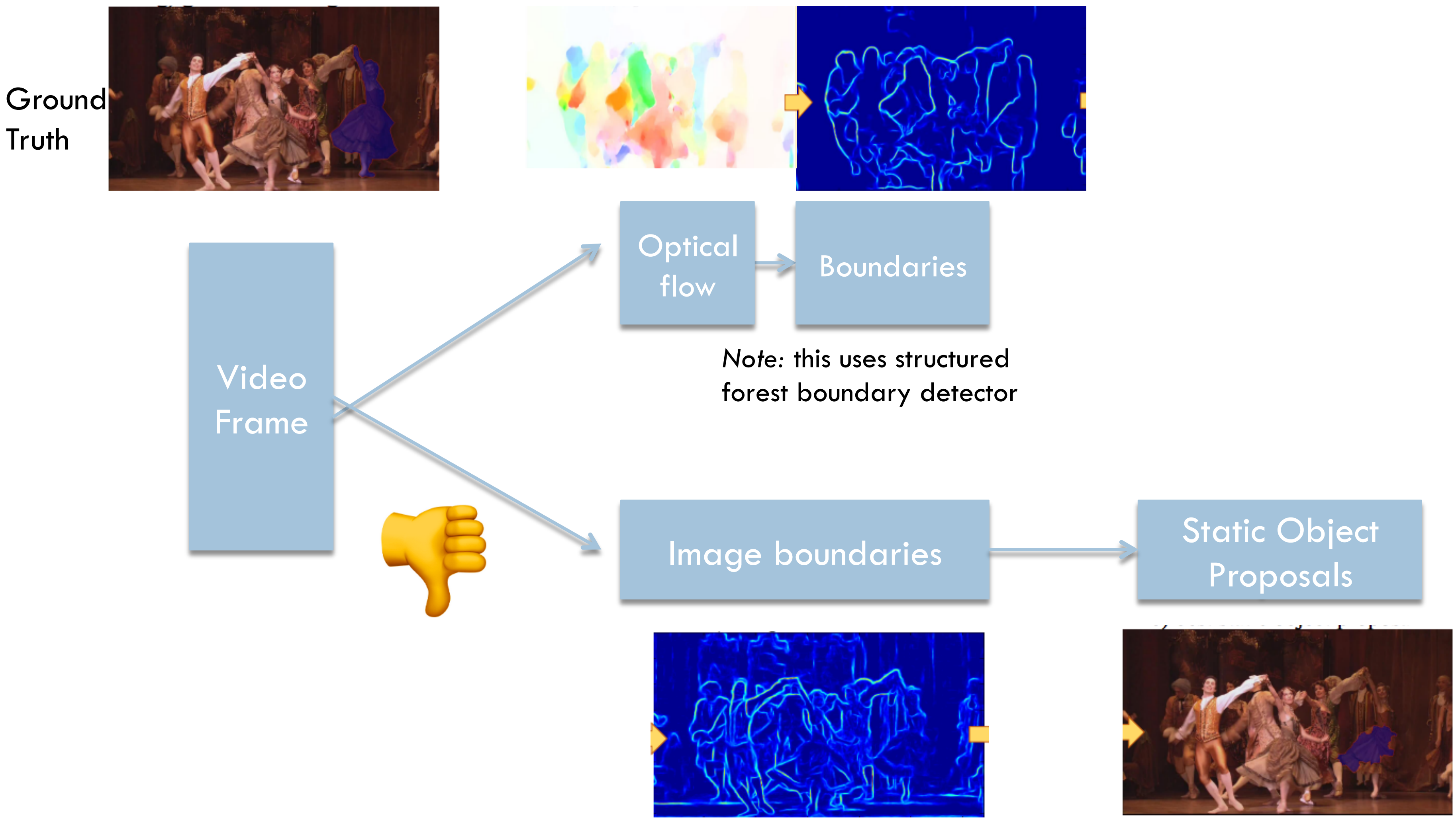


Image credit: Fragkiadaki et. al

Approach: Step 1

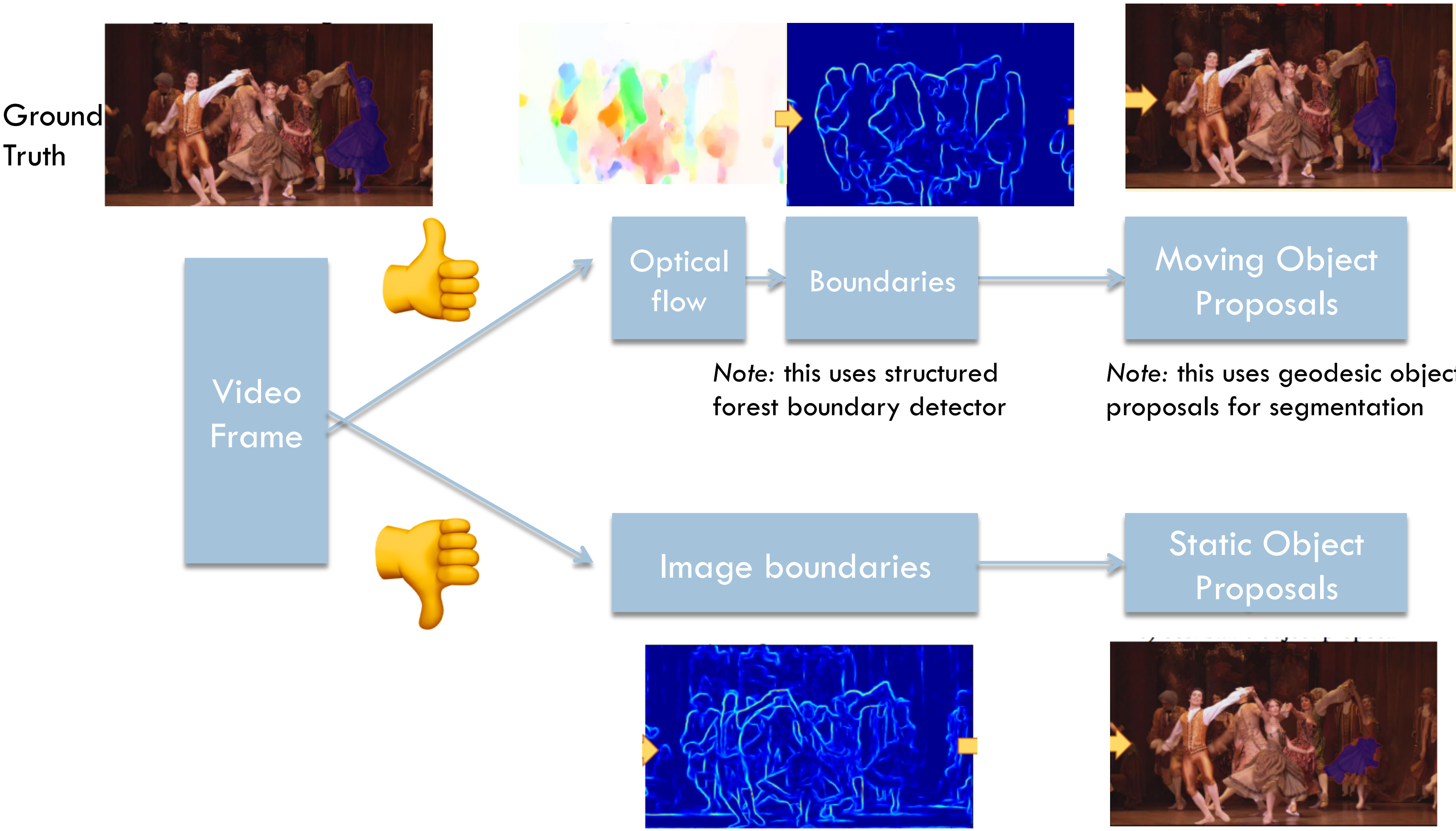
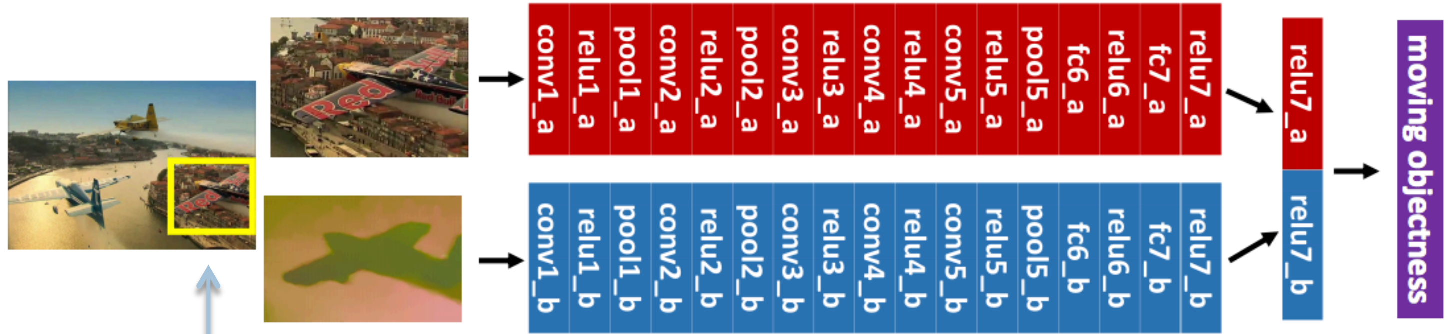


Image credit: Fragkiadaki et. al

Approach: Step 2a

Moving Objectness Detector
with dual pathway architecture
on optical flow + RGB channels

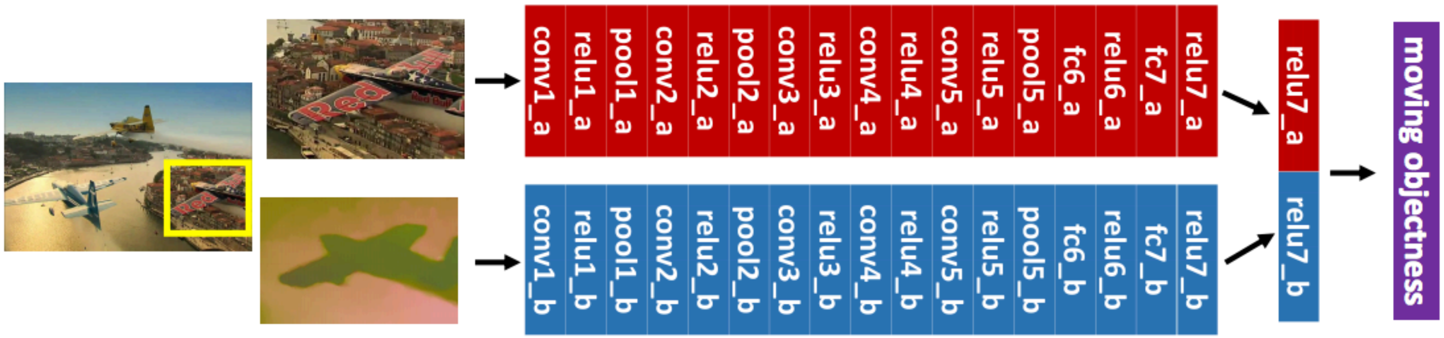
Outputs
score in
[0, 1]



Moving Object Proposal

Image credit: Fragkiadaki et. al

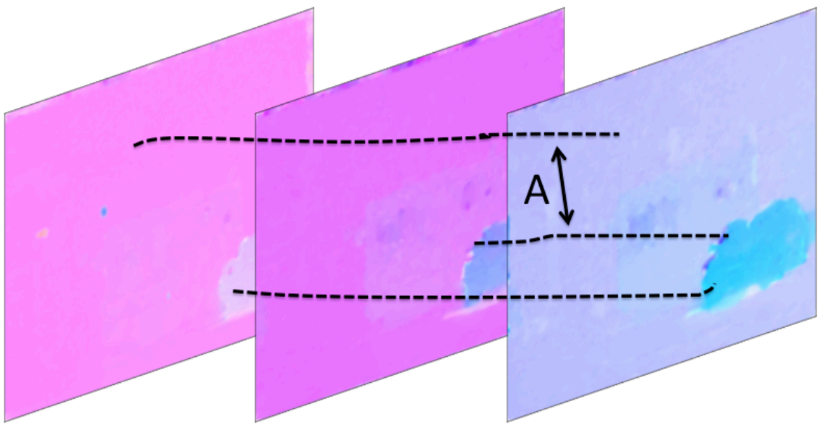
Approach: Step 2b



- Weights in each network stack initialized to pretrained Imagenet 200 category network (R-CNN)
- Finetuned with small collection of moving object boxes + background boxes from VSB100 and Moseg video datasets

Image credit: Fragkiadaki et. al

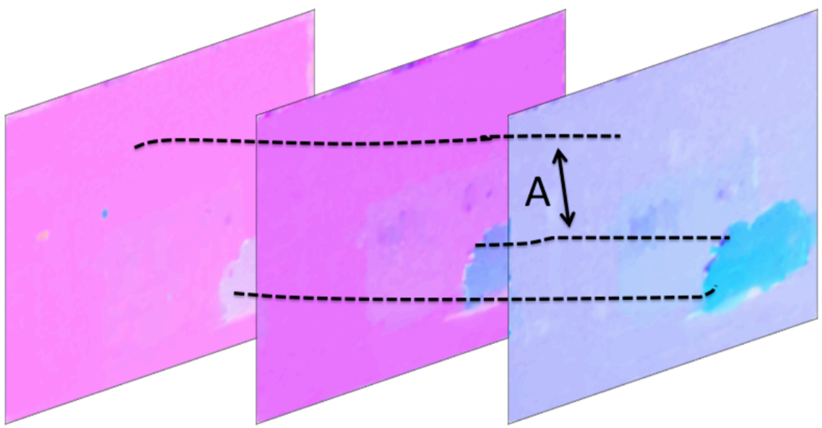
Approach: Step 3



Obtain dense point trajectories
by linking optical flow fields.

Image credit: Fragkiadaki et. Al (https://www.cs.cmu.edu/~katef/videoseg.html)

Approach: Step 3



Obtain dense point trajectories
by linking optical flow fields.

N = # trajectories

$$
\begin{array}{cc}
N & \begin{bmatrix} 0.5 & 1 & 0.25 & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} \\
 & N
\end{array}
$$

Compute pairwise trajectory affinity matrix **A** (affinity = fn of maximum velocity difference)

Image credit: Fragkiadaki et. Al (https://www.cs.cmu.edu/~katef/videoseg.html)

Approach: Step 4a

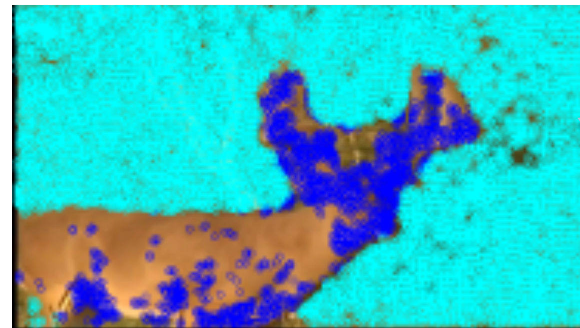
Moving Object Proposal

Image credit: Fragkiadaki et. al

Approach: Step 4a

Moving Object Proposal

Trajectories intersection with MOP



background

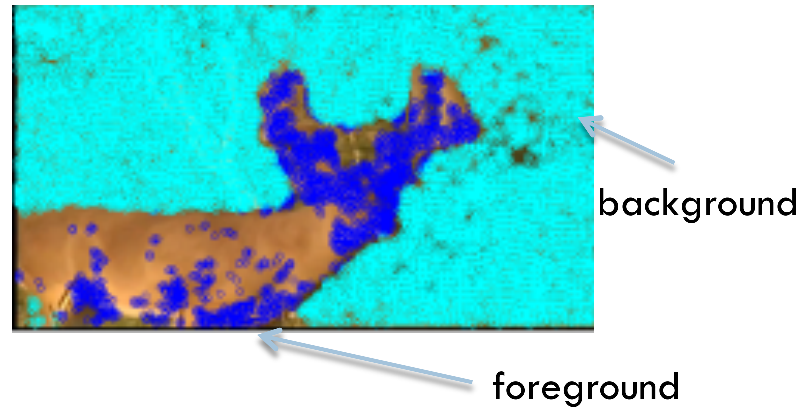
foreground

Image credit: Fragkiadaki et. al

Approach: Step 4a

Moving Object Proposal

Trajectories intersection with MOP



- Problem: Frames around F temporally might not have apparent motion (trajectories not overlap with MOP as shown below)

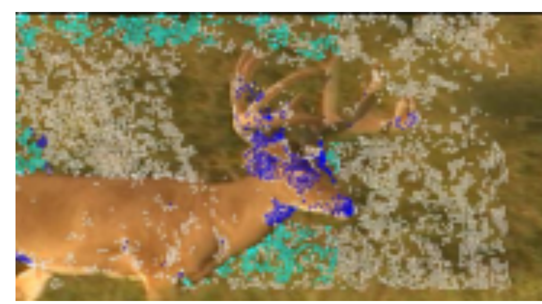


Image credit: Fragkiadaki et. al

Approach: Step 4b

- Propagate pixel labels through trajectory motion affinities using Random Walkers and minimizing cost function

$$\min_x . \quad \sum_{i,j}^{n} \mathbf{A}_{ij}(x_i - x_j)^2$$
$$\text{subject to} \quad x_B = 0, \quad x_F = 1$$

x denotes trajectory labels (fg or bg)

- Perform series of label diffusions (~50) to propagate trajectory labels and get better segmentations

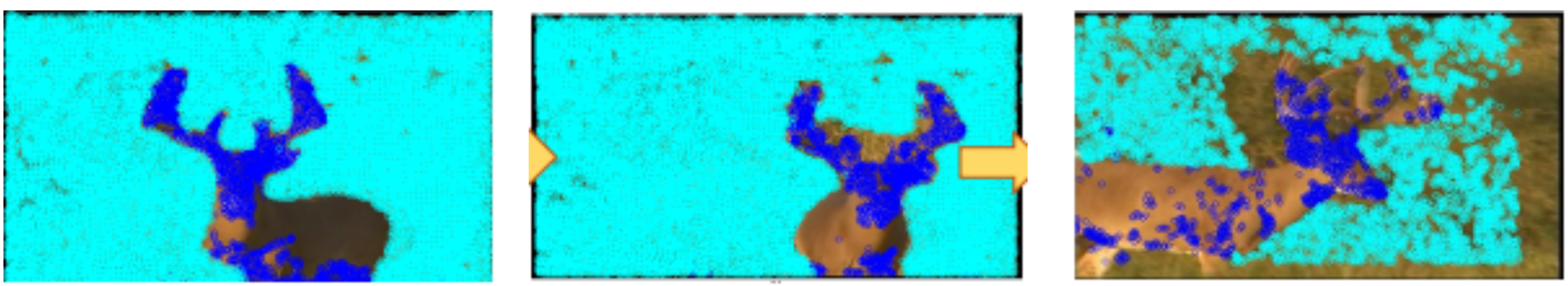


Image credit: Fragkiadaki et. al

Approach: Step 5

- Map trajectory clusters to pixels used weighted average over superpixels that extend across multiple frames
- Final goal: Maximize Intersection over Union (IOU) of spatio-temporal tubes with ground truth objects using fewest tube proposals

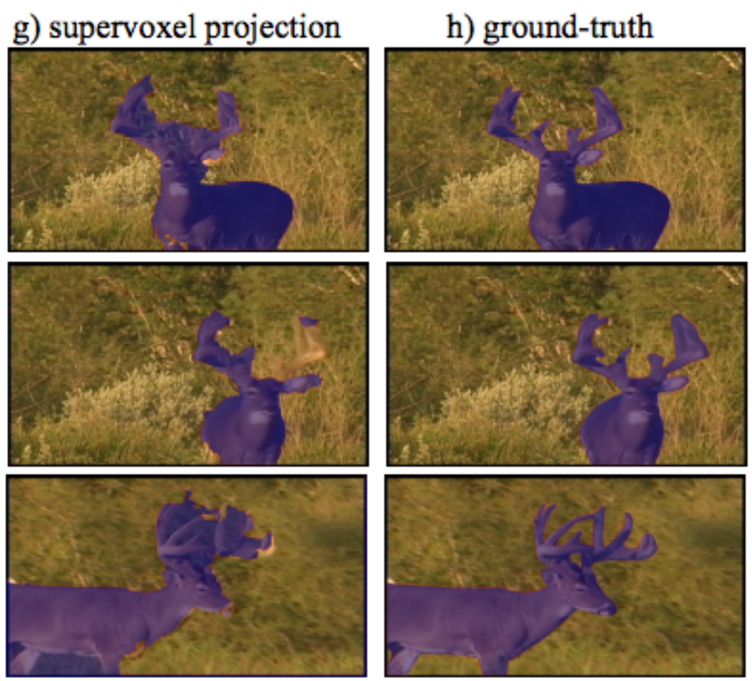


Image credit: Fragkiadaki et. al

Datasets

- VSB100
  - 100 HD human-annotated videos
  - Many crowded scenes (parade, cycling, etc.)
    - More challenging
- Moseg
  - 59 video sequences (720 frames) with pixel-accurate segmentation
  - Scenes from movie “Miss Marple” + cars and animals
  - Uncluttered scenes (one or two objects per video)

Experiments/Results

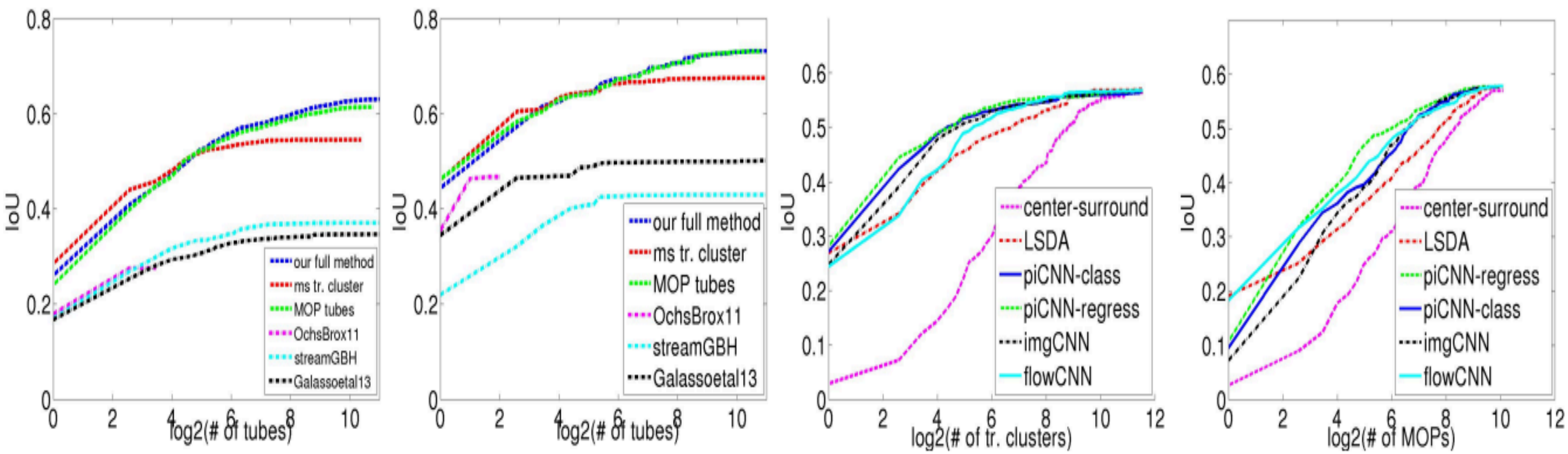


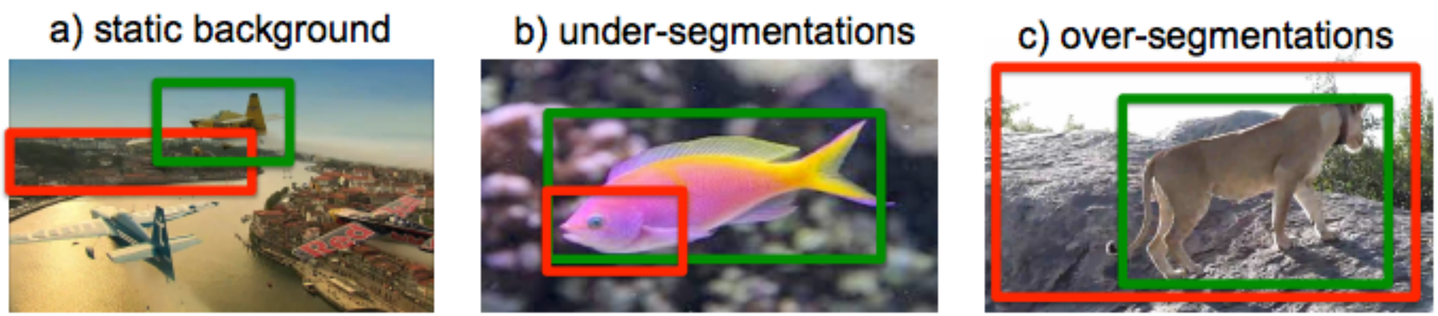
Image credit: Fragkiadaki et. al

Experiments/Results

Image credit: Fragkiadaki et. Al (https://www.cs.cmu.edu/~katef/videolearn.html)

Advantages

- Moving Objectness Detector learns to suppress these cases (in red)



- Not all frames will have moving objects because objects are not constantly in motion
  - Trajectory clustering propagates segmentation to frames with little motion
- Bridges gap between “bottom-up” motion segmentation and object-specific detectors

Image credit: Fragkiadaki et. Al (https://www.cs.cmu.edu/~katef/posters/CVPR2015_LearnVideoSegment.pdf)

Disadvantages/Extensions

- Same boundary detector used on both optical flow map and video frame
- Temporal Fragmentations caused by large motion or full object occlusions
- Inaccurate mapping of trajectory clusters to pixel tubes

Summary Points

- Video segmentation method with great looking results that are rarely undersegmented
- Opinion: Frame by frame MOP approach seems inherently flawed
  - Input to MOD could be n consecutive frames itself
- Trajectory clustering is noisy
  - Random walk depends on dataset and how long objects typically remain static