YouTube-VOS: Sequence-to-Sequence Video Object Segmentation
ECCV 2018
[arXiv: Sep 3, 2018]

YouTube-VOS: A Large-Scale Video Object Segmentation Benchmark
[arXiv: Sep 6, 2018]

Presented By:
Jyoti Kini

Video Object Segmentation

- Segmenting a particular object instance throughout the entire video sequence given only the object mask on the first frame

YouTube-VOS Dataset

[Partial dataset used for model implementation]

Scale	YouTube-VOS (Ours)	YouTube-VOS (Ours)
Videos	3,252	4,453
Categories	78	94
Objects	6,048	7,755
Annotations	133,886	197,272
Duration	217.21	334.81

[Final dataset]

- Human annotators select up to five objects per video clip and annotate every five frames in a 30fps frame rate, resulting in 6fps sampling rate

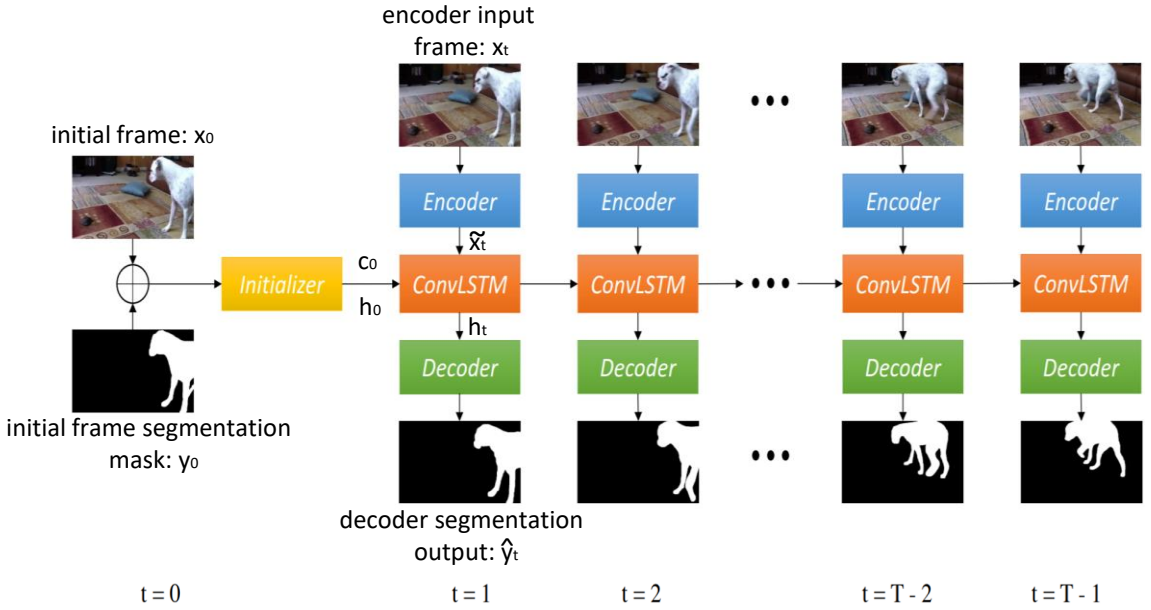
Videos - 4,453

Training - 3,471

Validate – 474 (26 unseen)

Testing – 508 (26 unseen)

Algorithm



$$\mathbf{c}_0, \mathbf{h}_0 = Initializer(\mathbf{x}_0, \mathbf{y}_0)$$
$$\tilde{\mathbf{x}}_t = Encoder(\mathbf{x}_t)$$
$$\mathbf{c}_t, \mathbf{h}_t = ConvLSTM(\tilde{\mathbf{x}}_t, \mathbf{c}_{t-1}, \mathbf{h}_{t-1})$$
$$\hat{\mathbf{y}}_t = Decoder(\mathbf{h}_t)$$
$$\mathcal{L} = -(\mathbf{y_t} \log(\hat{\mathbf{y}}_\mathbf{t})) + ((1 - \mathbf{y_t}) \log(1 - \hat{\mathbf{y}}_\mathbf{t}))$$

Implementation Details

❑ Data Loader:

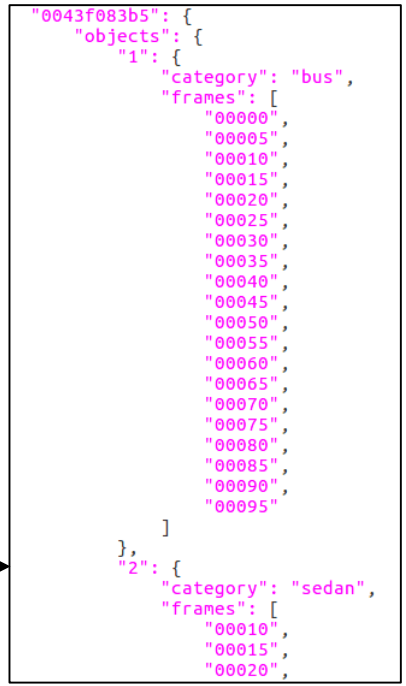
- Dataset directories:

Annotation-6fps
JPEGImages-6fps

Annotation-6fps
JPEGImages-30fps

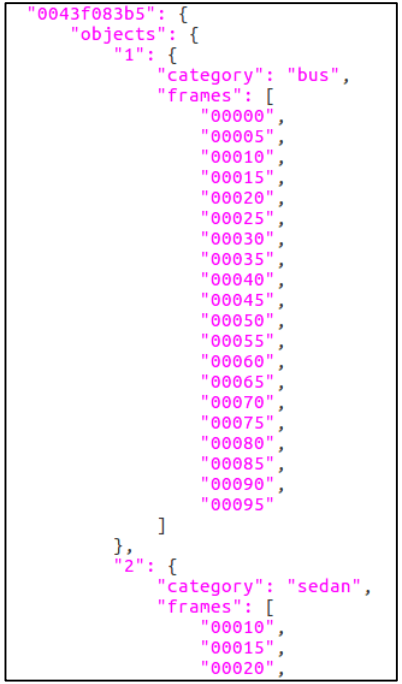
Annotation-6fps
JPEGImages-6fps

Annotation-6fps
JPEGImages-30fps



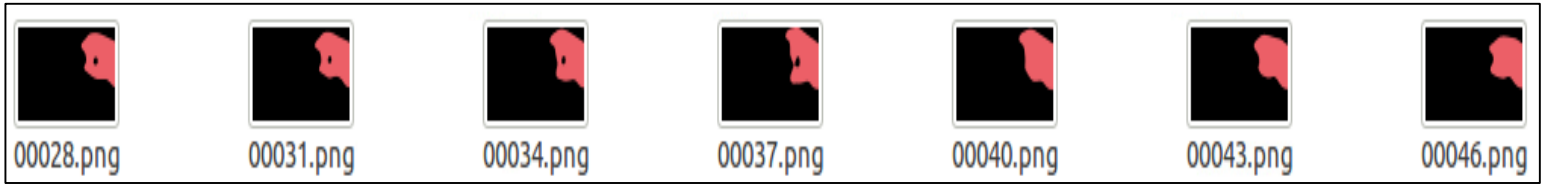
Implementation Details

- **Data Loader:**
  - JSON:
    - Not all objects appear from the 1st frame onwards
    - Colors to binary masks (PIL/OpenCV)
    - Handle occlusion



Implementation Details

- **Data Loader:**
  - CPU threads
  - Random selection of single object instance (PIL/OpenCV)

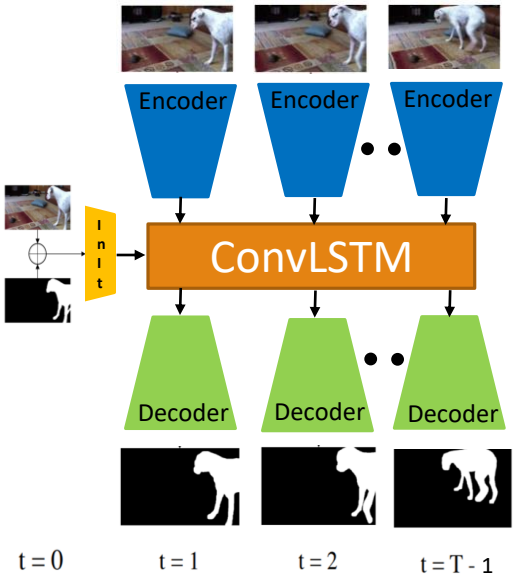


  - Select time frames T (5 ~ 11) from a random training video sequence
  - Resize original RGB frames and annotations to 256×448

Implementation Details

❑ Model structure:

Component	Layer	Filters	Filter size	Initialization	Activation
Initializer	VGG-16 (all convolution)				
	VGG-16 (first fully-connected)		1x1		
	└→Convolution	512	1x1	Xavier	Relu
	└→Convolution	512	1x1	Xavier	Relu
Encoder	VGG-16 (all convolution)				
	VGG-16 (first fully-connected)		1x1		
	Convolution	512	1x1	Xavier	Relu
ConvLSTM	ConvLSTM	512	3x3	Xavier Forget bias=1	Sigmoid - gate outputs Relu - state outputs
Decoder	Deconvolution	512	5×5	Xavier	
	Deconvolution	256	5×5	Xavier	
	Deconvolution	128	5×5	Xavier	
	Deconvolution	64	5×5	Xavier	
	Deconvolution	64	5×5	Xavier	
	Convolution	1	5×5	Xavier	Sigmoid



Implementation Details

- **Training:**
  - Primarily, trained on 6fps, followed by 30fps as the loss stabilizes
  - Loss for frames without ground-truth is set to 0
  - Model converges in 80 epochs
  - Initial learning rate is set to 10^{-5}

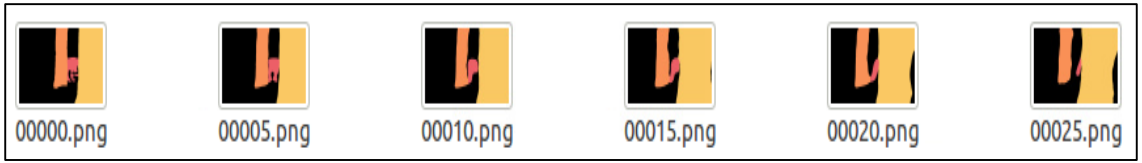
Implementation Details

❑ Evaluation:

- Submission access URL

(https://competitions.codalab.org/competitions/19544)

- Submission format



Input:

```
<submission>.zip
        |- Annotations
            |- <video_id>
                |- <frame_id>.png
                |- <frame_id>.png
            |- <video_id>
                |- <frame_id>.png
                |- <frame_id>.png
```

Output:

Overall: 0.474763296366
J_seen: 0.548071686125
J_unseen: 0.419488520321
F_seen: 0.495602588958
F_unseen: 0.315890390059

Segmentation Metrics

- **Region similarity:** Defined as the intersection-over-union between the estimated segmentation and the ground truth mask

$$\mathcal{J} = \frac{|M \cap G|}{|M \cup G|}$$

Region Similarity is the intersection-over-union between mask M and ground truth G

- **Contour accuracy:** Interprets the masks as a set of closed contours and computes the contour-based F-measure which is a function of precision and recall

$$\mathcal{F} = \frac{2 P_c R_c}{P_c + R_c}$$

Contour Accuracy is the F-measure for the contour based precision and recall

Questions