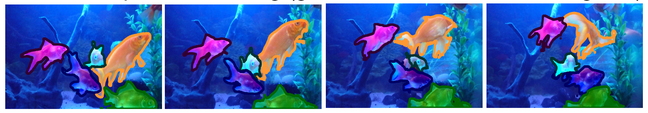
# VideoMatch: Matching based Video Object Segmentation

Yuan-Ting Hu[1] Jia-Bin Huang[2] Alexander G. Schwing[1]
[1] University of Illinois Urbana-Champaign [2] Virginia Tech

## 1. Introduction

### Problem

- Instance level segmentation of multiple objects in videos
- Semi-supervised setting (ground truth of the 1st frame given)



**Challenges:** occlusion, deformation, dynamic background

**Existing methods:** require fine-tuning -> **slow**

### Our work

- Formulates as a matching problem
- Requires no fine-tuning -> **fast**
- On par performance compared to fine-tuned methods

## 2. Overview

### Problem definition

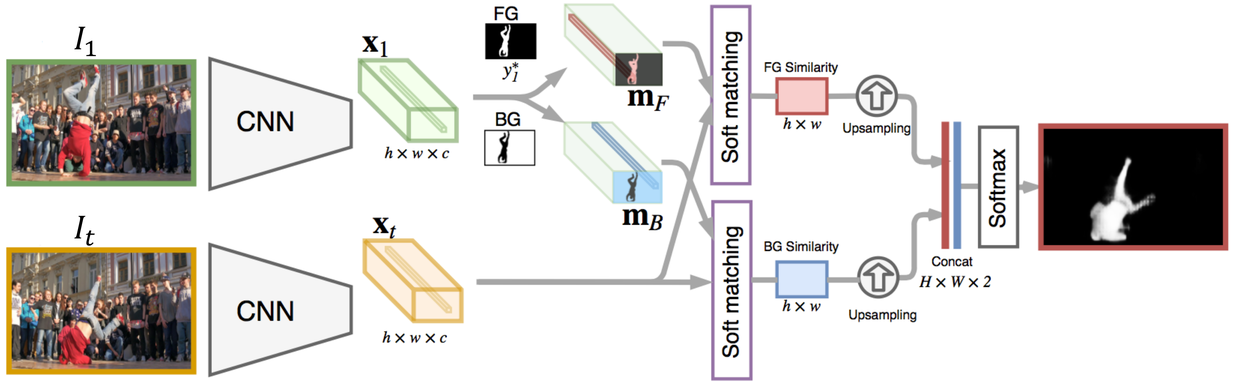
- Input: a video sequence $\{I_1, I_2, \dots, I_T\}$ + ground truth mask for the first frame $y_1^*$
- Goal: predict segmentation mask $y_2, y_3, \dots, y_T$

### Approach

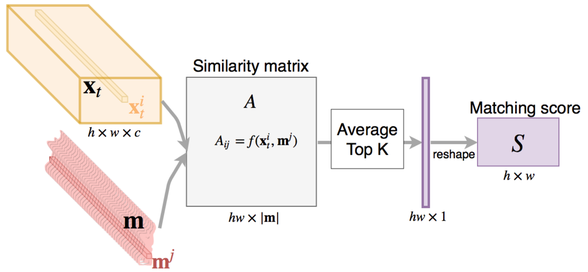
- Match between image $I_t$ and the template $I_1$ using the proposed soft matching layer

### Notation

- $\mathbf{x}_t$: features extracted from frame $I_t$
- $\mathbf{m}_F = \{\mathbf{x}_1^i : i \in \delta(y_1^* = 1)\}$, the set of FG features
- $\mathbf{m}_B = \{\mathbf{x}_1^i : i \in \delta(y_1^* = 0)\}$, the set of BG features




## 3. Soft Matching Layer




**Input:** two sets of features $\mathbf{x}_t$ and $\mathbf{m}$

**Output**: a matching score matrix measuring the compatibility of every pixel in the frame $I_t$ with the FG or BG pixels

- $f(\mathbf{x}_t^i, \mathbf{m}^j)$: a function measuring the similarity between two features $\mathbf{x}_t^i$ and $\mathbf{m}^j$; we use cosine similarity
- Compute average top K along the second axis
- End-to-end trainable

## 4. Online Update

- Remove outliers using the last prediction
- Update $\mathbf{m}_B$: add the features of pixels that are predicted as FG but not in $\hat{y}_{t-1}$
- Update $\mathbf{m}_F$: add the features of pixels that are predicted as FG with high confidence and far from object boundary

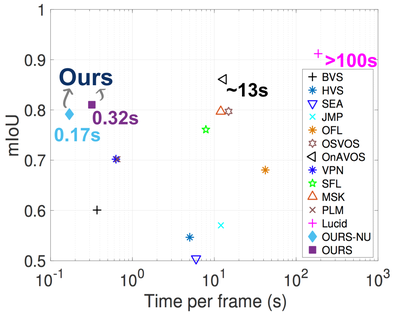
(a) FG pred. $y_{t.\text{init}}$ (b) FG pred. $y_{t-1}$

(c) Extruded pred. $\hat{y}_{t-1}$ (d) Output pred. $y_t$

## 5. Experimental Results

### Quantitative results

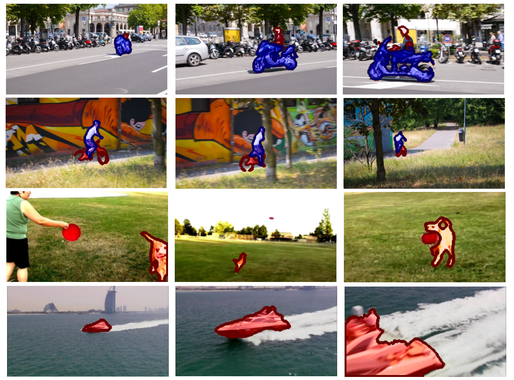
- Intersection over union (IoU) vs speed on DAVIS-16




- IoU on Youtube-Object dataset

| | OURS | OnAVOS | MSK | OSVOS | OFL | JFS |
|---|---|---|---|---|---|---|
| Fine-tuned? | - | Yes | Yes | Yes | - | - |
| Average | **0.797** | 0.793 | 0.718 | 0.783 | 0.776 | 0.74 |

### Qualitative results of our method



### Failure cases

- Similar appearance/tiny objects

### Ablation study

- On DAVIS-16

| RM Outliers | BG Update | FG Update | mIoU |
|---|---|---|---|
| - | - | - | 0.792 |
| ✓ | - | - | 0.805 |
| ✓ | ✓ | - | 0.809 |
| ✓ | ✓ | ✓ | 0.810 |

- Effect of K in the soft matching layer