

# One-Shot Learning for Semantic Segmentation

Amirreza Shaban  
amirreza@gatech.edu

Shray Bansal  
sbansal34@gatech.edu

Zhen Liu  
liuzhen1994@gatech.edu

Irfan Essa  
irfan@gatech.edu

Byron Boots  
bboots@cc.gatech.edu

College of Computing,  
Georgia Institute of Technology,  
Georgia, USA

---

## Abstract

Low-shot learning methods for image classification support learning from sparse data. We extend these techniques to support dense semantic image segmentation. Specifically, we train a network that, given a small set of annotated images, produces parameters for a Fully Convolutional Network (FCN). We use this FCN to perform dense pixel-level prediction on a test image for the new semantic class. Our architecture shows a 25% relative meanIoU improvement compared to the best baseline methods for one-shot segmentation on unseen classes in the PASCAL VOC 2012 dataset and is at least 3× faster.

## 1 Introduction

Deep Neural Networks are powerful at solving classification problems in computer vision. However, learning classifiers with these models requires a large amount of labeled training data, and recent approaches have struggled to adapt to new classes in a data-efficient manner. There is interest in quickly learning new concepts from limited data using one-shot learning methods [21, 37]. One-shot image classification is the problem of classifying images given only a single training example for each category [22, 39].

We propose to undertake *One-Shot Semantic Image Segmentation*. Our goal is to predict a pixel-level segmentation mask for a semantic class (like horse, bus, *etc.*) given only a single image and its corresponding pixel-level annotation. We refer to the image-label pair for the new class as the support set here, but more generally for  $k$ -shot learning, support set refers to the  $k$  images and labels.

A simple approach to performing one-shot semantic image segmentation is to fine-tune a pre-trained segmentation network on the labeled image [3]. This approach is prone to overfitting due to the millions of parameters being updated. It also introduces complications in optimization, where parameters like step size, momentum, number of iterations, *etc.* may be

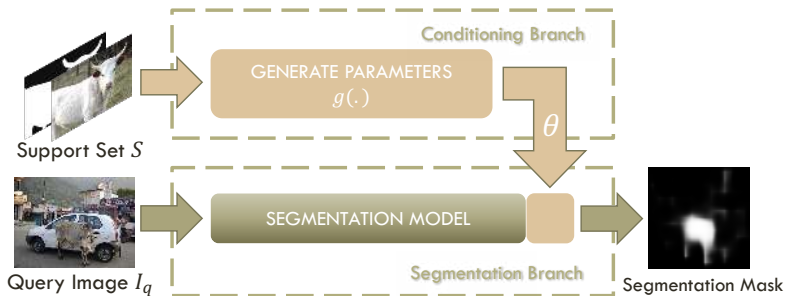


Figure 1: Overview.  $S$  is an annotated image from a new semantic class. In our approach, we input  $S$  to a function  $g$  that outputs a set of parameters  $\theta$ . We use  $\theta$  to parameterize part of a learned segmentation model which produces a segmentation mask given  $I_q$ .

difficult to determine. Recent one-shot image categorization methods [22, 39] in contrast, meta-learn a classifier that, when conditioned on a few training examples, can perform well on new classes. Since Fully Convolutional Neural Networks (FCNs) [26] perform segmentation as pixel-wise classification, we could extend these one-shot methods directly to classify at the pixel level. However, thousands of dense features are computed from a single image and one-shot methods do not scale well to this many features. We illustrate this issue by implementing an extension to the Siamese Network from [22] as a baseline in Section 6.

We take inspiration from few-shot learning and propose a novel two-branched approach to one-shot semantic image segmentation. The first branch takes the labeled image as input and produces a vector of parameters as output. The second branch takes these parameters as well as a new image as input and produces a segmentation mask of the image for the new class as output. This is illustrated in Figure 1. Unlike the fine tuning approach to one-shot learning, which may require many iterations of SGD to learn parameters for the segmentation network, the first branch of our network computes parameters in a single forward pass. This has several advantages: the single forward pass makes our method fast; our approach for one-shot learning is fully differentiable, allowing the branch to be jointly trained with the segmentation branch of our network; finally, the number of parameters  $\theta$  is independent of the size of the image, so our method does not have problems in scaling.

To measure the performance for one-shot semantic segmentation we define a new benchmark on the PASCAL VOC 2012 dataset [11] (Section 5). The training set contains labeled images from a subset of the PASCAL classes and the testing set has annotations of classes that were not present in training. We show significant improvements over the baselines on this benchmark in terms of the standard meanIoU (mean Intersection over Union) metric as described in Section 7.

We extend to  $k$ -shot learning by applying our one-shot approach for each of the  $k$  images independently to produce  $k$  segmentation masks. We then aggregate these masks by performing a logical-OR operation at the pixel level. This approach, apart from being easy to implement and fast, requires no retraining to generalize to any number of images in the support set. We show its effectiveness in terms of increasing meanIOU accuracy per added image to the support set in section 7.

PASCAL VOC contains only 20 classes, which is small when compared to standard

datasets used for training one-shot classification methods like Omniglot (1623) [24] and ImageNet (1000) ([10]). Simulating the one-shot task during training, even with such a limited number of classes performs well. This is in contrast to the common notion that training models for few-shot learning requires a large number of classes. We hypothesize that part of our algorithm’s ability to generalize well to unseen classes comes from the pre-training performed on ImageNet, which contains weak image-level annotations for a large number of classes. We perform experiments on the pretraining in section 7.1.

This paper makes the following contributions: (1) we propose a novel technique for one-shot segmentation which outperforms baselines while remaining significantly faster; (2) we show that our technique can do this without weak labels for the new classes; (3) we show that meta-learning can be effectively performed even with only a few classes having strong annotations available; and (4) we set up a benchmark for the challenging  $k$ -shot semantic segmentation task on PASCAL.

## 2 Related Work

**Semantic Image Segmentation** is the task of classifying every pixel in an image into a predefined set of categories. Convolutional Neural Network (CNN) based methods have driven recent success in the field. Some of these classify super-pixels [13, 15, 27], others classify pixels directly [6, 16, 26, 28]. We base our approach on the Fully Convolutional Network (FCN) for Semantic Segmentation [26] which showed the efficiency of pixel-wise classification. However, unlike FCN and the other approaches above, we do not assume a large set of annotated training data for the test classes.

**Weak Supervision.** Weak and semi-supervised methods for Semantic Segmentation reduce the requirement on expensive pixel-level annotations, thus attracting recent interest. Weak supervision refers to training from coarse annotations like bounding boxes [9] or image labels [30, 31, 33]. A notable example is co-segmentation, where the goal is to find and segment co-occurring objects in images from the same semantic class [12, 35]. Many co-segmentation algorithms [8, 17, 34] assume object visual appearances in a batch are similar and either rely on hand-tuned low-level features or high-level CNN features trained for different tasks or objects [34]. In contrast, we meta-learn a network to produce a high-level representation of a new semantic class given a single labeled example. Semi-supervised approaches [18, 19, 30] combine weak labels with a small set of pixel-level annotations. However, they assume a large set of weak labels for each of the desired objects. For instance, Pathak *et al.* [32] use image-level annotations for all classes and images in the PASCAL 2012 training set [11], while we exclude all annotations of the testing classes from the PASCAL training set.

**Few-Shot Learning** algorithms seek to generalize knowledge acquired through classes seen during training to new classes with only a few training examples [25, 36, 39]. Discriminative methods in which the parameters of the base classifier (learned on training classes) are adapted to the new class [1, 2, 14, 40] are closely related to our work. The main challenge is that the adapted classifier is prone to over-fit to the newly presented training examples. Wang and Herbert [40] address this challenge by learning to predict classifiers which remain close to the base classifier. Bertinetto *et al.* [2] trained a two-branch network, in which one branch receives an example and predicts a set of dynamic parameters. The second branch classifies the query image using the dynamic parameters along with a set of learned static parameters. A similar approach was used by Noh *et al.* in [29] for question answering. We

draw several ideas from these papers and adapt them for the task of dense classification to design our model. Metric learning is another approach to low-shot learning [22, 39]. It aims to learn an embedding space that pulls objects from the same categories close, while pushing those from different categories apart. Koch *et al.* [22] show that a Siamese architecture trained for a binary verification task can beat several classification baselines in  $k$ -shot image classification. We adapt their approach for image segmentation as one of our baselines.

### 3 Problem Setup

Let the support set  $S = \{(I_s^i, Y_s^i(l))\}_{i=1}^k$  be a small set of  $k$  image-binary mask pairs where  $Y_s^i \in L_{test}^{H \times W}$  is the segmentation annotation for image  $I_s^i$  and  $Y_s^i(l)$  is the mask of the  $i^{th}$  image for the semantic class  $l \in L_{test}$ . The goal is to learn a model  $f(I_q, S)$  that, when given a support set  $S$  and query image  $I_q$ , predicts a binary mask  $\hat{M}_q$  for the semantic class  $l$ . An illustration of the problem for  $k = 1$  is given Figure 1.

During training, the algorithm has access to a large set of image-mask pairs  $D = \{(I^j, Y^j)\}_{j=1}^N$  where  $Y^j \in L_{train}^{H \times W}$  is the semantic segmentation mask for training image  $I^j$ . At testing, the query images are only annotated for new semantic classes i.e.  $L_{train} \cap L_{test} = \emptyset$ . This is the key difference from typical image segmentation where training and testing classes are the same. While the problem is similar to  $k$ -shot learning, which has been extensively studied for image classification [36, 39], applying it to segmentation requires some modification.

In this problem, unlike image classification, examples from  $L_{test}$  might appear in training images. This is handled naturally when an annotator unaware of some object class, labels it as background. Annotations of  $L_{test}$  objects are excluded from the training set, while the images are included as long as there is an object from  $L_{train}$  present. State-of-the-art algorithms for image segmentation [4, 5] use networks pre-trained on large-scale image classification datasets like [10]. Although these weights give the models a better starting point, they still require many segmented images and thousands of weight updates to learn a good model for pixel classification. This is true even for the classes that directly overlap. We allow similar access to weak annotations for our problem by initializing VGG with weights pre-trained on ImageNet [10]. In section 7.1 however, we show that even excluding all the overlapping classes from pre-training does not degrade the performance of our approach.

### 4 Proposed Method

We propose an approach where the first branch receives as input a labeled image from the support set  $S$  and the second branch receives the query image  $I_q$ . In the first branch, we input the image-label pair  $S = (I_s, Y_s(l))$  to produce a set of parameters,

$$w, b = g_\eta(S). \quad (1)$$

In the other branch, we extract a dense feature volume from  $I_q$  using a parametric embedding function  $\phi$ . Let  $F_q = \phi_\zeta(I_q)$  be that feature volume extracted from  $I_q$ , then  $F_q^{mn}$  is the feature vector at the spatial location  $(m, n)$ . Pixel level logistic regression is then performed on the features using the parameters from the first layer to get the final mask,

$$\hat{M}_q^{mn} = \sigma(w^\top F_q^{mn} + b). \quad (2)$$

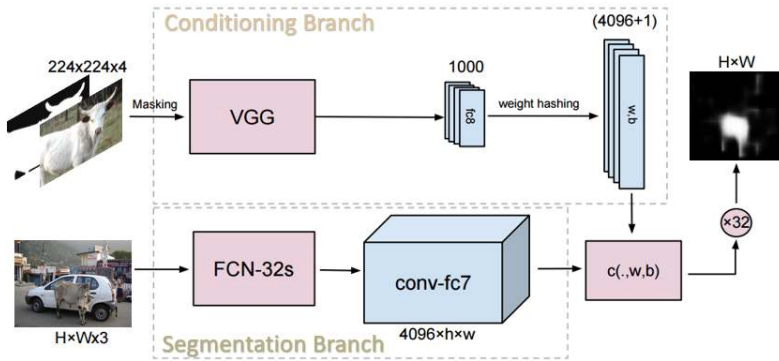


Figure 2: Model Architecture. The conditioning branch receives an image-label pair and produces a set of parameters  $\{w, b\}$  for the logistic regression layer  $c(\cdot, w, b)$ . The segmentation branch is an FCN that receives a query image as input and outputs strided features of conv-fc7. The predicted mask is generated by classifying the pixel-level features through  $c(\cdot, w, b)$ , which is then upsampled to the original size.

Here,  $\sigma(\cdot)$  is the sigmoid function and  $\hat{M}_q^{mn}$  is the  $(m, n)$  location of the predicted mask for the query. This can be understood as a convolutional layer with parameters  $\{w, b\}$  followed by a sigmoid activation function, where the parameters are not fixed after training and get computed through the first branch for each image in the support set. The predicted mask is then upsampled back to the original image size using standard bilinear interpolation. The final binary mask is produced by using a threshold of 0.5 on  $\hat{M}_q$ . The overall architecture is illustrated in Figure 2. We explain each part of the architecture in more detail in the following subsections.

## 4.1 Producing Parameters from Labeled Image

We modify the VGG-16 architecture from [38] to model the function  $g_\eta(\cdot)$ .

**Masking.** We chose to mask the image with its corresponding label so it contains only the target object instead of modifying the first layer to receive the four channel image-mask pair as input. We do this for the following two empirical reasons. (1) Even in the presence of the mask the network response tends to be biased towards the largest object in the image which may not be the object we would like to segment. (2) Including the background information in the input increased the variance of the output parameters  $\{w, b\}$  which prevented the network from converging.

**Weight Hashing.** Inspired by Noh *et al.* [29], we employed the weight hashing layer from [7] to map the 1000-dimensional vector output from the last layer of VGG to the 4097 dimensions of  $\{w, b\}$ . This mapping avoids the overfitting which would occur due to the massive number of extra parameters that a fully connected layer will introduce if used instead. We implemented it efficiently as a fully connected layer with fixed weights. This is explained in more detail in the supplementary material.

## 4.2 Dense Feature Extraction

We model the embedding function  $F_q = \phi_\zeta(I_q)$  by the FCN-32s fully convolutional architecture [26] excluding the final prediction layer. The 4096 channel feature volume at conv-fc7 is then fed to the logistic pixel classifier described above. In section 7 we also evaluate performance of the high resolution dilated-FCN [41] with stride 8.

## 4.3 Training Procedure

We simulate the one shot task during training by sampling a support set  $S$ , a query image  $I_q$  and its corresponding binary mask  $M_q$  from the training set  $D_{train}$  at each iteration. First, an image-label pair  $(I_q, Y_q)$  is sampled uniformly at random from  $D_{train}$ , then we sample a class  $l \in L_{train}$  uniformly from the classes present in the semantic mask and use it to produce the binary mask  $Y_q(l)$ .  $S$  is formed by picking one image-mask pair at random from  $D_{train} - \{(I_q, Y_q)\}$  with class  $l$  present. We can then predict the mask  $\hat{M}_q$  with a forward pass through our network. We maximize the log likelihood of the ground-truth mask

$$\mathcal{L}(\eta, \zeta) = \mathbb{E}_{S, I_q, M_q \sim D_{train}} \left[ \sum_{m, n} \log p_{\eta, \zeta}(M_q^{mn} | I_q, S) \right]. \quad (3)$$

Here  $\eta$  and  $\zeta$  are the network parameters,  $p_{\eta, \zeta}$  is the probability of the mask given the neural network output, and  $S$ ,  $I_q$ , and  $M_q$  are sampled by the sampling strategy described above. We use Stochastic Gradient Descent with a fixed learning rate of  $10^{-10}$ , momentum 0.99 and batch size of 1. The VGG network overfits faster than the fully-convolutional branch; therefore, we set the learning rate multiplier to 0.1 for learning the parameter  $\eta$ . We stop training after 60k iterations.

## 4.4 Extension to $k$ -shot

In the case of  $k$ -shot segmentation the support set contains  $k$  labeled images,  $S = \{I_s^i, Y_s^i(l)\}_{i=1}^k$ . We use these images to produce  $k$  sets of the parameters  $\{w^i, b^i\}_{i=1}^k$ . Each of them can be understood to be an independent classifier of an ensemble. These classifiers we noticed have high precision but low recall. We believe this is because each is produced by one example from the support set and a single image can only contain a small subset of the possible appearances of the object. So, we combine the decision of these classifiers by including a pixel in the final mask if it was considered an object by any of these classifiers. This is implemented as a logical OR operation between the  $k$  binary masks. This approach has the benefit that it does not require any retraining and can be generalized to any  $k$ . It is also much faster than the baselines as shown in section 7.

## 5 Dataset and Metric

**Dataset:** We create a new dataset, **PASCAL-5<sup>i</sup>**, for the problem of  $k$ -shot Image Segmentation using images and annotations from PASCALVOC 2012 [11] and extended annotations from SDS<sup>1</sup> [15]. From  $L$ , the set of twenty semantic classes in PASCALVOC, we sample five and consider them as the test label-set  $L_{test} = \{4i + 1, \dots, 4i + 5\}$ , with  $i$  being the fold

<sup>1</sup>For creating the training set, we only include images that do not overlap with the PASCALVOC 2012 validation set.

$i = 0$	$i = 1$	$i = 2$	$i = 3$		
aeroplane, bicycle, bird, boat, bottle	bus, car, cat, chair, cow	diningtable, dog, horse, motorbike, person	potted plant, sheep, sofa, train, tv/monitor		
Methods (1-shot)	PASCAL-5 <sup>0</sup>	PASCAL-5 <sup>1</sup>	PASCAL-5 <sup>2</sup>	PASCAL-5 <sup>3</sup>	Mean
1-NN	25.3	44.9	<b>41.7</b>	18.4	32.6
LogReg	26.9	42.9	37.1	18.4	31.4
Finetuning	24.9	38.8	36.5	30.1	32.6
Siamese	28.1	39.9	31.8	25.8	31.4
Ours	<b>33.6</b>	<b>55.3</b>	40.9	<b>33.5</b>	<b>40.8</b>
Methods (5-shot)	PASCAL-5 <sup>0</sup>	PASCAL-5 <sup>1</sup>	PASCAL-5 <sup>2</sup>	PASCAL-5 <sup>3</sup>	Mean
Co-segmentation	25.1	28.9	27.7	26.3	27.1
1-NN	34.5	53.0	<b>46.9</b>	25.6	40.0
LogReg	<b>35.9</b>	51.6	44.5	25.6	39.3
Ours	<b>35.9</b>	<b>58.1</b>	42.7	<b>39.1</b>	<b>43.9</b>

Table 1: Mean IoU results on PASCAL-5<sup>i</sup>. **Top**: test classes for each fold of PASCAL-5<sup>i</sup>. The **middle** and **bottom** tables contain the semantic segmentation meanIoU on all folds for the 1-shot and 5-shot tasks respectively.

number, and the remaining fifteen forming the training label-set  $L_{train}$ . Test and training class names are shown in Table 5. We form the training set  $D_{train}$  by including all image-mask pairs from PASCALVOC and SDS training sets that contain at least one pixel in the semantic mask from the label-set  $L_{train}$ . The masks in  $D_{train}$  are modified so that any pixel with a semantic class  $\neq L_{train}$  is set as the background class  $l_{\emptyset}$ . We follow a similar procedure to form the test set  $D_{test}$ , but here the image-label pairs are taken from PASCALVOC validation set and the corresponding label-set is  $L_{test}$ . Thus, apart from a few exclusions, the set of images is similar to those used in Image Segmentation papers, like FCN [26]. However, the annotations are different. Given the test set  $D_{test}$ , we use the same procedure that is described in Section 4.3 to sample each test example  $\{S, (I_q, Y_q(l))\}$ . We sample  $N = 1000$  examples and use it as the benchmark for testing each of the models described in the next section.

**Metric:** Given a set of predicted binary segmentation masks  $\{\hat{M}_q\}_{i=1}^N$  and the ground truth annotated mask  $\{M_q\}_{i=1}^N$  for a semantic class  $l$  we define the per-class Intersection over Union ( $IoU_l$ ) as  $\frac{tp_l}{tp_l + fp_l + fn_l}$ . Here,  $tp_l$  is the number of true positives,  $fp_l$  is the number of false positives and  $fn_l$  is the number of false negatives over the set of masks. The **meanIoU** is just its average over the set of classes, i.e.  $(1/n_l) \sum_l IoU_l$ . This is the standard metric of meanIoU defined in Image Segmentation literature adapted for our binary classification problem.

## 6 Baselines

We evaluate the performance of our method with different baselines. Since one-shot image segmentation is a new problem, we adapt previous work for dense pixel prediction to serve as baselines to compare against.

- **Base Classifiers:** CNNs learn deep representations of images, so these models are an intuitive starting point for classification. Specifically, we first fine-tune FCN-32s pretrained on ILSVRC2014 data to perform 16-way (15 training foreground classes + 1 background class) pixel-wise predictions on the **PASCAL-5<sup>i</sup>** dataset. During testing, we extract dense pixel-level features from both images in the support set and the query image. We then train classifiers to map dense fc-7 features from the support set to their

corresponding labels and use it to generate the predicted mask  $\hat{M}_q$ . We experimented with various classifiers including 1-NN and logistic regression<sup>2</sup>

- **Fine-tuning:** As suggested by [3], for each test iteration we fine-tune the trained segmentation network on examples in the support set and test on the query image. We only fine-tune the fully connected layers (fc6, fc7, fc8) to avoid overfitting and reducing the inference time per query. We also found that the fine-tuned network converges faster if we normalize the fc-7 features by a batch normalization layer.
- **Co-segmentation by Composition:** To compare with these techniques, we include the results of the publicly available implementation<sup>3</sup> of [12] on PASCAL-5<sup>i</sup>.
- **Siamese Network for One-shot Dense Matching:** Siamese Networks trained for image verification, i.e. predicting whether two inputs belong to the same class, have shown good performance on one-shot image classification [22]. We adapt them by using two FCNs to extract dense features and then train it for pixel verification. A similarity metric from each pixel in the query image to every pixel in the support set is also learned and pixels are then labeled according to their nearest neighbors in the support set. Implementation details are provided in the supplementary document.

## 7 Experiments

We conduct several experiments to evaluate the performance of our approach on the task of  $k$ -shot Image segmentation by comparing it to other methods. Table 1 reports the performance of our method in 1-shot and 5-shot settings and compares them with the baseline methods. To fit a 5-shot Siamese network into memory we sampled from features in the support set with a rate of 0.3. However, sub-sampling considerably degraded the performance of the method and 5-shot results were worse than the 1-shot version so we exclude those results.

Our method shows better generalization performance to new classes. The difference is very noticeable in 1-shot learning as other methods overfit to only the image in the support set. Specifically, our method outperforms 1-NN and fine-tuning in one-shot image segmentation by 25% relative meanIoU. We also provide some qualitative result from our method in Figure 4. Surprisingly, the results for 1-NN are almost as good as the fine-tuning baseline, which overfits quickly to the data in the support set.

In Table 1, we also compare Co-segmentation by Composition [35] for 5-shot segmentation to our approach. As expected, using the strong pixel-level annotations enables our method to outperform the unsupervised co-segmentation approach, by 16%. In fact, we can outperform co-segmentation results that require 5 weakly annotated images with just a single strongly annotated image.

**Dilated-FCN:** In addition to the low-resolution version of our method, we also trained the dilated-FCN with higher resolution on PASCAL-5<sup>0</sup> and achieved 37.0% and 37.43% meanIoU for 1-shot and 5-shot respectively. We notice a 3.4% improvement over low-resolution for one-shot, however, the gap between 1-shot and 5-shot is small at this resolution. We believe this is due to our training being specific to the 1-shot problem. We do not use dilated-FCN architecture for other methods due to the impracticality caused by their high computational cost or memory footprint.

<sup>2</sup>We also trained linear SVM, but could not get a comparable result to logistic regression.

<sup>3</sup><http://www.wisdom.weizmann.ac.il/~vision/CoSegmentationByComposition.html>



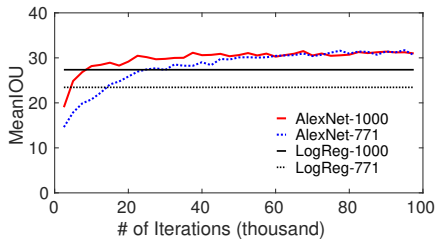


Figure 3: Pretraining Effect on AlexNet.

Methods	1-shot	5-shot
1-NN	1.10	4.55
Logistic Reg	0.66	3.50
Finetune	5.56	-
Siamese	5.65	-
Ours-32s	0.19	0.21

Table 2: Inference Time (in s).

**Running Time:** In Table 2 we include the running time of each algorithm. All the experiments were executed on a machine with a 4GHz Intel Core-i7 CPU, 32GB RAM, and a Titan X GPU. In one-shot setting our method is  $\sim 3\times$  faster than the second fastest method logistic regression. For 5-shot our method is  $\sim 10\times$  faster than logistic regression.

## 7.1 Pretraining Effect

The models compared above have two sources of information, the image-level labels for the classes in ImageNet [10] through the pretraining and the pixel-level annotation of classes in  $L_{train}$ . Although the test classes  $L_{test}$  do not overlap with  $L_{train}$ , they have partial overlap with some ImageNet classes. To understand this effect, we use a dataset which excludes all the classes in ImageNet with any overlap with PASCAL categories called PASCAL-removed-ImageNet as in [20]. This dataset contains only 771 classes as compared to 1000 originally since each class in PASCAL usually overlaps with multiple ImageNet classes. We use AlexNet [23] trained on ImageNet and PASCAL-removed-ImageNet (from Huh *et al.* [20]) with the suffixes 1000 and 771 respectively. We replaced the VGG and FCN from both branches of our approach with AlexNet to give us AlexNet-1000 and AlexNet-771. We also have a baseline in the form of Logistic Regression performed on convolutional AlexNet features finetuned on PASCAL, similar to the Base Classifiers described in section 6. We refer to these as LogReg-1000 and LogReg-771. Figure 3 contains the results for these models on the first fold, i.e. PASCAL-5<sup>0</sup>. Note that the results for the two baselines are constant because we evaluate the networks only once they converge.

In Figure 3 we observe that AlexNet-1000 is better initially and shows faster convergence. However, after convergence AlexNet-771 performs on par with AlexNet-1000. The initial gap could be understood by the fact that even the  $L_{train}$  classes were not presented during the pre-training. AlexNet being a simpler model performs worse than VGG, meanIOU was 33.6% in Table 5. However, AlexNet-771 outperforms even our best VGG baseline, which was Siamese at 28.1% for PASCAL-5<sup>0</sup>. This result shows that we can generalize to new categories without any weak supervision for them. In contrast, LogReg-1000 outperforming LogReg-771 shows its incapacity to learn a good representation without seeing weak labels for test categories. This highlights the importance of meta-learning for this task.

## 8 Conclusion

Deep learning approaches have achieved top performance in many computer vision problems. However, learning a new concept given few examples is still a very challenging task.



Figure 4: Some qualitative results of our method for 1-shot. Inside each tile, we have the support set at the top and the query image at the bottom. The support is overlaid with the ground truth in yellow and the query is overlaid with our predicted mask in red.

In this paper we developed a new architecture to address this problem for image segmentation. Our architecture learns to learn an ensemble classifier and use it to classify pixels in the query image. Through comprehensive experiments we show the clear superiority of our algorithm. The proposed method is considerably faster than the other baselines and has a smaller memory footprint.

## References

- [1] Evgeniy Bart and Shimon Ullman. Cross-generalization: Learning novel classes from a single example by feature replacement. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 672–679. IEEE, 2005.
- [2] Luca Bertinetto, João F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. In *Advances in Neural Information Processing Systems*, pages 523–531, 2016.
- [3] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Computer Vision and Pattern Recognition*, 2017.
- [4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

- [5] Liang-Chieh Chen, Jonathan T Barron, George Papandreou, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4545–4554, 2016.
- [6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.
- [7] Wenlin Chen, James T Wilson, Stephen Tyree, Kilian Q Weinberger, and Yixin Chen. Compressing neural networks with the hashing trick. In *ICML*, pages 2285–2294, 2015.
- [8] Xinlei Chen, Abhinav Shrivastava, and Abhinav Gupta. Enriching visual knowledge bases via object discovery and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2027–2034, 2014.
- [9] Jifeng Dai, Kaiming He, and Jian Sun. Boxesup: Exploiting bounding boxes to super-vise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1635–1643, 2015.
- [10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- [11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>.
- [12] Alon Faktor and Michal Irani. Co-segmentation by composition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1297–1304, 2013.
- [13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [14] Bharath Hariharan and Ross B. Girshick. Low-shot visual object recognition. *CoRR*, abs/1606.02819, 2016.
- [15] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014.
- [16] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 447–456, 2015.
- [17] Dorit S Hochbaum and Vikas Singh. An efficient algorithm for co-segmentation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 269–276. IEEE, 2009.
- [18] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 1495–1503, 2015.

- [19] Seunghoon Hong, Junhyuk Oh, Honglak Lee, and Bohyung Han. Learning transferable knowledge for semantic segmentation with deep convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3204–3212, 2016.
- [20] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614*, 2016.
- [21] Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. Learning to remember rare events. *arXiv preprint arXiv:1703.03129*, 2017.
- [22] Gregory Koch. *Siamese neural networks for one-shot image recognition*. PhD thesis, University of Toronto, 2015.
- [23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. URL <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>.
- [24] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.
- [25] Fei-Fei Li, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.
- [26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.
- [27] Mohammadreza Mostajabi, Payman Yadollahpour, and Gregory Shakhnarovich. Feed-forward semantic segmentation with zoom-out features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3376–3385, 2015.
- [28] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015.
- [29] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 30–38, 2016.
- [30] George Papandreou, Liang-Chieh Chen, Kevin Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. *arXiv preprint arXiv:1502.02734*, 2015.

- [31] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*, 2014.
- [32] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1796–1804, 2015.
- [33] Pedro O Pinheiro and Ronan Collobert. Weakly supervised semantic segmentation with convolutional networks. In *CVPR*, volume 2, page 6. Citeseer, 2015.
- [34] Rong Quan, Junwei Han, Dingwen Zhang, and Feiping Nie. Object co-segmentation via graph optimized-flexible manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 687–695, 2016.
- [35] Carsten Rother, Tom Minka, Andrew Blake, and Vladimir Kolmogorov. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 993–1000. IEEE, 2006.
- [36] Ruslan Salakhutdinov, Joshua B Tenenbaum, and Antonio Torralba. One-shot learning with a hierarchical nonparametric bayesian model. In *ICML Unsupervised and Transfer Learning*, pages 195–206, 2012.
- [37] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1842–1850, 2016.
- [38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
- [39] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016.
- [40] Yu-Xiong Wang and Martial Hebert. Learning to learn: Model regression networks for easy small sample learning. In *European Conference on Computer Vision*, pages 616–634. Springer, 2016.
- [41] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.