The Role of Data for One-Shot Semantic Segmentation

Timo Lüddecke and Alexander Ecker
University of Goettingen
Computer Science Department, Neural Data Science Group
{lueddecke, ecker}@cs.uni-goettingen.de

Abstract

In this work we investigate the potential of larger datasets for one-shot semantic segmentation. While computer vision models are often trained on millions of diverse samples, current one-shot semantic segmentation datasets encompass only a small number of samples (PASCAL-5i), a small number of classes (PASCAL-5i and COCO-20i) or have little variability (FSS-1000). To improve this situation, we introduce LVIS-OneShot, a one-shot variant of the LVIS dataset. With 718 classes and 114,347 images, it exceeds previous datasets substantially in terms of size. By controlled experiments we show that not only the number of images but also the number of different classes is crucial. We analyze transfer learning across common datasets and find that by training on LVIS-OneShot we outperform current state-of-the-art models on PASCAL-5i. In particular, we observe that a simple baseline model (MaRF) learns to perform one-shot segmentation when trained on a large dataset although it has a generic architecture without strong inductive biases. Code and dataset are available here: eckerlab.org/code/one-shot-segmentation

1. Introduction

In many fields of computer vision, we witnessed a trend towards training conceptually simple models with a large number of parameters on large-scale datasets involving millions of samples, starting from AlexNet [12] on the ImageNet dataset [5] to the recent Big Transfer [11] and CLIP [20] on text segments from the Internet. Instead of designing mechanisms to solve specific tasks (for example fine-grained recognition), features are learned by scaling up the dataset and applied on downstream tasks by fine-tuning, learning a linear classifier from frozen features or zero-shot transfer [20]. Very recently, similar findings were made for one-shot image classification [25] and object detection [15].

One may wonder to what degree these findings can be replicated for one-shot segmentation, which requires dynamic adaptation at inference time to make pixel-wise predictions. One-shot segmentation models are currently trained and evaluated on datasets with relatively few classes (PASCAL-5i [23]: 20, COCO-20i [17]: 80) or little variation in object size and position (FSS-1000 [28]). At the same time, models use customized mechanisms to tackle one-shot segmentation, such as operating on multiple scales [26] or modeling object parts [14, 30, 4] (see Table 1 for an overview). Thus, it is quite possible that scaling up datasets in terms of labels and classes could similarly boost performance in one-shot segmentation and dwarf the advances made by previous attempts at engineering customized architectures for small datasets. In particular, we believe the diversity of labels to play a pivotal role.

In order to investigate these questions, we introduce the LVIS-OneShot dataset which is a variant of the LVIS dataset [7], originally intended for long-tail instance segmentation. Contrary to previous one-shot segmentation datasets, it combines a large number of samples with a large number of classes. Furthermore, we intentionally place all PASCAL-5i classes into the test set to simplify evaluating performance on PASCAL-5i. In addition to the dataset, we present the primitive baseline model MaRF without bells and whistles which relies on very simple design: ResNet backbones [9], masked (global) pooling and feature-wise linear modulation (FiLM) [6]. Based on both, LVIS-OneShot dataset and MaRF baseline, as well as avail-
and, due to an overlap between PASCAL-5 and L VIS, the small number of classes limits the utility of this dataset for one-shot (and few-shot) segmentation dataset. However, we address both short-comings with a novel dataset (Fig. 2) based on the recently proposed LVIS [7] labels for COCO images [13].

**2. A Diverse Dataset: LVIS-OneShot**

With COCO-2014 [18] there already exists a large-scale one-shot (and few-shot) segmentation dataset. However, the small number of classes limits the utility of this dataset and, due to an overlap between PASCAL-5 and COCO-2014 splits, transfer learning is cumbersome since a new model needs to be trained for each split. We address both shortcomings with a novel dataset (Fig. 2) based on the recently proposed LVIS [7] labels for COCO images [13].

**PASCAL-5 Overlap** As a first step, we identify all classes in LVIS that overlap with Pascal. To automate this process, we make use of the WordNet [16] synset assignments of LVIS. First, we manually assign the corresponding synset for all 20 Pascal classes. For each LVIS class we recursively traverse the set of hypernyms (i.e. more general meanings) and check if it has an intersection with the Pascal synsets.

**3. Complex and Simple Models**

Arguably the best performing model for one-shot segmentation is PFENet [26]. It contains several custom mechanisms, e.g. multi-scale processing and a prior based on high-level feature similarity. Contrary, our baseline model (Fig. 4) relies on only three basic and well-known components: Masked pooling [32], the ResNet architecture [9] and the FiLM conditioning mechanism [6], hence we call it MaRF. It is meant as a low-modelling, generic counterpoint to the complex architecture defined by PFENet. Query and support branch only interact through a single vector.

**Support encoding** The support encoder $s$ takes the support tuple $t$, containing an image with corresponding segmentation, and encodes it into a conditional vector $c$. For this, features from a ResNet50 or ResNet18 [9] are extracted at the 3rd or 4th residual block. Then, masked pooling transforms these feature maps into a single vector by averaging all feature vectors that pertain to the object (indicated by the support segmentation map). Given a support image, segmentation pair $t = (t_{img}, t_{seg})$ and an operator which resizes $t_{seg}$ to match the spatial size of the output of $s_{enc}$, we obtain the conditional vector $c = \text{AvgPool}(s_{enc}(t_{img}) \odot \text{resize}(t_{seg}))$, where $\odot$ is the pointwise multiplication.

**Query processing** The query network $q$ takes the query image $x$ and the generated conditional vector $c$ and outputs a binary segmentation mask $y$. $q$ consists of a CNN encoder which generates a high level representation of the query image and a decoder which forms the output using skip connections, similar to the U-Net [22]. The conditional vector $c$ is fused into the query network $q_{enc}$ at layer $l$ through feature-wise linear modulation [6]. Afterwards, using information from the support image/segmentation, the decoder generates an output tensor of the same spatial size as the input. It consists of four blocks with channel sizes of 128, 128, 128 and 32, each incorporates a skip connection from the decoder. Instead of using transposed convolutions, spatial resolution is increased using bilinear interpolation followed by a $5 \times 5$ convolution. We can describe the computation of a segmentation mask $y$ by:

$$y = q_{dec}(q_{enc}^{[L]}(q_{enc}^{[0:L]}(x)\phi(c) + \pi(c)))$$  \hspace{1cm} (1)$$

where the query network $q_{enc}$ is be decomposed into layers before $(q_{enc}^{[0:L]}$) and after $(q_{enc}^{[L]}$) conditioning layer $L$ where information from the support image is fused. Analogously to the support encoder $s$, the query encoder $q_{enc}$ is implemented by a ResNet with 18 or 50 layers. Following Tian et al. [26], we replace the standard ResNet50 with a ResNet50 from PSPNet [33] in both encoder and query network. If not stated otherwise, ImageNet weights are frozen in both encoders up to the 4th residual block.

**4. Experiments**

**Datasets and Metrics** We perform our analysis on three datasets: LVIS-OneShot, FSS-1000 [29] and PASCAL-
Figure 4: Overview of MaRF (with FiLM at 3): An encoder-decoder network (blue) processes the query image and generates an output segmentation. Information about the search target is introduced through another encoder (yellow) which uses masked pooling.

LVIS-OneShot Training Instead of a fixed assignment of training image pairs, we randomly sample pairs of one category during training. Images are scaled to have a minimal side length of 480px and are then cropped to a square-shaped image. We apply mild augmentation [3] to the images involving horizontal flip as well as HSV and gamma change. For validation and test, 1000 and 10,000 fixed pairs are used without augmentation to reduce variance.

Technical Details We use PyTorch [19], Adam optimizer [10] with varying learning rates (LR), batch sizes (BS) and early stopping patience (ES) as shown below:

<table>
<thead>
<tr>
<th>dataset</th>
<th>epochs</th>
<th>ES</th>
<th>LR</th>
<th>BS</th>
</tr>
</thead>
<tbody>
<tr>
<td>LVIS-OneShot</td>
<td>70</td>
<td>9</td>
<td>0.0001</td>
<td>32</td>
</tr>
<tr>
<td>PASCAL-5\textsuperscript{i}</td>
<td>2</td>
<td>-</td>
<td>0.0003</td>
<td>32</td>
</tr>
<tr>
<td>FSS-1000</td>
<td>25</td>
<td>5</td>
<td>0.0001</td>
<td>64</td>
</tr>
</tbody>
</table>

4.1. MaRF Configurations

We find that early conditioning (FiLM at layer 3) yields slightly better results on LVIS-OneShot than late conditioning (layer 5). For the smaller PASCAL-5\textsuperscript{i} dataset, the opposite is true. This means a network relying on earlier conditioning is generally favorable but more data is required to learn such a mechanism. This insight would not be visible if only a small dataset was used. Also freezing weights, like in [26], improves the performance compared to training all weights. This is likely because it prevents overfitting as LVIS-OneShot is still much smaller than image classification datasets the ResNets are normally trained on. The PSPNet [33] modification of ResNet50 turns out to be an important factor as it performs much better than conventional ResNet50 (labeled “original” in Table 2). The quality of features is an essential predictor of final performance.

4.2. Sample-efficiency and label diversity

LVIS-OneShot is larger than competitive datasets in terms of the number of categories contained. This allows us to analyze the effect of label diversity for one-shot segmentation: Given a fixed budget of samples, is there an advantage of having a diverse set of images? In order to answer this question we generate a set of categories \( C \) containing a specific number of samples using an iterative algorithm.

Regarding number of samples (Fig. 5, left), we find a positive relationship between number of samples and performance, which was expected as more samples generally improve performance. We observe a similarly strong correlation between performance and sample diversity (Fig. 5 right), even when the number of samples is kept constant. This result supports our intuition that not only sample size but also label diversity is crucial. The fine-grained division of classes conveys additional information useful for one-shot segmentation.

4.3. Comparison with State-of-the-art

PASCAL-5\textsuperscript{i} For evaluation, we use the PASCAL-5\textsuperscript{i} implementation provided by Tian et al. [26] using their train-
### Table 3: One-shot segmentation performance on PASCAL-5\textsuperscript{i}.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>IoU\textsubscript{BIN}</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trained on PASCAL-5\textsuperscript{i}</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DAN [4]</td>
<td>RN101</td>
<td>71.9</td>
<td>58.2</td>
</tr>
<tr>
<td>PFENet [26]</td>
<td>RN50</td>
<td>73.3\textsuperscript{*}</td>
<td>60.8</td>
</tr>
<tr>
<td>PFPM [26]</td>
<td>RN50</td>
<td>71.2\textsuperscript{*}</td>
<td>60.6\textsuperscript{*}</td>
</tr>
<tr>
<td>RePRI [2]</td>
<td>RN50</td>
<td>-</td>
<td>56.3</td>
</tr>
<tr>
<td>MardR (FiLM 3, ours)</td>
<td>RN50</td>
<td>58.3</td>
<td>33.3</td>
</tr>
</tbody>
</table>

### Table 4: Comparison between PFENet and MaRF (RN50) on the new LVIS-OneShot dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>Backbone</th>
<th>IoU\textsubscript{BIN}</th>
<th>IoU\textsubscript{mIoU}</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFENet</td>
<td>RN50</td>
<td>78.5</td>
<td>64.3</td>
</tr>
<tr>
<td>MardR (FiLM 3, ours)</td>
<td>RN50</td>
<td>77.4</td>
<td>61.1</td>
</tr>
</tbody>
</table>

Table 3: One-shot segmentation performance on PASCAL-5\textsuperscript{i}. *using weights provided by the PFENet authors.

Table 4: Comparison between PFENet and MaRF (RN50) on the new LVIS-OneShot dataset.

Transfer Learning LVIS-OneShot → PASCAL-5\textsuperscript{i} All existing approaches on PASCAL-5\textsuperscript{i} are outperformed by training MaRF LVIS-OneShot (without fine-tuning), despite the distribution shift (Fig. 3). The results on PASCAL-5\textsuperscript{i} (Table 3) show the utility of training on LVIS-OneShot for the simple MaRF model. With an mIoU score of 61.1 and IoU\textsubscript{BIN} of 77.4, we clearly outperform the best reported result: PFENet trained on PASCAL-5\textsuperscript{i}. Also PFENet benefits from training on LVIS-OneShot, establishing a slightly better score. However, its improvements are much smaller than those of MaRF.

### 4.3.1 LVIS-OneShot

As shown in Table 4, PFENet exhibits a better performance on LVIS-OneShot than MaRF. However, MaRF achieves a greater performance gain from using more samples (also see Fig. 1). On one hand, this gain supports our intuition that data can outweigh model design to some extent. On the other hand, there remains a gap to PFENet, suggesting that the inductive biases of PFENet are generally useful for one-shot segmentation and are not overfit to the classes present in small segmentation datasets.

### 4.3.2 FSS1000

The results on FSS-1000 (Table 5) show that MaRF outperforms earlier work, except for the DAN model [4] which uses a larger encoder. This supports the claim that simple models match state-of-the-art performance with sufficient training data, evading the need of model design.

Surprisingly, a baseline of our model that did not receive a support image/segmentation (labeled “no support” in Table 5) achieved decent performance, and even outperformed all previously published approaches except DAN [4]. This result suggests that FSS-1000 is strongly biased towards centered objects and has little variation in object size and location.

To further investigate the biases of FSS-1000, we introduce 50% negative samples to the test set. We observe a strong drop in performance, while the models trained on LVIS-OneShot cope best with this setting. Possibly due to these different statistics, transfer learning from LVIS-OneShot does not work as well for FSS-1000 as for PASCAL-5\textsuperscript{i}.

### 5. Discussion and Conclusion

Large-scale training can replace model design and strong inductive biases in one-shot segmentation. This result is consistent with previous findings in computer vision [24, 20, 15] and NLP [8, 21]. We find conceptually simple models to profit to a much greater extent from more samples and more diverse samples than the complex PFENet. The latter achieved only small gains from a substantial increase of dataset richness. However, PFENet performance remains slightly better than our baseline, suggesting that inductive biases still matter in the large data-regime (although to a smaller extent) and PFENet modeled the right ones. For future research in one-shot semantic segmentation, our findings represent a strong argument in favor of using large and diverse datasets. We recommend to consider PASCAL-5\textsuperscript{i} primarily a test dataset.
References


[28] Tianhan Wei, Xiang Li, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-


