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-5^i*), a small number of classes (PASCAL-*5^i *and COCO-*20^i*) 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-*5^i*. 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

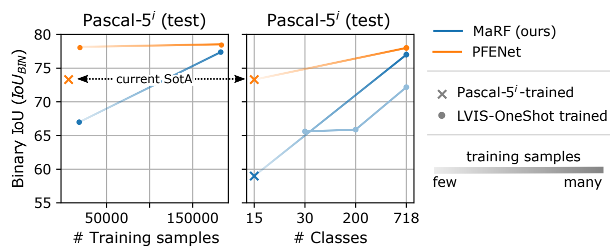

Figure 1: One-shot segmentation on PASCAL-5^i benefits from the larger and more diverse LVIS-OneShot dataset. The simple MaRF baseline profits from more (left) data and more diverse (right) data. Both, PFENet and MaRF exceed the current state-of-the-art (SotA).

predictions. One-shot segmentation models are currently trained and evaluated on datasets with relatively few classes (PASCAL-5^i [23]: 20, COCO-20^i[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 mutliple 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-5^i classes into the test set to simplify evaluating performance on PASCAL-5^i. 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-

Mechanism	Used by
modeling of object parts	[14, 30]
graph attention on regions/ parts	[4, 31]
multi-scale features / feature pyramid	[26, 1, 31, 4]
high-level feature similarity prior	[26]
metric learning & prototype alignment loss	[27]

Table 1: Overview of mechanisms used for one-shot segmentation.

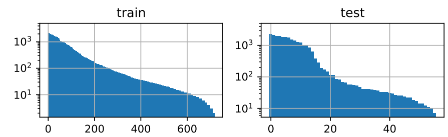

Figure 2: Distribution of the class frequencies in LVIS-OneShot

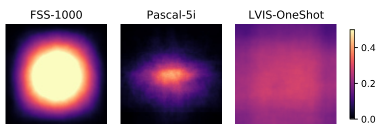

Figure 3: Average images of 5,000 random ground truth segmentations of FSS-1000, PASCAL-$5^i$and LVIS-OneShot reveal dataset bias.

able datasets we conduct extensive experiments seeking to understand the role of data in one-shot segmentation.

2. A Diverse Dataset: LVIS-OneShot

With COCO-20^i[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^i and COCO-20^i 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^i 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 $\mathbf{t}$, containing an image with corresponding segmentation, and encodes it into a conditional vector $\mathbf{c}$. 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 $\mathbf{t} = (\mathbf{t}_{\text{img}}, \mathbf{t}_{\text{seg}})$ and an operator which resizes $\mathbf{t}_{\text{seg}}$ to match the spatial size of the output of s_{enc}, we obtain the conditional vector $\mathbf{c} = \text{AvgPool}(s_{\text{enc}}(\mathbf{t}_{\text{img}}) \odot \text{resize}(\mathbf{t}_{\text{seg}}))$, where $\odot$ is the pointwise multiplication.

Query processing The query network q takes the query image $\mathbf{x}$ and the generated conditional vector $\mathbf{c}$ and outputs a binary segmentation mask $\mathbf{y}$. 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 $\mathbf{c}$ 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 encoder. Instead of using transposed convolutions, spatial resolution is increased using bilinear interpolation followed by a 5×5 convolution. We can describe the computation of a segmentation mask $\mathbf{y}$ by:

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

where the query network q_{enc} is be decomposed into layers before ($q_{\text{enc}}^{[0:L]}$) and after ($q_{\text{enc}}^{[L:]}$) 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-

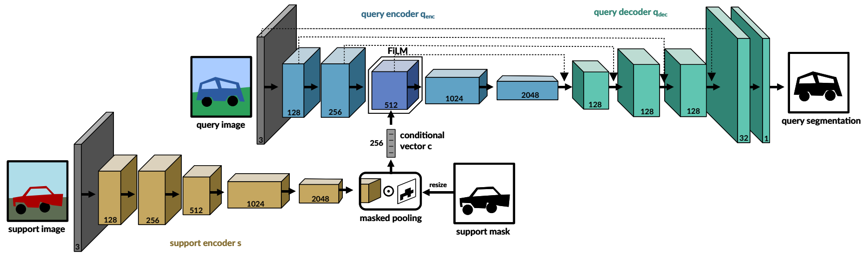

Figure 4: Overview of MaRF (with FiLM at 3): An **encoder**-**decoder** network (blue) processes the query image and generates an output segmentation. Information about the search target is introduced through another **encoder** (yellow) which uses masked pooling.

5^i[23]. These datasets vary strongly in the average size and position of the object to be segmented (see Fig. 3). If not stated otherwise, both encoders q_{enc} and s_{enc} are initialized with weights obtained through ImageNet-pretraining as this is a common practice in one-shot segmentation. We exclusively use binary segmentation problems which have only a single foreground class, i.e. one-shot, one-way segmentation. We use the intersection over union-based (IoU) metrics mean IoU (mIoU), binary (or foreground-background) IoU and foreground IoU.

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:

dataset	epochs	ES	LR	BS
LVIS-OneShot	70	9	0.0001	32
PASCAL-5^i	2	-	0.0003	32
FSS-1000	25	5	0.0001	64

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^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.

	Backbone	FiLM	IoU_{BIN}	mIoU
LVIS-OneShot	RN50	5	68.2	38.7
	RN50	3	**71.4**	**42.0**
	RN50 (no freeze)	3	67.1	35.9
	RN50 (original)	3	66.3	35.1
PASCAL-5^i	RN50	3	58.3	33.3
	RN50	5	**59.4**	**35.6**

Table 2: Performance on different configurations of MaRF. FiLM: conditioning layer L.

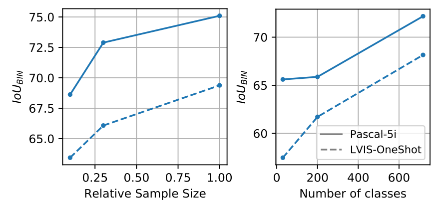

Figure 5: A larger number of samples (left) as well as more classes (right) have a positive impact on one-shot segmentation performance of MaRF (training on LVIS-OneShot).

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^i For evaluation, we use the PASCAL-5^i implementation provided by Tian et al. [26] using their train-

Method	Backbone	IoU_{BIN}	mIoU
Trained on PASCAL-5^i			
DAN [4]	RN101	71.9	58.2
PFENet [26]	RN50	**73.3**	**60.8**
PFENet [26]	RN50	71.2*	60.6*
RPMM [30]	RN50	-	56.3
RePRI [2]	RN50	-	59.7
MaRF (FiLM 3, ours)	RN50	58.3	33.3
Trained on FSS-1000			
FSS Basel. [29]	VGG16 -	-	**58.6**
MaRF (FiLM 3, ours)	RN50	67.3	45.4
Trained on LVIS-OneShot			
PFENet [26]	RN50	**78.5**	**64.3**
MaRF (FiLM 3, ours)	RN50	77.4	61.1

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

	10% of data		100% of data		Δ	
Model	IoU_{BIN}	mIoU	IoU_{BIN}	mIoU	IoU_{BIN}	mIoU
PFENet	**71.7**	**45.1**	**72.3**	**46.6**	0.6	1.5
MaRF	64.4	31.8	71.4	42.0	**7.0**	**10.2**

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

ing and validation sets. We find our MaRF model to perform quite poorly when it was trained on PASCAL-5^i. This is expected due to its simplicity without explicit mechanisms (or inductive biases) for one-shot segmentation and the small size of the dataset.

Transfer Learning LVIS-OneShot → PASCAL-5^i All existing approaches on PASCAL-$5^i$are outperformed by training MaRF LVIS-OneShot (without fine-tuning), despite the distribution shift (Fig. 3). The results on PASCAL-5^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_{BIN} of 77.4, we clearly outperform the best reported result: PFENet trained on PASCAL-5^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.

Model	Backbone	mIoU	$mIoU_{neg}$	IoU_{BIN}	IoU_{FG}
Trained on FSS-1000					
PFENet [26]	RN50	-	-	-	80.8*
DAN [4]	RN101	-	-	-	**85.2**
MaRF (FiLM 3) + aug.	RN50	**81.2**	42.3	**88.9**	83.3
MaRF (F. 3) (no support)	RN50	79.5	41.1	87.5	81.2
Trained on LVIS-OneShot					
PFENet [26]	RN50	76.8	**55.3**	85.5	78.2
MaRF (FiLM 3, ours)	RN50	70.6	51.7	81.5	72.1

Table 5: Performance on FSS-1000 in comparison with state-of-the-art methods.

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-1000is 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^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^i primarily a test dataset.

References

[1] Reza Azad, Abdur R Fayjie, Claude Kauffmann, Ismail Ben Ayed, Marco Pedersoli, and Jose Dolz. On the texture bias for few-shot cnn segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[2] Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? *arXiv preprint arXiv:2012.06166*, 2020.

[3] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 2020.

[4] Xianbin Cao and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. *European Conference on Computer Vision (ECCV)*, 2020.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009.

[6] Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. Feature-wise transformations. *Distill*, 2018.

[7] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[8] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. *arXiv preprint arXiv:1912.11370*, 6, 2019.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C.Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

[14] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. Part-aware prototype network for few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 142–158. Springer, 2020.

[15] Claudio Michaelis, Matthias Bethge, and Alexander S. Ecker. Closing the generalization gap in one-shot object detection. 2020.

[16] George A Miller. Wordnet: a lexical database for english. 38(11):39–41, 1995.

[17] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[18] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 622–631, 2019.

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[21] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 2015.

[23] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *BMVC*, 2017.

[24] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, 2017.

[25] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: A good embedding is all you need? In *European Conference on Computer Vision (ECCV)*, Cham, 2020. Springer International Publishing.

[26] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *IEEE transactions on pattern analysis and machine intelligence*, PP, August 2020.

[27] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9197–9206, 2019.

[28] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1328–1338, 2019.

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

shot segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[30] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *European Conference on Computer Vision (ECCV)*, 2020.

[31] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9587–9595, 2019.

[32] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybernetics*, 2020.

[33] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.