FS-CIS Net: Few-shot Instance Segmentation without Proposals

Divyansh Garg (dg595) Abhay Singh (as2626) Vineet Parikh (vap43) Andreas Schoenleben (as3932)

Abstract

Modern deep learning techniques excel with vasts amount of labeled data. However, in real-time autonomous driving scenarios, labeled samples will never exist for every possible object in a scene. To this end, we propose a novel network architecture that performs instance segmentation with few labeled samples.

Our Few-shot Clustering Instance Segmentation Network (FS-CIS Net) tackles the problem of proposal-free few-shot instance segmentation: given reference images of a previously unseen object of interest, directly segment all distinct instances of this same object in some query image, without the utilization of regional proposals.

We utilize an additional prototypical branch to extend existing pixel-embedding-based instance segmentation methods, and validate our approach on the PASCAL-5i dataset. We provide a strong baseline result for the task of proposal-free few-shot instance segmentation, and hope to inspire further research relevant to the task. We also find that, compared with existing instance segmentation model that extend Mask R-CNN, we achieve a significant speed-up in inference with regards to the one-shot setting, with comparable performance results.

1. Introduction

Infants do not possess prior knowledge of names of objects to categorize them. Nonetheless, they can easily recognize and distinguish between different objects/entities. The same holds true for an adult, who is able to observe a previously unseen object (*zero-shot learning*) or one seen previously with only a few examples (*few-shot learning*) and still distinguish between it and its surroundings (*instance segmentation*). Humans are remarkably good at using their prior knowledge of objects to distinguish them in novel situations. This ability to demarcate newly-seen instances in a scene from only a small number of training examples is what we hope to further explore, motivated by the fact that systems should be able to use prior knowledge of previously seen object instances to segment new instances in an image.

Such a few-shot learning approach to instance segmentation would have immediate benefit. It would enable segmentation of objects that have no large-scale annotated datasets, such as MS COCO (Lin et al., 2014), making it extremely useful in real-world applications. Furthermore, a proposal-free instance segmentation method is more suited for real-time applications such as autonomous driving, which requires fast execution with sufficiently high accuracy. An example would be avoiding oddly-sized objects that cannot necessarily be categorized (for e.g. traffic cones or roadblocks), but a few reference images of objects-to-avoid could aid in this situation.

Although current state-of-the-art instance segmentation methods, such as Mask R-CNN (He et al., 2017), are proposal-based, they are slow and generate masks at a fixed and low resolution. Other proposal-free works in the past such as bottom-up top-down combined segmentation (Levin & Weiss, 2006), have attempted to generalize well to unseen classes rely heavily on hand-crafted cues and can at best supplement existing class-based approaches; such works are thus limited in adaptability and accuracy. The amount of existing modern research in few-shot instance segmentation is limited, with few models (Michaelis et al., 2018) that tackle the task, and no such models that we know of that are proposal-free.

There have been few approaches to few-shot instance segmentation, most notably proposal-based ones such as Siamese Mask R-CNN (Shaban et al., 2017). However, proposal-based instance segmentation networks suffer slow speeds, making them hard to use and thus impractical in constrained, real-world applications; this is especially true for classifying objects that may be amorphous in shape. There are no models for this task that are intrinsically proposal-free.

For the remainder of this paper we pose and tackle few-shot instance segmentation without generated proposals. As input, we are given reference images of a previously unseen object of interest, and as output we wish to directly segment all distinct instances of this same object in some given query, all without regional proposals. Such a task is inherently class-agnostic.

More concretely, we formulate the task of few-shot instance segmentation as follows: Given a scene image (*query*) and a

previously unknown object category defined by *k* reference instances (*k-shot learning*), generate a segmentation mask for every instance of that category in the image. This task can be seen as an example-based extension of the typical instance segmentation setup, eliminating the need of fixed class categories.

Our new model, Few-shot Clustering Instance Segmentation Network (FS-CIS Net), Figure 3 incorporates ideas from few-shot learning, specifically prototypical networks (Snell et al., 2017), into proposal-free instance segmentation methods (Neven et al., 2019), to learn this task in a class-agnostic manner to generalize to previously unknown object categories.

2. Related Work

Bottom-Up Top-Down Instance Segmentation Traditional approaches attempted to combine low level features like color with high level features to generate segmentation masks. This task can be formulated in the form of conditional random fields, but these are also hand-engineered rather than being truly "learned" methods of performing instance segmentation (Levin & Weiss, 2006).

Siamese Mask R-CNN. To enable one-shot instance segmentation a 'Siamese', prototypical-inspired, branch is added to extend the Mask R-CNN model. The existing and added branches extract features from the reference and the query image respectively. The embedding of the query image, as well as the L1 difference of the output of embeddings from both branches are then fed into the Mask R-CNN head which performs the segmentation (Michaelis et al., 2018).

Spatial Embeddings. Davy Neven et al. suggest a fast and accurate proposal free instance segmentation method, utilizing a clustering-based approach. Their network possesses a seed branch and an instance branch. The seed map calculates class heat maps and the instance branch calculates offset vectors that point to the center of an instance. These embeddings are then clustered using a learnable bandwidth (Neven et al., 2019).

Prototypical Networks. To be able to segment objects with very few examples, prototypical networks learn a metric space where the distance between samples represents their dissimilarity. An additional 'Siamese' branch is used to calculate class prototypes from a given support set in the learned metric space. Query images are classified according to these prototypes (Snell et al., 2017).

3. Approach

As training input, we take a reference image containing some object, the object's associated instance mask(s) in the reference image, a query image, with the query's associated instance mask(s) as labels.

During inference, we only have the reference image, a query image, and instance mask(s) of some object in the reference; the objective is to directly output all instance masks in the query image of this object. In this stage, the object in the reference image need not necessarily be in the query image.

We combine semantic-based few-shot learning (Snell et al., 2017) and embedding-based segmentation methods (Neven et al., 2019). From the former we utilize the prototypical network architecture, and from the latter we utilize pixel embedding-based grouping methods. In conjunction, we train globally to tackle the issue of instance segmentation in a proposal-free manner, directly optimizing the intersection-over-union of each instance mask by clustering embedded pixels.

We optimize pixel embeddings such that embeddings from the same instance mask form a cluster in the pixel space, and that the embeddings in different instance-specific clusters are pushed away from each other. To this end, we utilize a clustering-based contrastive loss function for this task of proposal-free instance segmentation. Specifically, we use the Lovasz hinge loss (Berman & Blaschko, 2017) to optimize the model for maximum mAP IoU, a standard metric used in segmentation tasks.

With FS-CIS Net's fast architecture, our constructed prototypical network performs few-shot instance segmentation in real-time with sufficiently high accuracy, utilizing knowledge of the clusters in our pixel space.

Our model uses a proposal-free, instance-segmentation network proposed in (Neven et al., 2019) and adds an additional prototypical branch that takes a reference image and outputs a prototype for a specific class (Fig. 3). For k-shot learning, the prototypes are averaged over the k given reference images. For the tasks in the few-shot scenario, where a network must generalize to inputs previously unseen from a given training set, the few number of examples of each new class given directly is not sufficient. The prototypical branch enables FS-CIS Net to instead learn a metric space such that masks can be generated by utilizing distances to so-called 'prototype' representations of each class.

After learning these pixel embeddings, we then convert the distances between embeddings $e_i = x_i + o_j$ and the instance centroid C_k into a probability that the given pixel belongs to that instance with the equation

$$\phi_k(e_i) = exp(-\frac{||e_i - C_k||^2}{2\sigma_k^2})$$

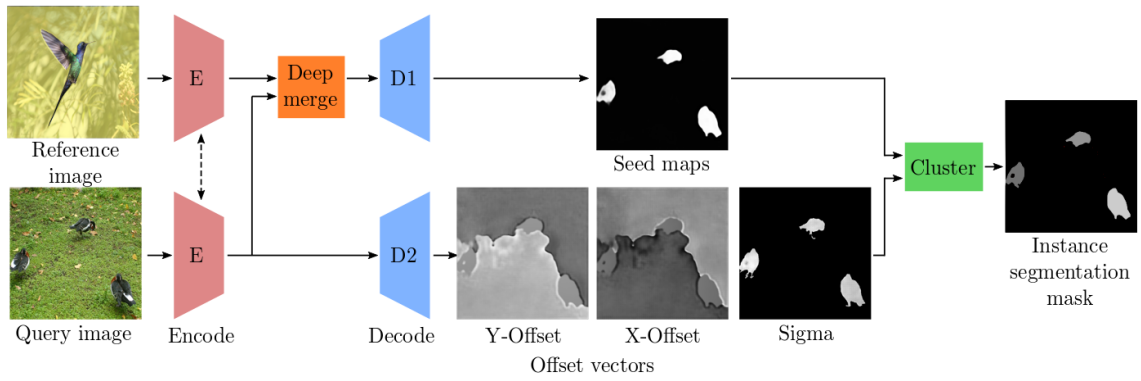


Figure 1. **FS-CIS Net** Architecture. Encoders E share weights and calculate embeddings of the reference and query images respectively. The Distance between the embeddings is calculated using some distance measure. Decoders generated class seed map, center offset vectors and clustering bandwidth. Both branches are merged with a clustering algorithm.

In the process, we can also derive the seed loss as

$$L_{seed} = \frac{1}{N}\Sigma_i^N(\mathbb{1}_{[s_i \in S_k]}||s_i - \phi_k(e_i)||^2 + \mathbb{1}_{[s_i \in bg]}||s_i - 0||^2)$$

The above equation essentially shows that if a pixel embedding has a high seed score, it's fairly close to an object's center.

We combine the embeddings using a 'DeepMerge' strategy. The DeepMerge strategy uses different matching loss between the reference prototypes and the query image's embeddings. Examples include taking the L1 or L2 loss between the reference and query image embeddings and concatenating them. These features are then fed into decoder D1, which generates a class-specific probability distribution (seed maps) over the pixels of the query image. Furthermore, decoder D2 operates on the class-agnostic features of the query image and produces offset vectors for each pixel, indicating the x and y offset to the nearest instance centroid. The sigma values are predicted per pixel, and correspond to the learnable clustering margin for each cluster, which controls the size of the representative clusters in the image space. The class-specific seed maps are used to pick the centers for each instance. The offset maps and the sigma maps are then used to cluster and predict instances around these centers.

4. Experiments/Results

We evaluate our approach against the performance of other learned forms of few-shot instance segmentation, particularly considering Siamese Mask R-CNN (Michaelis et al., 2018). We primarily base accuracy measurements off of the Pascal VOC's Jaccard Index, which measures the mean intersection over union (mean IOU) for pixels belonging to a single instance.(Everingham et al., 2015).

To this end, we base our tests on a standard benchmark in the few-shot learning scenario: Pascal-5i, an extension of the existing Pascal-VOC dataset for few-shot scenarios. The PASCAL Visual Object Classes dataset has 20 classes for which instance segmentation masks are provided. Pascal-5i has 5 different folds each containing some specific classes, so that a network can be trained and tested on different folds.

When testing for accuracy with regards to mean AP on the fifth fold (unseen classes) in order to ensure a true few-shot setting, we have a mean IOU of 0.21 in the 1-shot setting, and a mean IOU of .35 in the 5-shot setting. While we are currently in the process of comparing against Siamese Mask R-CNN's accuracy metrics for this benchmark, we do believe that this poses a strong initial baseline for the setting of few-shot instance segmentation without proposals and is significantly better than random chance in both settings (roughly less than or equal to .1 for most categories). As an initial comparison, we broadly estimate Siamese accuracy for the 1-shot case and 5-shot case by using the proportional change in accuracy for a similarly-structured Mask R-CNN model between the Pascal-VOC and COCO datasets, but we plan to fully verify these estimates for both the one-shot and five-shot IoU cases.

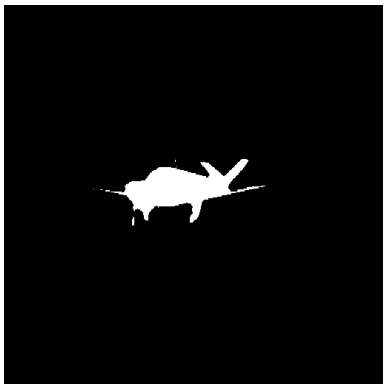
We also find that, within the 1-shot setting, we achieve an inference time of roughly 44.6ms per image on the fifth fold, while for Siamese Mask R-CNN we estimate the inference time on Pascal-VOC to be 129ms per image. Thus we believe we can achieve a 3x improvement in speed without a significant change in accuracy, making FS-CIS ideal for real-time applications.

5. Discussion and Conclusion

We proposed FS-CIS Net, a method to do real-time few-shot instance segmentation of images. On experimenting and

(a) Reference image (mask not shown but is overlaid over plane)

(b) Query image (different-looking plane)



(c) Mask produced on query image

Figure 2. Reference, query, and produced mask on the four trained folds

In this paper, we want to tackle the problem of proposal-free few-shot instance segmentation: given reference images of a previously unseen object of interest, directly segment all distinct instances of this same object in some query image, without the utilization of regional proposals. Such a task is inherently class-agnostic.

Table 1. Results and estimated benchmarks

Metric	One-shot IoU	Five-shot IoU	One-shot Timing (ms)
FS-CIS	0.21	0.35	44.9ms
Siamese Mask R-CNN (estimated)	N/A(estimated .36)	N/A(estimated 0.42)	129ms

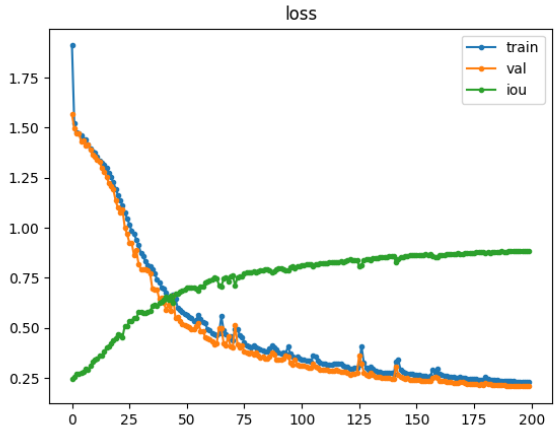


Figure 3. Results of training/validation loss and IoU from training over 200 epochs on the four trained folds, where IoU saturates to 0.88

testing our methods, we found that we can significantly improve the speed of the system by 3x while limiting accuracy loss.

Nevertheless, our results remain lower in accuracy compared to Siamese Mask R-CNN, a slower proposal-based method. There might be a potential to improve our method's accuracy by using a better strategy for merging the reference and query features for the few-shot setting.

6. Future Work

There are multiple options for extending this project beyond the limits of this course:

1. Experimenting with different strategies for merging reference and query features, in order to better optimize accuracy, and evaluating against other potential benchmarks.

2. Evaluating this approach against of cue-based attempts at grouping pixels to delineate objects, especially bottom-up top-down instance segmentation methods(Levin & Weiss, 2006), to understand whether hand-engineered groupings based off of cues can be more effectively replaced by learned embeddings.

3. Extending this platform to work on a noisy, real-world embedded system, to determine usability in different systems and ML settings, such as robotics, distributed systems, and cloud computing.

4. Considering how different a query image needs to be from a reference image before inference fails, and also

considering if similar objects in a reference image can be used to give masks for query images.

5. Embedding information given with this few-shot approach along with concept embeddings, and potentially using those to improve performance.

References

Berman, M. and Blaschko, M. B. Optimization of the jaccard index for image segmentation with the lovász hinge. *CoRR*, abs/1705.08790, 2017. URL http://arxiv.org/abs/1705.08790.

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, January 2015.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. B. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. URL http://arxiv.org/abs/1703.06870.

Levin, A. and Weiss, Y. Learning to combine bottom-up and top-down segmentation. In Leonardis, A., Bischof, H., and Pinz, A. (eds.), *Computer Vision – ECCV 2006*, pp. 581–594, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-33839-0.

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

Michaelis, C., Ustyuzhaninov, I., Bethge, M., and Ecker, A. S. One-shot instance segmentation. *CoRR*, abs/1811.11507, 2018. URL http://arxiv.org/abs/1811.11507.

Neven, D., Brabandere, B. D., Proesmans, M., and Gool, L. V. Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2019.

Shaban, A., Bansal, S., Liu, Z., Essa, I., and Boots, B. One-shot learning for semantic segmentation. *CoRR*, abs/1709.03410, 2017. URL http://arxiv.org/abs/1709.03410.

Snell, J., Swersky, K., and Zemel, R. S. Prototypical networks for few-shot learning. *CoRR*, abs/1703.05175, 2017. URL http://arxiv.org/abs/1703.05175.