Unsupervised learning from video to detect foreground objects in single images

Ioana Croitoru[1], Simion-Vlad Bogolin[1], Marius Leordeanu[1, 2]

[1]Institute of Mathematics of the Romanian Academy and [2]University "Politehnica" of Bucharest

Published at the IEEE International Conference on Computer Vision (ICCV) 2017

Objectives

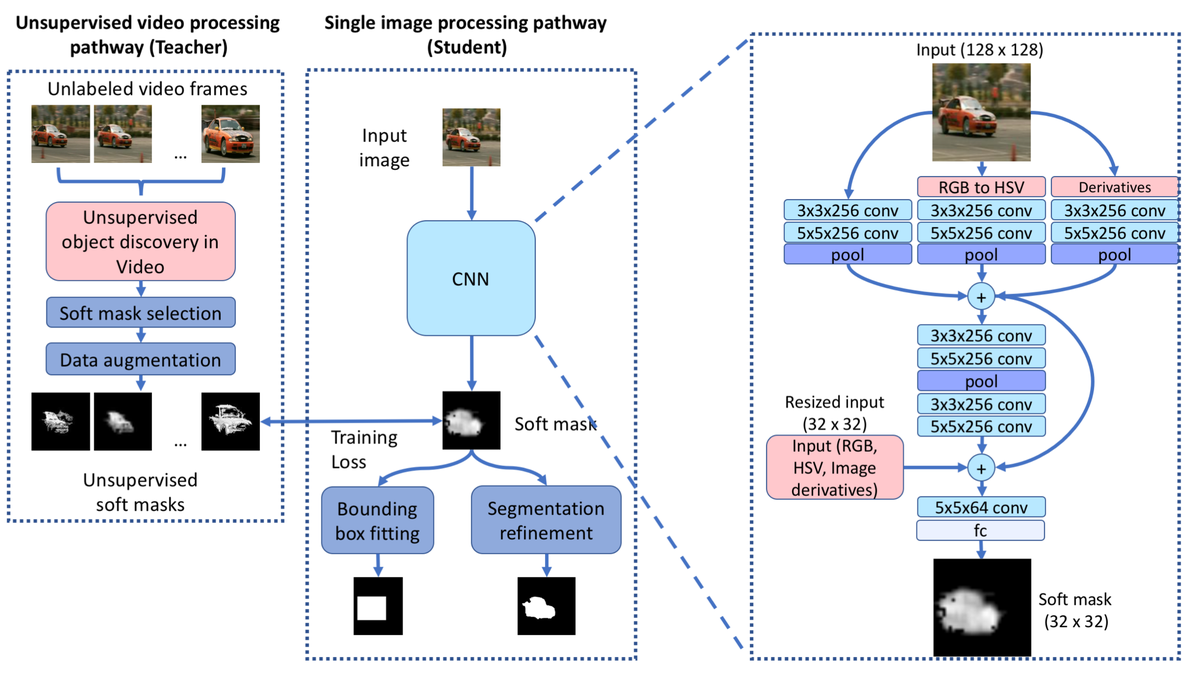
We propose a system that learns to **detect foreground objects in single images in an unsupervised way**. It consists of:

- **Unsupervised video processing pathway (teacher)** that discovers objects in video.
- **Single image processing pathway (student)** that learns from the video discoverer to detect objects in single images.

Introduction

A deep convolutional net learns from a teacher that performs unsupervised object segmentation in video, to produce similar object masks in single frames. While the teacher method takes advantage of the consistency in appearance, shape and motion of objects in video, over space and time, the student eventually learns general object features at higher semantic levels.

System architecture

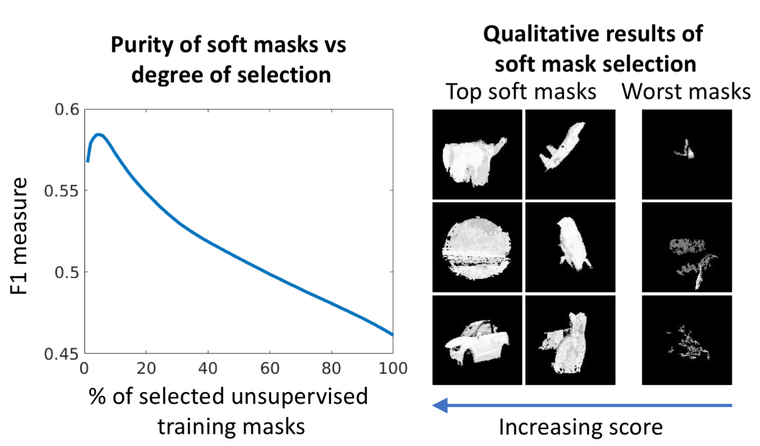


Unsupervised discovery in video

We used the fast VideoPCA algorithm, part of the system in [1]. VideoPCA models the background in video frames with PCA. Foreground and background pixels are estimated, in each frame, from PCA reconstruction error. Color models of foreground and background are learned and used to quickly produce soft object masks.

Data selection and augmentation

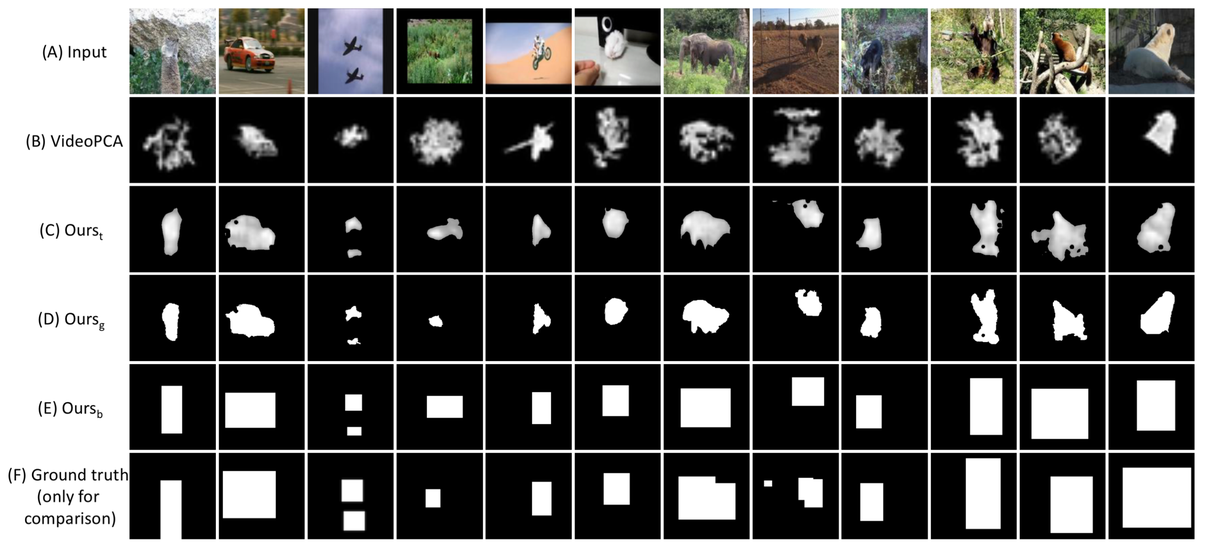
Often the resulting VideoPCA soft masks are noisy due to various challenges present in video. We applied a simple measure of masks quality based on the observation that soft masks that are closer to the ground truth have a higher mean pixel value. We used this mean value to automatically select for training, masks that are more likely to be correct. We also augment the training set of good quality pairs (image, masks) by scaling and random crop of fixed size.



Results on video datasets

Quantitative results on Youtube Objects [2] and qualitative results on VID [3]. Note that our method outperforms its teacher, despite being limited to a single image input.

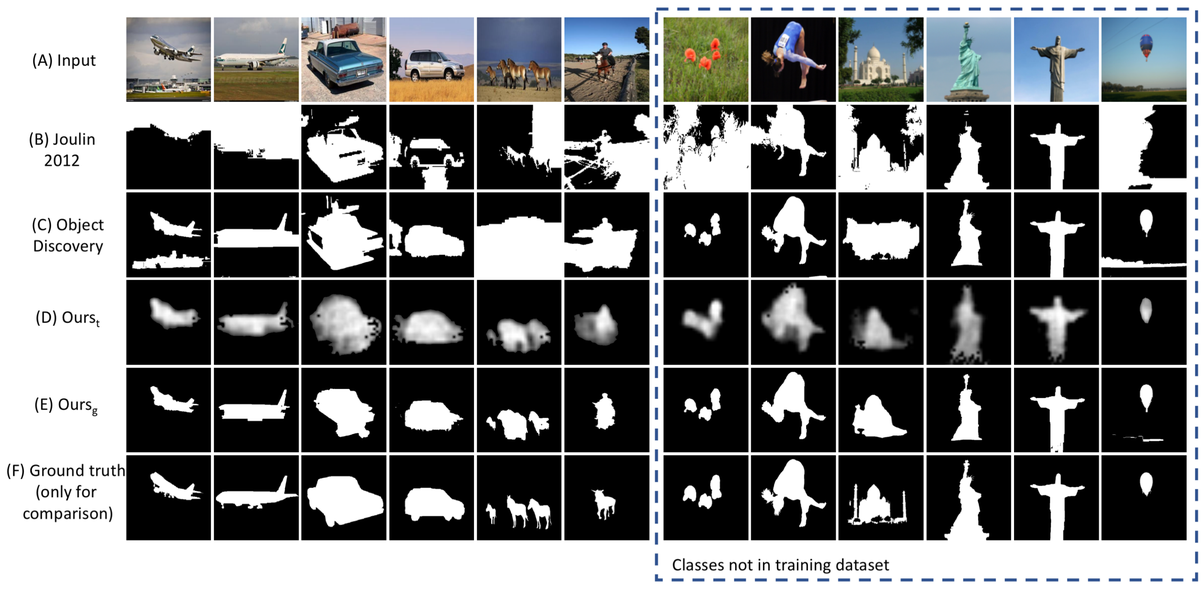
Method	Aero	Bird	Boat	Car	Cat	Cow	Dog	Horse	Mbike	Train	Avg	Time
[4]	64.3	63.2	73.3	**68.9**	44.4	62.5	71.4	52.3	**78.6**	23.1	60.2	N/A
Ours_{VID}	69.8	59.7	65.4	57.0	50.0	**71.7**	**73.3**	46.7	32.4	34.9	56.1	0.04s
$\text{Ours}_{\text{VID+YTO}}$	**77.0**	**67.5**	**77.2**	68.4	**54.5**	68.3	72.0	**56.7**	44.1	**34.9**	**62.1**	0.04s



Results on image datasets

	Airplane		Car		Horse	
	P	J	P	J	P	J
[6]	90.25	40.33	**87.65**	64.86	86.16	33.39
Ours_{VID}	90.92	**62.76**	85.15	66.39	**87.11**	54.59
$\text{Ours}_{\text{VID+YTO}}$	**91.41**	61.37	86.59	**70.52**	87.07	**55.09**

Results on the Object Discovery in Internet images [5] dataset.



Comparison with [7] (B) and [5] (C). Sharp masks could be obtained after refining with GrabCut [8](E). Note that the student net is able to segment objects in images for classes it has not seen during training.

The **code is available** on our project page: `http://sites.google.com/view/unsupervisedlearningfromvideo`

References

[1] O. Stretcu and M. Leordeanu, "Multiple frames matching for object discovery in video.," in *BMVC*, 2015.
[2] A. Prest *et al.*, "Learning object class detectors from weakly annotated video," in *CVPR*, pp. 3282–3289, IEEE, 2012.
[3] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, 2015.
[4] J. Koh *et al.*, "Pod: Discovering primary objects in videos based on evolutionary refinement of object recurrence, background, and primary object models," in *CVPR*, 2016.
[5] M. Rubinstein *et al.*, "Unsupervised joint object discovery and segmentation in internet images," in *CVPR*, 2013.
[6] X. Chen *et al.*, "Enriching visual knowledge bases via object discovery and segmentation," in *CVPR*, 2014.
[7] A. Joulin *et al.*, "Multi-class cosegmentation," in *CVPR*, 2012.
[8] C. Rother *et al.*, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM Transactions on Graphics*, vol. 23, pp. 309–314, 2004.