Learning Instance Segmentation by Interaction

Deepak Pathak*, Yide Shentu*, Dian Chen*, Pulkit Agrawal*,
Trevor Darrell, Sergey Levine, Jitendra Malik
{pathak,fredshentu,dianchen,pulkitag,trevor,slevine,malik}@berkeley.edu

1. Introduction

Objects are a fundamental component of visual perception. How are humans able to effortlessly reorganize their visual observations into a discrete set of objects is a question that has puzzled researchers for centuries. The Gestalt school of thought put forth the proposition that humans use similarity in color, texture and motion to group pixels into individual objects [21]. Various methods for object segmentation based on color and texture cues have been proposed [3, 6, 7, 14, 16]. These approaches are, however, known to over-segment multi-colored and textured objects.

The state of the art overcomes these issues by making use of detailed class-specific segmentation annotations for a large number of objects in a large dataset of images to get impressive results on segmenting objects in web images [10, 12, 19]. A typical system in this paradigm first makes use of 1M human annotated images in ImageNet [20] to pretrain a deep neural network. This network is then finetuned using over 700K object instances belonging to eighty semantic classes from the COCO dataset. Such data is laborious and extremely time consuming to collect.

In this work, we take the first step towards putting this developmental hypothesis to test and investigate if it is possible for an active agent to learn class agnostic instance segmentation of objects by starting off with two axioms: (a) there are objects in the world; (b) principle of common fate [21], i.e. pixels that move together, group together. To that end, we set up an agent, shown in Figure 1, to interact with its environment and record the resulting RGB images. The agent maintains a belief about how images can be decomposed into objects, and actively tests its belief by attempting to grasp potential objects in the world. Through such self-supervised interaction, we show that it is possible to learn to segment novel objects kept on textured backgrounds into individual instances. We will publicly release the collected data (i.e. over 50K interactions recorded from four views) along with a set of 1700 human labelled images containing 9.3K object segments to serve as a benchmark for evaluating self-supervised, weakly supervised or unsupervised class agnostic instance segmentation method.

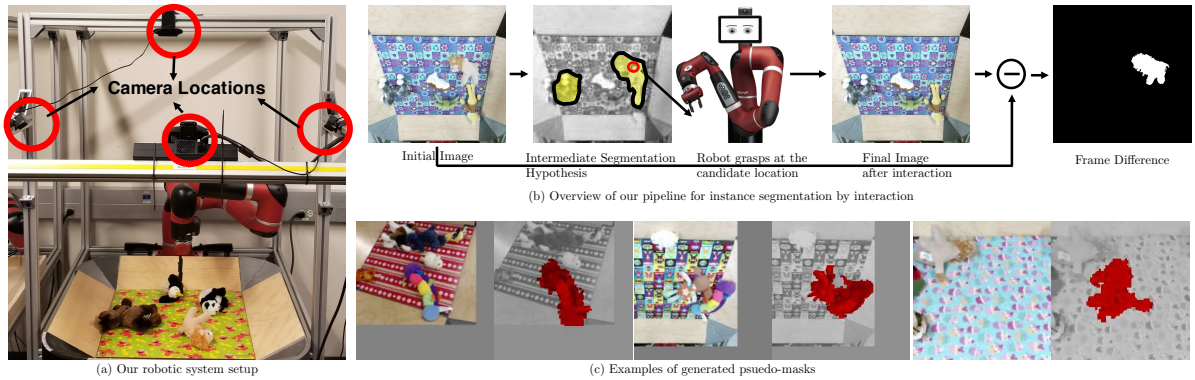
An interesting technical challenge we encountered is that the object masks discovered through interaction are noisy compared to the object masks marked by human annotators. For instance, sometimes the agent may mistakenly think of two nearby objects as one object. This requires the training to be robust to label noise, analogous to how in regression, we need to be robust to outliers in the data. We developed a technique, "robust set loss", to handle this during the training, with the general idea being that the segmenter is not required to predict exactly the pixels in the candidate object mask, rather that the predicted pixels as a set have a good Jaccard index overlap with the candidate mask.

Related Work In the past, active perception [2, 4, 5, 9] has garnered a lot of interest. However, much of this work is concerned with using interaction to segment a specific scene [8, 11, 15]. In contrast, our system uses interactions to actively gather supervision to train a segmentation system that can be used to segment objects in new images. Instead of using interaction with the environment, the work of [17] used optical flow to generate pseudo ground truth masks from passively observed videos. As far as we are aware, ours is the first work that aims to learn to segment objects using self-supervision from active robotic interaction.

2. Experimental Setup

The basic interaction primitive used by the robot allows it to attempt to pick and place objects in the scene. We set up the robot to interact autonomously with its environment without human supervision. Overall, the robot performed more than 50,000 interactions with its environment. Approximately the first 15,000 interactions were recorded using the main camera and the remainder of interactions were recorded using auxiliary four cameras. Data recording from

*Denotes equal contribution. This is workshop version of the full paper which is available at https://pathak22.github.io



(a) Our robotic system setup

(b) Overview of our pipeline for instance segmentation by interaction

(c) Examples of generated psuedo-masks

Figure 1. (a): Overview of our approach: a robot conducts experiments in its environment to learn a model for segmenting its visual observation into individual object instances. Our robot maintains a belief about what groups of pixels might constitute an object and actively tests its belief by attempting to grasp this set of pixels (for e.g. attempts a grasp at the location shown by the yellow circle). Interaction with objects causes motion, whereas interaction with background results in no motion. This motion cue is utilized by the robot to train a deep neural network for segmenting objects. (b),(c): Visualization of the set of thirty six objects used for training (b) and sixteen objects used for testing (c). Validation objects can be seen in supp. materials. Separate sets of backgrounds were used for training, validation and testing.

four cameras was used to increase invariance to viewpoint.

Datasets: We use 24 backgrounds for training, 6 for validation and 10 for testing. We use 36 different objects for training, 8 for validation and 15 for testing. The validation set consisted of 30 images (5 images per background) and the test set consisted of 50 images (5 images per background). We manually annotated object masks in these images for the purpose of evaluation; no labels for training.

3. Instance Segmentation by Interaction

The training procedure is summarized in Algorithm 1. The major challenge in training a model with such self-generated masks is that they are far from perfect (Figure 1). Typical error modes include: (a) false negatives due to complete failure to grasp an object; (b) failure in grasping that slightly perturb the object resulting in incomplete masks; (c) in case two objects are located near each other, picking one object moves the other one, resulting in masks that span multiple objects; (d) erroneous masks due to variation in lighting, shadows and other nuisance factors. Any method attempting to learn object segmentation from interaction must deal with such imperfections in the self-generated *pseudo ground truth* masks.

Algorithm 1: Segmentation by Interaction

Pre-train network with passive unsupervised data
for *iteration* $t = 1$ *to* T **do**
  Record current observation I_t
  Generate object hypothesis: $\{ s_1^t, \ldots s_K^t \} \Leftarrow \text{CNN}(I_t)$
  Randomly choose one hypothesis $s_j^t \in \{ s_1^t, \ldots s_K^t \}$
  Interact with hypothesized object $(move(s_j^t))$
  Record observation I_{t+1}
  mask $\Leftarrow$ frame_difference(I_t, I_{t+1})
  if *mask is empty* **then**
    {(x,y), mask, I^t} is negative training example
  else
    {(x,y), mask, I^t} is positive training example
  end
  if t % *update_interval* $== 0$ **then**
    Update CNN using positive/negative examples
  end
end

3.1. Robust Set Loss

Attempting to exactly fit the noisy masks is adversarial for the learning process, as (a) overfitting to noise would hamper the ability to generalize to unseen examples, and (b) inability to fit noise would increase variance in the gradients and thereby make training unstable.

The principled approach of learning with noisy training data is to use a robust loss for mitigating the effect of outliers. Robust loss functions have been extensively studied in statistics, in particular, Huber loss [13] applied to regression problems. However, such ideas have mostly been explored in the context of regression and classification for modeling independent outputs. Unfortunately, segmentation mask is a "set of pixels", where a statistic of interest such as the similarity between two sets of pixels (e.g., ground-truth and predicted masks) measured for instance using Jaccard Index (i.e., intersection over union (IOU)) depends on all the pixels. The dependence of the statistic on a set of pixels makes it non-trivial to generalize ideas such as Huber loss in a straightforward manner. We formulate Robust Set Loss

(a) Performance vs. Interactions (b) Successes vs. Interactions (c) Precision vs. Recall

Figure 2. Quantitative evaluation of the segmentation model on the held-out test. (a) The performance of our system measured as mAP at IoU of 0.3 steadily increases with the amount of data. After 50K iterations our system significantly beats GOP tuned with domain knowledge (i.e. GOP-Tuned; section **??**). (b) The efficacy of experimentation performed by the robot is computed as the recall of ground truth objects that have IoU of more than 0.3 with the group of pixels that the robot believes to be objects. The steady increase in recall at different precision threshold shows that the robot learns to perform more efficient experiments with time. (c) Precision-Recall curves re-confirm the results.

(RSL) to deal with "set-level" noise. The key insight is to impose a soft constraint for only matching a subset of target pixels while ensuring that (potentially non-differentiable) some metric of interest, such as IOU, between the prediction and the noisy target is greater than or equal to a certain threshold. We generalize the CCNN constrained formulation proposed in Pathak et. al. [18] to achieve this loss. Please refer to full paper for the details of RSL [1].

3.2. Bootstrapping via Passive Self-Supervision

Without any prior knowledge, the agent's initial beliefs about objects will be arbitrary, causing it to spend most of its time interacting with the background. This process would be very inefficient. We address this issue by assuming that initially our agent can passively observe objects moving in its environment. For this purpose we use a prior robotic pushing dataset [1] that was constructed by a robot randomly pushing objects in a tabletop environment. We apply the method of [17] to automatically extract masks from this data, which we use to pre-train our ResNet-18 network (initialized with random-weights). Note that this method of pre-training is completely self-supervised.

4. Results and Evaluations

We compare the performance of our method against a state-of-the-art bottom up segmentation method called Geodesic Object Proposals (GOP [16]), and a top-down instance segmentation method trained in a class agnostic manner using over 700K strongly supervised masks obtained from the COCO dataset (DeepMask [19]) and pre-trained on 1M ImageNet, using the AP at IOU 0.3 metric on the held-out testing set as shown in Figure 2(a). Our system significantly outperforms vanilla GOP and GOP with domain knowledge (tuned). These results are re-confirmed by the precision-recall curves shown in Figure 2(c). The performance of our system steadily increases with the amount of data and from the performance curve (see Figure 2(a)). Our method performs similar to vanilla DeepMask, but worse than the one tuned to our domain for scaling and position of objects. This result is significant because DeepMask was trained with perfect ground truth segmentation masks for 700K COCO objects after being pre-trained to classify 1M imagenet images, whereas our system was trained using imperfect masks (section 3) using only 50K self-supervised active interactions after pre-training with approximately 60K passive observations of moving objects. AP evaluation at IOU 0.5 (see Table 1) reveals that while our method significantly outperforms GOP, it is outperformed by DeepMask. We believe the main reason is that the masks obtained by robot interaction are imperfect. Refer to full paper more results and generalization[1].

Method	Property	AP at IU 0.3	AP at IU 0.5
GOP	Bottom up	07.4	05.6
GOP (**tuned**)	Bottom up	24.3	20.7
DeepMask	Strong Sup.	44.5	34.3
DeepMask (**tuned**)	Strong Sup.	61.8	47.3
Ours	Self-sup.	44.1	20.2
Ours+Human	Semi-sup.	47.0	25.1
Ours+Robust Set Loss	Self-sup.	48.7	24.6

Table 1. Quantitative comparison of our method with bottom-up (GOP [16]), learned top-down (DeepMask [19]) segmentation methods and optimization without robust set loss on the full test set. Note that our approach significantly outperforms GOP, but is outperformed by DeepMask that uses strong manual supervision of 700K+ COCO segments and 1M ImageNet images. Adding approx. 1500 images with clean segmentation masks improves performance of our base system.

[1]Full paper available at https://pathak22.github.io

References

[1] P. Agrawal, A. Nair, P. Abbeel, J. Malik, and S. Levine. Learning to poke by poking: Experiential learning of intuitive physics. *NIPS*, 2016. 3

[2] J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International journal of computer vision*, 1(4):333–356, 1988. 1

[3] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, 2014. 1

[4] R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988. 1

[5] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos. Revisiting active perception. *Autonomous Robots*, pages 1–20, 2016. 1

[6] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *PAMI*, 2012. 1

[7] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004. 1

[8] P. Fitzpatrick. First contact: an active vision approach to segmentation. In *IROS*, 2003. 1

[9] J. J. Gibson. *The ecological approach to visual perception*. Psychology Press, 1979. 1

[10] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, 2014. 1

[11] K. Hausman, D. Pangercic, Z.-C. Márton, F. Bálint-Benczédi, C. Bersch, M. Gupta, G. Sukhatme, and M. Beetz. Interactive segmentation of textured and textureless objects. In *Handling Uncertainty and Networked Structure in Robot Control*. Springer, 2015. 1

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. *arXiv preprint arXiv:1703.06870*, 2017. 1

[13] P. J. Huber. Robust estimation of a location parameter. *The annals of mathematical statistics*, 1964. 2

[14] P. Isola, D. Zoran, D. Krishnan, and E. H. Adelson. Crisp boundary detection using pointwise mutual information. In *European Conference on Computer Vision*, pages 799–814. Springer, 2014. 1

[15] J. Kenney, T. Buckley, and O. Brock. Interactive segmentation for manipulation in unstructured environments. In *ICRA*, 2009. 1

[16] P. Krähenbühl and V. Koltun. Geodesic object proposals. In *ECCV*, 2014. 1, 3

[17] D. Pathak, R. Girshick, P. Dollár, T. Darrell, and B. Hariharan. Learning features by watching objects move. *CVPR*, 2017. 1, 3

[18] D. Pathak, P. Krähenbühl, and T. Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015. 3

[19] P. O. Pinheiro, R. Collobert, and P. Dollár. Learning to segment object candidates. In *Advances in Neural Information Processing Systems*, pages 1990–1998, 2015. 1, 3

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 1

[21] M. Wertheimer. Laws of organization in perceptual forms. 1938. 1