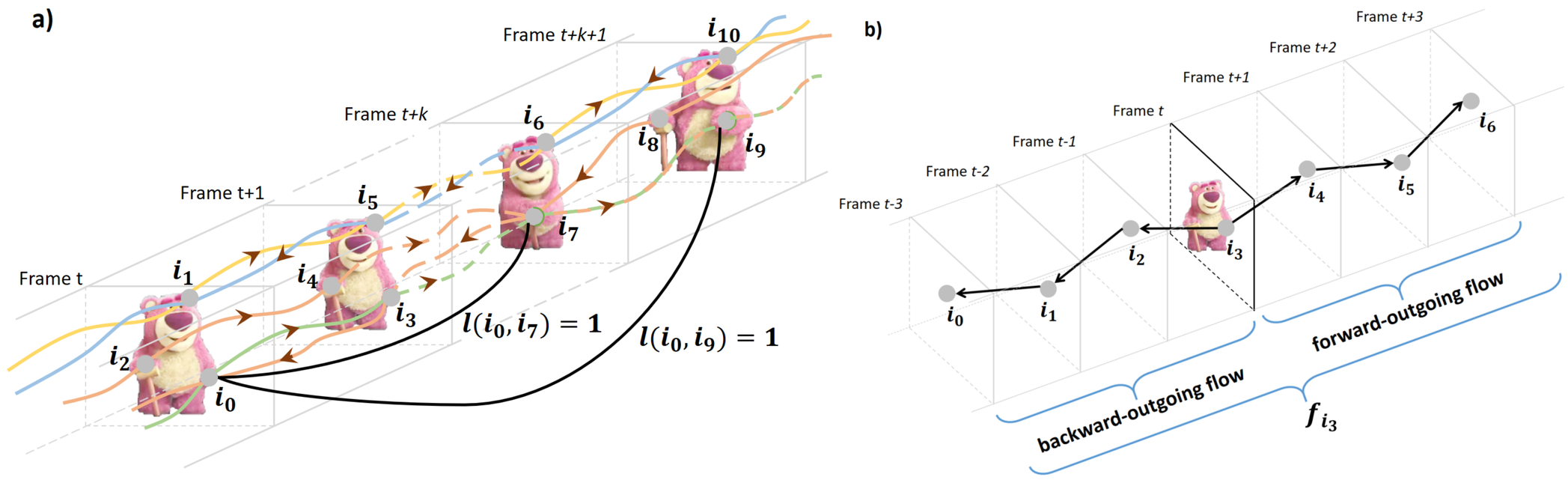
# Spacetime Graph Optimization for Video Object Segmentation

Emanuela Haller and Marius Leordeanu
ehaller@bitdefender.com, marius.leordeanu@imar.ro
Submitted to IEEE International Conference on Computer Vision (ICCV), 2019

## Spacetime graph




- **Motion flows**
- **Graph of pixels** $G = (V, E)$ nodes correspond to video pixels ($|V| = n = m \cdot h \cdot w$)
- **Adjacency matrix** $\mathbf{M} \in \mathbb{R}^{n \times n}$, $\mathbf{M}_{i,j} = l(i,j) \cdot k(i,j)$
- **Nodes features** $\mathbf{F} \in \mathbb{R}^{n \times d}$ ($\mathbf{f}_i$ collected along outgoing chains)
- **Nodes labels** $\mathbf{x} \in \mathbb{R}^{n \times 1}$ ($x_i \in [0,1]$ - (soft) segmentation labels)

## Motivation

- Move beyond frame by frame approaches
- Exploit spacetime data
  - Spacetime coherence as self-supervision signal
  - Alignments are usually non-accidental

## Problem Formulation

$$(\mathbf{x}^*, \mathbf{w}^*) = \text{argmax}_{\mathbf{x},\mathbf{w}}\, S(\mathbf{x}, \mathbf{w}) \quad \text{s.t.} \quad \|\mathbf{x}\|_2 = 1$$

$$S(\mathbf{x}, \mathbf{w}) = \mathbf{x}^T\mathbf{M}\mathbf{x} - \alpha(\mathbf{F}\mathbf{w} - \mathbf{x})^T(\mathbf{F}\mathbf{w} - \mathbf{x}) - \beta\mathbf{w}^T\mathbf{w}$$

- Maximize $S_C(\mathbf{x}) = \mathbf{x}^T\mathbf{M}\mathbf{x}$
  clustering score
- Minimize $\|\mathbf{F}\mathbf{w} - \mathbf{x}\|_2$
  enforce features consistency
- Under constraint $\|\mathbf{x}\|_2 = 1$

## Optimal Segmentation

- $\mathbf{x}$ - fixed point iteration scheme
- $\mathbf{w}$ - closed form solution

$$\begin{cases} \mathbf{x}^{(it+1)} = \frac{1}{p}(\mathbf{M}\mathbf{x}^{(it)} + 2\alpha\mathbf{F}\mathbf{w}^{(it)}) \\ \mathbf{w}^{(it+1)} = 2\alpha(\alpha\mathbf{F}^T\mathbf{F} - \beta\mathbf{I}_d)^{-1}\mathbf{F}^T\mathbf{x}^{(it+1)} \end{cases}$$

$$p = \|\mathbf{M}\mathbf{x}^{(it)} + 2\alpha\mathbf{F}\mathbf{w}^{(it)}\|_2$$

## Algorithm

**Propagation step**

$$\mathbf{x}^{(it+1)} \leftarrow \mathbf{M}\mathbf{x}^{(it)}$$

(implemented as label propagation)

**Regression step**

$$\mathbf{w}^{(it+1)} \leftarrow (\mathbf{F}^T\mathbf{F} - \beta\mathbf{I}_d)^{-1}\mathbf{F}^T\mathbf{x}^{(it+1)}$$

**Projection step**

$$\mathbf{x}^{(it+1)} \leftarrow \mathbf{F}\mathbf{w}^{(it+1)}$$

## Convergence

- Leading eigenvector of a specific matrix
- Solution is independent of the initialization

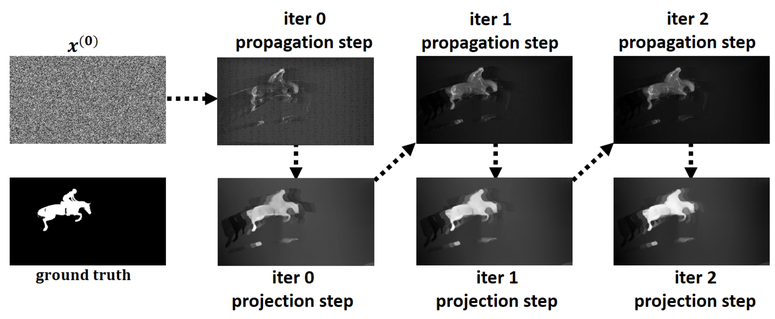
$$\mathbf{x}^{(it+1)} = \frac{\mathbf{A}\mathbf{x}^{(it)}}{\|\mathbf{A}\mathbf{x}^{(it)}\|_2}$$

$$\mathbf{A} = \mathbf{F}(\mathbf{F}^T\mathbf{F} - \beta\mathbf{I})^{-1}\mathbf{F}^T\mathbf{M} = \mathbf{P}\mathbf{M}$$

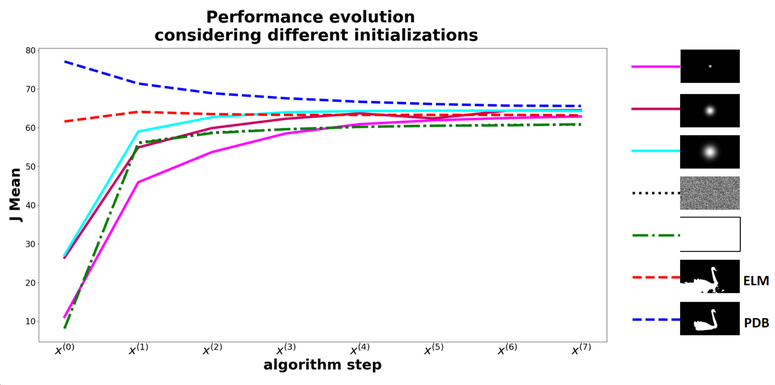
$\mathbf{P}$ - depends only on features

$\mathbf{M}$ - depends only on motion flows

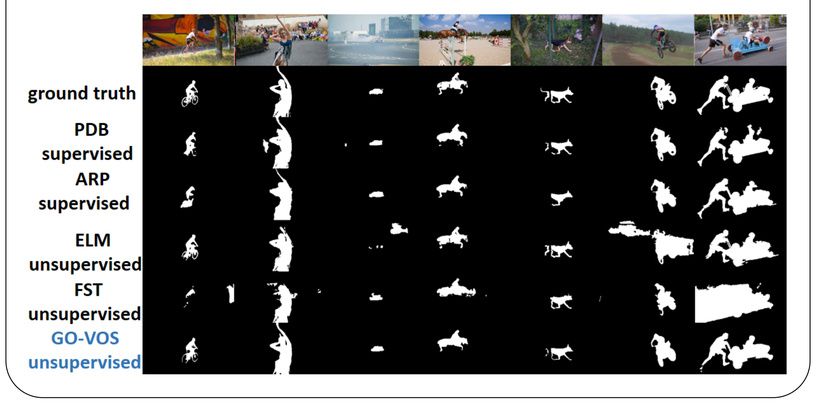
## Convergence in practice



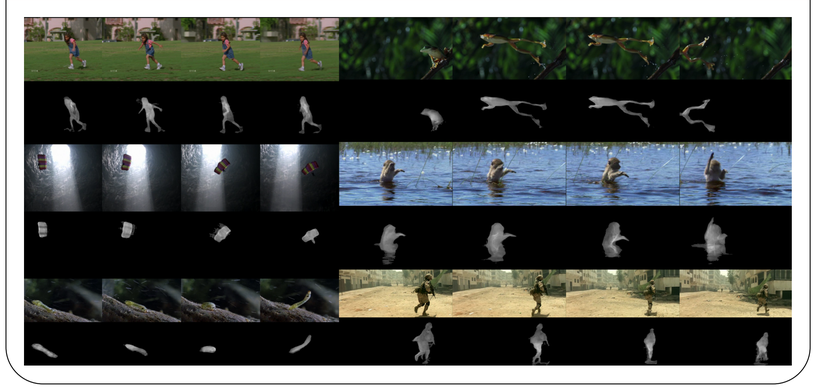

## Independence of initialization




## Qualitative Comparison - DAVIS




## Qualitative Results - SegTrack



## Quantitative Results

### DAVIS dataset

| Task | Method | | J Mean | F Mean | sec/frame |
|---|---|---|---|---|---|
| Unsupervised | Supervised features | PDB[13] | 77.2 | 74.5 | **0.05** |
| | | ARP[7] | 76.2 | 70.6 | N/A |
| | | LVO[14] | 75.9 | 72.1 | N/A |
| | | FSEG[4] | 70.7 | 65.3 | N/A |
| | | LMP[15] | 70.0 | 65.9 | N/A |
| | | **GO-VOS supervised** + features of [13] | **79.9** (+2.7) | **78.1** | 0.61 |
| | | **GO-VOS supervised** + features of [7] | 78.7 (+2.5) | 73.1 | 0.61 |
| | | **GO-VOS supervised** + features of [14] | 77.0 (+1.1) | 73.7 | 0.61 |
| | | **GO-VOS supervised** + features of [4] | 74.1 (+3.5) | 69.9 | 0.61 |
| | | **GO-VOS supervised** + features of [15] | 73.7 (+3.7) | 69.2 | 0.61 |
| | Unsupervised | ELM[8] | 61.8 | **61.2** | 20 |
| | | FST[11] | 55.8 | 51.1 | 4 |
| | | CUT[5] | 55.2 | 55.2 | ≈1.7 |
| | | NLC[2] | 55.1 | 52.3 | 12 |
| | | **GO-VOS unsupervised** | **65.0** | 61.1 | **0.91** |

### YouTube-Objects dataset v1.0

| Method | aero | bird | boat | car | cat | cow | dog | horse | moto | train | avg | sec/frame |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [12] | 51.7 | 17.5 | 34.4 | 34.7 | 22.3 | 17.9 | 13.5 | 26.7 | 41.2 | 25.0 | 28.5 | N/A |
| [11] | 65.4 | 67.3 | 38.9 | 65.2 | 46.3 | 40.2 | 65.3 | 48.4 | 39.0 | 25.0 | 50.1 | 4 |
| [16] | 75.8 | 60.8 | 43.7 | 71.1 | 46.5 | 54.6 | 55.5 | 54.9 | 42.4 | 35.8 | 54.1 | N/A |
| [6] | 64.3 | 63.2 | 73.3 | 68.9 | 44.4 | 62.5 | 71.4 | 52.3 | 78.6 | 23.1 | 60.2 | N/A |
| HPP[3] | 76.3 | 71.4 | 65.0 | 58.9 | 68.0 | 55.9 | 70.6 | 33.3 | 69.7 | 42.4 | 61.1 | 0.35 |
| [1] | 77.0 | 67.5 | **77.2** | 68.4 | 54.5 | **68.3** | 72.0 | **56.7** | 44.1 | 34.9 | 62.1 | 0.04 |
| **GO-VOS unsupervised** | **88.2** | **82.5** | 62.7 | **76.7** | **70.9** | 50.0 | **81.9** | 51.8 | **86.2** | **55.8** | **70.7** | 0.91 |

### YouTube-Objects dataset v2.2

| Method | aero | bird | boat | car | cat | cow | dog | horse | moto | train | avg | sec/frame |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [1] | 75.7 | 56.0 | 52.7 | 57.3 | 46.9 | **57.0** | 48.9 | 44.0 | 27.2 | 56.2 | 52.2 | **0.02** |
| HPP[3] | 76.3 | 68.5 | **54.5** | 50.4 | **59.8** | 42.4 | 53.5 | 30.0 | **53.5** | **60.7** | 54.9 | 0.35 |
| **GO-VOS unsupervised** | **79.8** | **73.5** | 38.9 | **69.6** | 54.9 | 53.6 | **56.6** | **45.6** | 52.2 | 56.2 | **58.1** | 0.91 |

### SegTrack v2 dataset

| Task | Method | | IoU | sec/frame |
|---|---|---|---|---|
| Unsupervised | Supervised features | KEY [9] | 57.3 | >120 |
| | | FSEG [4] | **61.4** | N/A |
| | | LVO [14] | 57.3 | N/A |
| | | [10] | 59.3 | N/A |
| | Unsupervised | FST [11] | 54.3 | 4 |
| | | CUT [5] | 47.8 | ≈1.7 |
| | | HPP [3] | 50.1 | **0.35** |
| | | **GO-VOS unsupervised** | **62.2** | 0.91 |

[1] Croitoru et al. In: *ICCV*. 2017. [2] Faktor et al. In: *BMVC*. 2014. [3] Haller et al. In: *ICCV*. 2017. [4] Jain et al. In: *CVPR* (2017). [5] Keuper et al. In: *ICCV*. 2015. [6] Koh et al. In: *CVPR*. 2016. [7] Koh et al. In: *CVPR*. 2017. [8] Lao et al. In: *ECCV*. 2018. [9] Lee et al. In: *ICCV*. 2011. [10] Li et al. In: *CVPR*. 2018. [11] Papazoglou et al. In: *ICCV*. 2013. [12] Prest et al. In: *CVPR*. 2012. [13] Song et al. In: *ECCV*. 2018. [14] Tokmakov et al. In: *ICCV* (2017). [15] Tokmakov et al. In: *CVPR*. 2017. [16] Zhang et al. In: *CVPR*. 2015.