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 [2]University "Politehnica" of Bucharest
ioana.croi@gmail.com vladbogolin@gmail.com marius.leordeanu@imar.ro

Abstract

Unsupervised learning from visual data is one of the most difficult challenges in computer vision. It is essential for understanding how visual recognition works. Learning from unsupervised input has an immense practical value, as huge quantities of unlabeled videos can be collected at low cost. Here we address the task of unsupervised learning to detect and segment foreground objects in single images. We achieve our goal by training a student pathway, consisting of a deep neural network that learns to predict, from a single input image, the output of a teacher pathway that performs unsupervised object discovery in video. Our approach is different from the published methods that perform unsupervised discovery in videos or in collections of images at test time. We move the unsupervised discovery phase during the training stage, while at test time we apply the standard feed-forward processing along the student pathway. This has a dual benefit: firstly, it allows, in principle, unlimited generalization possibilities during training, while remaining fast at testing. Secondly, the student not only becomes able to detect in single images significantly better than its unsupervised video discovery teacher, but it also achieves state of the art results on two current benchmarks, YouTube Objects and Object Discovery datasets. At test time, our system is two orders of magnitude faster than other previous methods.

1. Introduction

Unsupervised learning is one of the most difficult and intriguing problems in computer vision and machine learning today. Researchers believe that unsupervised learning from video could help decode hard questions regarding the nature of intelligence and learning. As unlabeled videos are easy to collect at low cost, solving this task would bring a great practical value in vision and robotics. Recent unsupervised methods follow two directions. One is to learn powerful features in an unsupervised way and then use them in a classic supervised learning scheme in combination with different classifiers, such as SVMs or CNNs [31, 24, 22]. In very recent work [28], developed independently from ours, a deep network learns, from an unsupervised system using motion cues in video, image features that are applied to several transfer learning tasks. The second main approach to unsupervised learning is to discover, at test time, common patterns in unlabeled data using clustering, feature matching or data mining formulations [11, 7, 40]. Unsupervised learning in video is also related to co-segmentation [13, 18, 36, 14, 20, 43, 37] and weakly supervised localization [9, 25, 39]. Earlier methods are based on local feature matching and detection of their co-occurrence patterns [41, 40, 21, 27, 23], while more recent ones [15, 33] discover object tubes by linking candidate bounding boxes between frames with or without refining their location. Traditionally, the task of unsupervised learning from image sequences, has been formulated as a feature matching or data clustering optimization problem, which is computationally very expensive.

Our system is presented in Figure 1. We have an unsupervised training stage, in which a student deep neural network (Figure 2) learns frame by frame from an unsupervised teacher, which performs object segmentation in videos, to produce similar object masks in single images. The teacher method takes advantage of the consistency in appearance, shape and motion manifested by objects in video. In this way, it discovers objects in the video and produces a foreground segmentation for each individual frame. Then, the student network tries to imitate for each frame the output of the teacher, while having as input only a single image - the current frame. The teacher pathway is much simpler in structure, but it has access to information over time. In contrast, the student is much deeper in structure, but has access only to one image. Thus, the information discovered by the teacher in time is captured by the student in depth, over neural layers of abstraction. In experiments, we show a very encouraging fact: the student easily learns to outperform its teacher and discovers by itself general knowledge about the shape and appearance properties of objects,

well beyond the abilities of the teacher. Thus, the student produces significantly better object masks, which generally have a good form, do not have holes and display smooth contours, while having an appearance that is often in contrast to the background scene. Since there are available methods for video discovery with good performance, the training task becomes immediately feasible. In this work we chose the VideoPCA algorithm introduced as part of the system in [41] because it is very fast (50-100 fps), uses very simple features (pixel colors) and it is unsupervised - with no usage of supervised pre-trained features. That method exploits the stability in appearance and location of objects, which is common in video shots. While the discovered object masks are far from being perfect and are often noisy, the student network manages to generalize and overcome some of these limitations. We propose a ten layer deep neural network for the student pathway (Figure 2). It takes as input the original RGB, HSV and image spatial derivatives channels. It outputs a low resolution soft segmentation mask of the main objects present in a given image.

Main contributions: Our main contributions are:

1) Our system, to our best knowledge, is the first that learns to detect and segment foreground objects in images in an unsupervised fashion, with no pre-trained features needed or manual labeling, while requiring only a single image at test time.

2) The proposed architecture is novel. It consists of two processing pathways, with complementary functions. The first pathway discovers foreground objects in videos in an unsupervised way and has access to all the video frames. It acts as a teacher. The second "student" pathway, which is a deep convolutional net, learns to predict the teacher's output for each frame while having access only to a single input image. The student learns to outperform its teacher, despite being limited to a single image input. Once trained, the student achieves state of the art results on two important datasets.

2. Approach and intuition

There are several observations that motivate the approach we take for addressing the unsupervised learning task. First, we notice that unsupervised learning methods are generally more effective when considering video input, in which objects satisfy spatio-temporal consistency, with smooth variations in shape, appearance and location over time. For that matter it is usually harder to learn about objects from collections of images that are independently taken. This motivates the video discovery pathway, based on the VideoPCA algorithm introduced in [41], which is both very fast, reasonably accurate and uses extremely simple cues - individual pixel colors, in combination with several stages that take advantage of spatio-temporal consistency and the contrasting properties of foreground and background. Next, if we want the student pathway to learn general principles about objects in images, we need to limit its access to a single input image. Otherwise, if given the entire video as input, a powerful deep network would easily overfit when trained to predict the teacher's output.

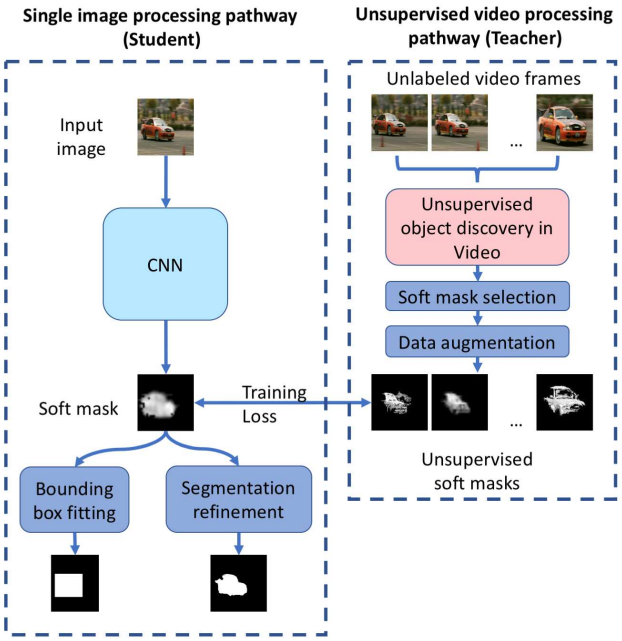

Figure 1. The dual student-teacher system proposed for unsupervised learning to detect foreground objects in images. It has two pathways: the teacher, on the right, discovers in an unsupervised fashion foreground objects in video. It outputs soft masks for each frame. The resulting masks, are then filtered based on a simple and effective unsupervised quality metric. The set of selected segmentations is then augmented in a relatively simple manner, automatically. The resulting final set of pairs - input image (a video frame) and soft mask (the mask for that particular frame which acts as an unsupervised label) - are used for training the student CNN pathway.

An important question that needs to be answered is whether the student can outperform its teacher. If this is the case, then the student has an important accuracy advantage over its teacher, besides being faster. We could envision the potential practical benefit of unsupervised learning - especially when there is so much unlabeled video data available. Thus, we first have to make sure that the student receives only the best quality input possible from the teacher. For that we add an extra module for unsupervised soft masks selection. It is based on a simple and intuitive measure of quality (explained later) which does a good job at ordering masks with respect to their true quality. Then, we also need to make sure that the student sees as much training data as possible. So, we design an automatic data augmentation

module, which creates extra training data by randomly scaling and shifting the masks provided by the teacher after the mask selection procedure.

In our experiments, the student indeed outperforms its teacher. Moreover, it achieves state of the art results on two different benchmarks. The success of this unsupervised learning paradigm is due to the fact that the student is forced to capture from appearance only (as it is limited to a single image) visual features that are good predictors for the presence of objects.

3. System architecture

We now detail the architecture of our system, module by module, as seen in Figure 1.

3.1. Teacher path: unsupervised discovery in video

There are several methods available for discovering objects and salient regions in images and videos [4, 6, 10, 12, 8, 3], with reasonably good performance. More recent methods for foreground objects discovery such as [26] are both relatively fast and accurate, with runtime above 4 seconds per frame. However, that runtime is still long and prohibitive for training the student CNN that requires millions of images. For that reason we used the VideoPCA algorithm, which is a part of the whole system introduced in [41]. It has lower accuracy than the full system, but it is much faster, running at $50 - 100$ fps. At this speed we can produce one million unsupervised soft segmentations in a reasonable time of about 5-6 hours.

VideoPCA models the background in video frames with Principal Component Analysis. It finds initial foreground regions as parts of the frames that are not reconstructed well with the PCA model. Foreground objects are smaller than the background and have more complex movements, which make them less likely to be captured well by the first PCA components. The initial soft masks are obtained from the error image, the difference between the original image and the PCA reconstruction. These "errors" are smoothed with a large Gaussian and thresholded. The binary masks obtained are used to learn color models of foreground and background, based on which individual pixels are classified as belonging to foreground or not. The object masks obtained are further multiplied with a large centered Gaussian, assuming that foreground objects are often closer to the image center. For more details the reader is invited to consult [41]. In this work, we use the method exactly as found online[1] without any parameter tuning.

3.2. Student path: single-image segmentation

The student processing pathway (Figure 1) consists of a deep convolutional network, with ten layers (seven convolutional, two pooling and one fully connected layer) and skip connections as shown in Figure 2. All layers use ReLU activation functions. We chose this CNN based on its relative simplicity and strong performance. Skip connections have proved to provide a boost in the network's performance [32, 29]. We also observed a slight improvement in our case ($\approx$ %1). The net takes as input a 128×128 color image (along with its hue, saturation, derivatives w.r.t. x and y) and produces a 32×32 soft segmentation of the main objects present in the image. While it does not identify the particular object classes, it learns from the unsupervised soft-masks provided by the teacher to detect and softly segment the main foreground objects present, regardless of their particular category, one frame at a time. Thus, as shown in experiments, it is also able to detect and segment classes it has never seen before.

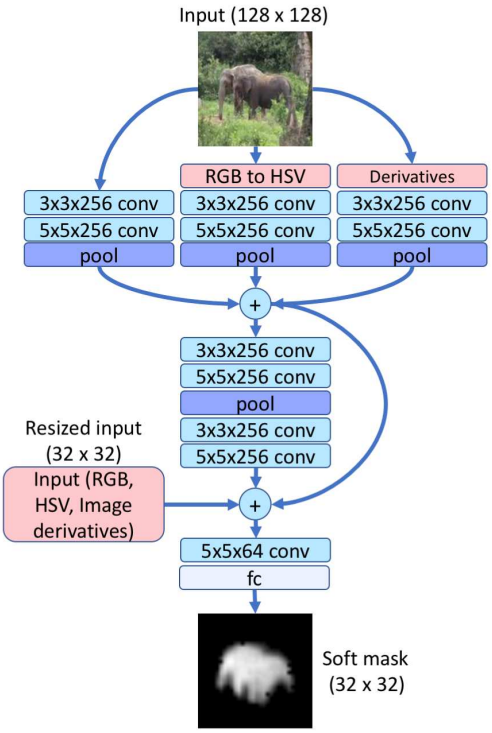

Figure 2. The "student" deep convolutional net that processes single images. It is trained to predict the unsupervised labels given by the teacher pathway, frame by frame. We observed that by adding at the last level the original input and mid-level features (skip connections) and resizing them appropriately, the performance increases.

We treat foreground object segmentation as a regression problem, where the soft mask given by the unsupervised video segmentation system acts as the desired output. Let $\mathbf{I}$ be the input RGB image (a video frame) and $\mathbf{Y}$ be the corresponding 0-255 valued soft segmentation given by the unsupervised teacher pathway for that particular frame. The

[1] https://sites.google.com/site/multipleframesmatching/

goal of our network is to predict a soft segmentation mask $\hat{\mathbf{Y}}$ of width $W = 32$ and height $H = 32$, that approximates as well as possible the mask $\mathbf{Y}$. For each pixel in the output image, we predict a 0-255 value, so that the total difference between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ is minimized. So, given a set of N training examples, let $\mathbf{I}^{(n)}$ be the input image (a video frame), $\hat{\mathbf{Y}}^{(n)}$ be the predicted output mask for $\mathbf{I}^{(n)}$, $\mathbf{Y}^{(n)}$ the soft segmentation mask (corresponding to $\mathbf{I}^{(n)}$) and $\mathbf{w}$ the network parameters. $\mathbf{Y}^{(n)}$ is produced by the video discoverer after processing the video that $\mathbf{I}^{(n)}$ belongs to. Then, our loss is:

$$L(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{p=1}^{W\times H}\left(\mathbf{Y}_p^{(n)} - \hat{\mathbf{Y}}_p^{(n)}(\mathbf{w}, \mathbf{I}^{(n)})\right)^2 \qquad (1)$$

where $\mathbf{Y}_p^{(n)}$ and $\hat{\mathbf{Y}}_p^{(n)}$ denotes the p-th pixel from $\mathbf{Y}^{(n)}$, respectively $\hat{\mathbf{Y}}^{(n)}$.

We observed that in our tests, the L2 loss performed better than the cross-entropy loss. We train our network using the Tensorflow [1] framework with the Adam optimizer [19]. All our models are trained end-to-end using a fixed learning rate of 0.001 for about 10 epochs. The training time for a given model is about 3 days on a Nvidia GeForce GTX 1080 GPU.

Post-processing: Our CNN outputs a 32×32 soft mask. In order to fairly compare our models with other methods, we have two different post processing steps: 1) bounding box fitting and 2) segmentation refinement. For fitting a box around our soft mask, we first up-sample the 32×32 output to the original size of the image, then threshold the mask (validated on a small subset), determine the connected components, filter out the small ones (smaller than half the size of the largest one) and finally fit a tight box around each of the remaining components. When we are interested in obtaining a fine object segmentation, we use the OpenCV implementation of the GrabCut [35] method to refine our soft mask, up-sampled to the original size.

3.3. Unsupervised soft masks selection

The performance of the student pathway is influenced by the quality of the soft masks provided as labels by the video discovery path. The cleaner the masks provided by the teacher, the more chances the student has to actually learn to segment well the objects in images. VideoPCA used by the video processing path usually has good results if the object present in the video stands out against the background scene, in terms of motion and appearance. However, if the object is occluded at some point, if it does not move w.r.t the scene or if it has a similar appearance to its background, the resulting soft masks might be poor. We used a simple measure of masks quality based on the following

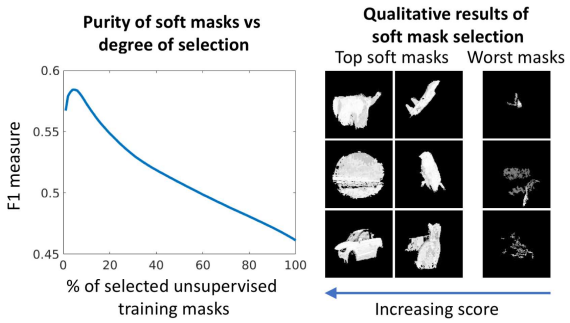

Figure 3. Purity of soft masks vs degree of selection. When selection rate decreases, the true purity of the training frames improves. Our automatic selection method is not perfect: some low quality masks have high scores and we remove some good segmentations.

observation: when masks are close to the ground truth, the mean of their nonzero values is usually high. Thus, when the discoverer is confident is more likely to be right. The mean value of non-zero pixels in the soft mask is then used as a score indicator for each segmented frame.

Next we sort all soft masks in the entire training dataset (e.g. VID [38], YTO [30]) in descending order of their mean score and keep only the top k percent. In this way, we obtain a very simple unsupervised selection method. In Figure 3 we present the dependency of segmentation performance w.r.t ground truth object boxes (used only for evaluation) vs. the percentile k of masks kept after the automatic selection. In other words, the fewer frames we select the more likely it is that they are correctly segmented. This procedure is not perfect, so we sometimes remove good segmentations during this masks selection step. Even though we can expect to improve the quality of the unsupervised masks by drastically pruning them, the fewer we are left with, the less training data we get, increasing the chance to overfit. We make up for the losses in training data by augmenting the set of training masks (Sec. 3.4) and by bringing in unlabeled videos from other datasets.

Thus, the more selective we are about what masks to accept for training, the more videos we need to collect and process with the teacher pathway, in order to improve generalization.

3.4. Data augmentation

Another drawback of VideoPCA is that it can only detect the main object if it is close to the center of the image. The assumption that the foreground is close to the center is often true and indeed helps that method to produce soft masks with a relatively high precision. It fails when the object is not in the center, therefore its recall is relatively low. Our data augmentation procedure also addresses this limitation. That module can be concisely described as fol-

lows: scale the input image and the corresponding soft mask given by the video discovery framework at a higher resolution (160×160) and randomly crop 128×128 patches from the scaled version. Finally, down-scale each soft mask to 32×32. This would produce slightly larger objects at locations that cover the whole image area, not just the center. As our experiments show, the student net is able to see objects at different locations in the image, unlike its raw teacher, which is strongly biased towards the image center. Data selection along with data augmentation of the training set significantly improve unsupervised learning, as shown in the experiments section (Sec. 4).

4. Experimental analysis

The experiments we conducted aim to highlight various aspects of the performance of our method. Firstly, we compare the quality of the segmentations obtained by the feed-forward CNN against its teacher, VideoPCA (Sec. 4.1). Secondly, we tested that adding extra unlabeled videos improves performance (Sec. 4.2). Finally, we compare the performance of our unsupervised system to state of the art approaches for object discovery in video, on the YouTube Objects Dataset [30] benchmark, and object discovery in images, on the Object Discovery in Internet Images [36] benchmark (Sec. 4.3).

4.1. Unsupervised learning from ImageNet

It is a well known fact that the performance of a convolutional network strongly depends on the amount of data used for training. Because of this, we chose to use as our primary training dataset the ImageNet Object Detection from Video (VID) dataset [38]. VID is one of the largest video datasets publicly available, being fully annotated with ground truth bounding boxes. The large set of annotations available allowed us to have a thorough evaluation of our unsupervised system. The dataset consists of about 4000 videos, having a total of about 1.2M frames. The videos contain objects that belong to 30 different classes. Each frame could have zero, one or multiple objects annotated. The benchmark challenge associated with this dataset focuses on the supervised object detection and recognition problem, which is different from the problem that we tackle here. Our system is not trained to identify different object categories. On the VID dataset we evaluated the student CNN against its teacher pathway. We measure performance of soft-masks by maximum F-measure computed w.r.t ground truth bounding box, by considering pixels inside the bounding box as true positives and those outside as true negatives. This simple metric allows us to evaluate the soft masks directly, without any post-processing steps.

We tested our unsupervised system on the validation split of the VID dataset. As it can be seen from Table 1 the student outperforms its teacher (VideoPCA) by a very significant margin. Also, in Figure 4 we present some qualitative results on this dataset as compared to VideoPCA. We can see that the masks produced by VideoPCA are of lower quality, often having holes, non-smooth boundaries and strange shapes. In contrast, the student learns more general shape and appearance characteristics of objects in images, reminding of the grouping principles governing the basis of visual perception as studied by the Gestalt psychologists [34] and the more recent work on the concept of "objectness" [2]. The object masks produced by the student are simpler, with very few holes, have nicer and smoother shapes and capture well the figure-ground contrast and organization. Another interesting observation is that the network is able to detect multiple objects, a feature that is less commonly achieved by the teacher.

Method	F1 measure	Dataset
VideoPCA [41]	41.83	-
Baseline	51.17	VID
Baseline	51.9	VID + YTO
Refined	52.51	VID
Data selection 5%	53.20	VID
Data selection 10%	53.82	VID
Data selection 30%	53.67	VID
Data selection 10%	54.53	VID + YTO

Table 1. Results on the VID dataset [38]. The "dataset" column refers to the datasets used for training the student network. Our baseline model is represented by a classic CNN having only the RGB image as input and no skip-connections. The refined model is our final student CNN model as presented in Figure 2. The data selection entries refer to the percentage of kept soft masks after applying our selection method. All masks selection experiments were conducted using the refined model. We highlight that the overall system performance improves with the amount of selectivity, which shows that a simple quality measure used for soft mask selection can improve the performance of the CNN image-based pathway. Thus, the data augmentation module makes up for the frames lost during the selection process.

4.2. Adding more data

We also tested how adding more unlabeled data affects the overall performance of our system. Therefore, we added the Youtube Objects(YTO) dataset to the existing VID dataset. The YTO dataset is a weakly annotated dataset that consists of about 2500 videos, having a total of about 720K frames, divided into 10 classes. Adding more unlabeled videos (from YTO, without annotations) to the unsupervised training set clearly improves performance as reported in Tables 3, 1 and 4. The capacity of our system to improve its performance in the presence of unlabeled data, without degradation or catastrophic forgetting is mainly due to the robustness of the teacher pathway combined with data selection and augmentation, in conjunction with the

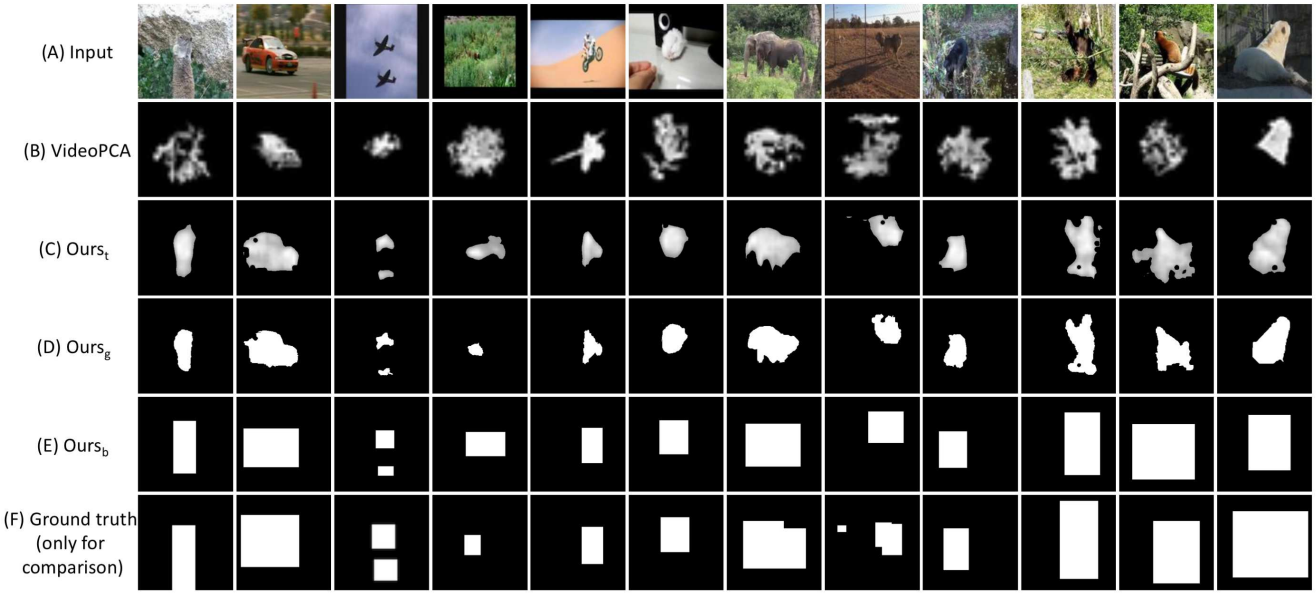

Figure 4. Visual results on the VID dataset [38] compared to the teacher method. A: current frame, B: soft mask produced by VideoPCA [41] for the current frame, after processing the entire video, C: thresholded soft mask produced by our network, D: segmentation mask produced after refining the soft output of our network with GrabCut [35], E: bounding box obtained from the soft segmentation mask; F: ground truth bounding box.

tendency of the single-image CNN net to improve over its teacher.

As it comes to the soft mask selection, our experiments show that we obtain the best overall results by using the top 10% soft masks with data augmentation. All the experiments are conducted using this setup for each dataset.

4.3. Comparisons with other methods

Single image discovery methods Next, we compare our unsupervised system with state of the art methods designed for the task of object discovery in collections of images, that might contain one or a few main object categories of interest. A representative current benchmark in this sense is the Object Discovery in Internet Images dataset. This set contains Internet images and it is annotated with high detail segmentation masks. In order to enable comparison with previous methods, we use the 100 images subsets.

The methods evaluated on this dataset, in the literature, aim to either discover the bounding box of the main object in a given image, or its fine segmentation mask. We evaluate our system on both. Different from other methods, we do not need a collection of images during testing, since each image is processed independently by our system, at test time. Therefore, our performance is not affected by the structure of the image collection or the number of classes of interest being present in the collection.

For evaluating the detection of bounding boxes the most used metric is CorLoc defined as the percentage of images correctly localized according to the PASCAL criterion: $\frac{B_p \cap B_{GT}}{B_p \cup B_{GT}} \geq 0.5$, where B_P is the predicted bounding box and B_{GT} is the ground truth bounding box. In Table 2 we present the performance of our method as compared to other unsupervised object discovery methods in terms of CorLoc on the Object Discovery dataset. We compare our predicted box against the tight box fitted around the ground-truth segmentation as done in [7, 42]. Our system can be considered in the mixed class category: it does not depend on the structure of the image collection. It treats each image independently. The performance of the other

Method	Airplane	Car	Horse	Avg
Kim *et al.* [18]	21.95	0.00	16.13	12.69
Joulin *et al.* [13]	32.93	66.29	54.84	51.35
Joulin *et al.* [14]	57.32	64.04	52.69	58.02
Rubinstein *et al.* [36]	74.39	87.64	63.44	75.16
Tang *et al.* [42]	71.95	93.26	64.52	76.58
Cho *et al.* [7]	82.93	94.38	*75.27*	84.19
Cho *et al.* [7] mixed	81.71	94.38	70.97	82.35
Ours$_{\text{VID}}$	**93.90**	93.26	70.97	**86.04**
Ours$_{\text{VID+YTO}}$	87.80	**95.51**	**74.19**	85.83

Table 2. Results on the Object Discovery in Internet images [36] dataset (CorLoc metric). Ours$_{\text{VID}}$ represents our network trained using the VID dataset (with 10% selection), while Ours$_{\text{VID+YTO}}$ represents our network trained on VID and YTO datasets (with 10% selection).

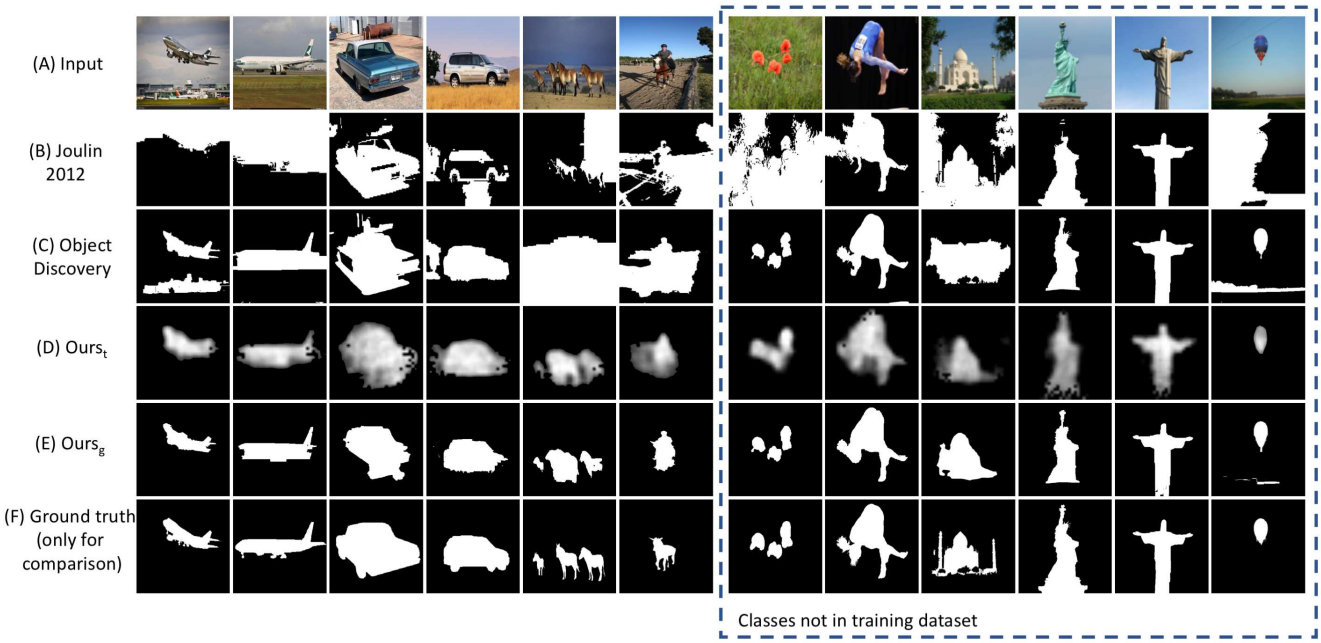

Figure 5. Visual results on the Object Discovery dataset. A: input image, B: segmentation obtained by [14], C: segmentation obtained by [36], D: thresholded soft mask produced by our network, E: segmentation mask produced after refining the soft output of our network with GrabCut [35], F: ground truth segmentation. More details and results: https://sites.google.com/view/unsupervisedlearningfromvideo.

	Airplane		Car		Horse	
	P	J	P	J	P	J
[18]	80.20	7.90	68.85	0.04	75.12	6.43
[13]	49.25	15.36	58.70	37.15	63.84	30.16
[14]	47.48	11.72	59.20	35.15	64.22	29.53
[36]	88.04	55.81	85.38	64.42	82.81	51.65
[5]	90.25	40.33	**87.65**	64.86	86.16	33.39
Ours$_1$	90.92	**62.76**	85.15	66.39	**87.11**	54.59
Ours$_2$	**91.41**	61.37	86.59	**70.52**	87.07	**55.09**

Table 3. Results on the Object Discovery in Internet images [36] dataset (P, J metric). Ours$_1$ represents our network trained using the VID dataset (with 10% selection), while Ours$_2$ represents our network trained on VID and YTO datasets (with 10% selection). We observe that Ours$_2$ has better results with mean P of **88.36** and mean J of **62.33** compared to Ours$_1$ (mean P: 87.73, mean J: 61.25).

algorithms degrades as the number of main categories increases in the collection (some are not even tested by their authors on the mixed-class case).

We obtain state of the art results on all classes (in the mixed class case), improving by a significant margin over the method of [7]. When the method in [7] is allowed to see a collection of images that are limited to a single majority class, its performance improves and outperforms ours on one class. However, the comparison is not truly appropriate since our method has no other information necessary besides the input image, at test time.

We also tested our system on the task of fine foreground object segmentation and compared to the best performers in the literature on the Object Discovery dataset in Table 3. For refining our soft masks we apply the GrabCut method, as it is available in OpenCV. We evaluate based on the same P, J evaluation metric as described by Rubinstein *et al*. [36] - the higher P and J, the better. P refers to the per pixel precision, while J is the Jaccard similarity (the intersection over union of the result and ground truth segmentations). In Figure 5 and 6 we present some qualitative samples from each class.

Video discovery methods We also performed comparisons with methods specifically designed for object discovery in video. For that, we choose the YouTube Objects dataset and compared to the best performers on this dataset in the literature (Table 4). Evaluations are conducted on both versions of YouTube Objects dataset, YTOv1 [30] and YTOv2.2 [17]. On YTOv1 we follow the same experimental setup as [16, 30], by running experiments only on the training videos. We have not included in Table 4 the results reported by [41] because they use a different setup, testing on all videos from YTOv1. It is important to stress out again the fact that while the methods presented here for comparison have access to whole video shots, ours only needs a single image at test time. Despite this limitation,

Figure 6. Qualitative results on the Object Discovery in Internet Images [36] dataset. For each example we show the input RGB image (first and third row) and immediately below (second and fourth row) our refined segmentation result obtained by applying GrabCut on the soft segmentation mask predicted by our network. Note that our method produces good quality segmentation results, even in cases with cluttered background.

Method	Aero	Bird	Boat	Car	Cat	Cow	Dog	Horse	Mbike	Train	Avg	Time	Version
[30]	51.7	17.5	34.4	34.7	22.3	17.9	13.5	26.7	41.2	25.0	28.5	N/A	v1 [30]
[26]	65.4	67.3	38.9	65.2	46.3	40.2	65.3	48.4	39.0	25.0	50.1	4s	
[16]	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	
$\text{Ours}_{\text{VID+YTO}}$	75.7	56.0	52.7	57.3	46.9	57.0	48.9	44.0	27.2	56.2	52.2	0.04s	v2.2 [17]

Table 4. Results on Youtube Objects dataset [30]. Ours_{VID} represents our network trained using the VID dataset (with 10% selection), while $\text{Ours}_{\text{VID+YTO}}$ represents our network trained on VID and YTO datasets (with 10% selection). Note that our system has a significantly lower per frame test time than [26] which we estimate that is the fastest method. Our method performs well on many classes, having state of the art results on 8 out of 10 and on average on YTOv1. The only class on which it has a significant lower score than the state of the art is motorbike, probably due to the fact that the objects are much smaller. On the last line we also present our results on YTOv2.2, which is the latest version of the dataset.

our method outperforms the others on 8 out of 10 classes and has the best overall average performance. Moreover, our CNN feed-forward net processes each image in 0.04 sec, being at least one to two orders of magnitude faster than all other methods (see Table 4). We also highlight that in all our comparisons, while our system is faster at test time, it takes much longer during its unsupervised training phase and requires large quantities of unsupervised training data.

5. Conclusions and Future Work

We have shown in extensive experiments that it is possible to use a relatively simple method for unsupervised object discovery in video to train a powerful deep neural network for detection and segmentation of objects in single images. The result is interesting and encouraging and shows how a system could learn, in an unsupervised fashion, general visual characteristics that predict well the presence and shape of objects in images. The network essentially discovers appearance object features from single images, at different levels of abstraction, that are strongly correlated with the spatiotemporal consistency of objects in video.

The student network, during the unsupervised training phase, is thus able to learn general "objectness" characteristics that are well beyond the capabilities of its teacher. These characteristics include good form, closure, smooth contours, as well as contrast with its background. What the simpler teacher discovers over time, the deep, complex student is able to learn across several layers of image features at different levels of abstraction. Therefore, our unsupervised learning model, tested in extensive experiments, brings a valuable contribution to the unsupervised learning problem in vision research.

Acknowledgements: This work was supported by UEFISCDI, under project PN-III-P4-ID-ERC-2016-0007.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, et al. Tensorflow: Large-scale machine learning on heterogeneous systems. *Software available from tensorflow.org*, 2015.

[2] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, 2010.

[3] O. Barnich and M. Van Droogenbroeck. Vibe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image processing*, 20(6):1709–1724, 2011.

[4] A. Borji, D. Sihite, and L. Itti. Salient object detection: A benchmark. In *ECCV*, 2012.

[5] X. Chen, A. Shrivastava, and A. Gupta. Enriching visual knowledge bases via object discovery and segmentation. In *CVPR*, 2014.

[6] M. Cheng, N. Mitra, X. Huang, P. Torr, and S. Hu. Global contrast based salient region detection. *PAMI*, 37(3), 2015.

[7] M. Cho, S. Kwak, C. Schmid, and J. Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *CVPR*, 2015.

[8] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts, and shadows in video streams. *PAMI*, 25(10), 2003.

[9] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 100(3), 2012.

[10] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007.

[11] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM computing surveys*, 31(3):264–323, 1999.

[12] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, 2013.

[13] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, 2010.

[14] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *CVPR*, 2012.

[15] A. Joulin, K. Tang, and L. Fei-Fei. Efficient image and video co-localization with Frank-Wolfe algorithm. In *ECCV*. 2014.

[16] Y. Jun Koh, W.-D. Jang, and C.-S. Kim. Pod: Discovering primary objects in videos based on evolutionary refinement of object recurrence, background, and primary object models. In *CVPR*, 2016.

[17] V. Kalogeiton, V. Ferrari, and C. Schmid. Analysing domain shift factors between videos and images for object detection. *PAMI*, 38(11), 2016.

[18] G. Kim, E. Xing, L. Fei-Fei, and T. Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, 2011.

[19] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] D. Kuettel, M. Guillaumin, and V. Ferrari. Segmentation propagation in imagenet. In *ECCV*, 2012.

[21] M. Leordeanu, R. Collins, and M. Hebert. Unsupervised learning of object features from video sequences. In *CVPR*, 2005.

[22] D. Li, W.-C. Hung, J.-B. Huang, S. Wang, N. Ahuja, and M.-H. Yang. Unsupervised visual representation learning by graph-based consistent constraints. In *ECCV*, 2016.

[23] D. Liu and T. Chen. A topic-motion model for unsupervised video object discovery. In *CVPR*, 2007.

[24] I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, 2016.

[25] M. Nguyen, L. Torresani, F. D. la Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *CVPR*, 2009.

[26] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, 2013.

[27] D. Parikh and T. Chen. Unsupervised identification of multiple objects of interest from multiple images: discover. In *Asian Conference on Computer Vision*, 2007.

[28] D. Pathak, R. Girshick, P. Dollar, T. Darrell, and B. Hariharan. Learning features by watching objects move. In *CVPR*, 2017.

[29] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár. Learning to refine object segments. In *ECCV*, 2016.

[30] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, pages 3282–3289. IEEE, 2012.

[31] F. Radenović, G. Tolias, and O. Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *ECCV*, 2016.

[32] T. Raiko, H. Valpola, and Y. LeCun. Deep learning made easier by linear transformations in perceptrons. In *AISTATS*, volume 22, pages 924–932, 2012.

[33] M. Rochan and Y. Wang. Efficient object localization and segmentation in weakly labeled videos. In *Advances in Visual Computing*, pages 172–181. Springer, 2014.

[34] I. Rock and S. Palmer. Gestalt psychology. *Sci Am*, 263:84–90, 1990.

[35] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics*, volume 23, pages 309–314, 2004.

[36] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*, 2013.

[37] J. Rubio, J. Serrat, and A. López. Video co-segmentation. In *ACCV*, 2012.

[38] O. Russakovsky et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3), 2015.

[39] P. Siva, C. Russell, T. Xiang, and L. Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *CVPR*, 2013.

[40] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering objects and their location in images. In *ICCV*, 2005.

[41] O. Stretcu and M. Leordeanu. Multiple frames matching for object discovery in video. In *BMVC*, 2015.

[42] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei. Co-localization in real-world images. In *CVPR*, 2014.

[43] S. Vicente, C. Rother, and V. Kolmogorov. Object cosegmentation. In *CVPR*, 2011.