ARNET: ACTIVE-REFERENCE NETWORK FOR FEW-SHOT IMAGE SEMANTIC SEGMENTATION

Guangchen Shi[1], *Yirui Wu*[1✉], *Shivakumara Palaiahnakote*[2], *Umapada Pal*[3] *and Tong Lu*[4]

[1]College of Computer and Information, Hohai University, Nanjing, 211106 China,
shi.guangchen@foxmail.com, wuyirui@hhu.edu.cn;
[2]Department of Computer System and Information Technology, University of Malaya,
Kuala Lumpur, 50603 Malaya, shiva@um.edu.my;
[3]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute,
Kolkata, 600029 India, umapada@isical.ac.in;
[4]National Key Lab for Novel Software Technology, Nanjing University,
Nanjing, 210093 China, lutong@nju.edu.cn.

ABSTRACT

To make predictions on unseen classes, few-shot segmentation becomes a research focus recently. However, most methods build on pixel-level annotation requiring quantity of manual work. Moreover, inherent information on same-category objects to guide segmentation could have large diversity in feature representation due to differences in size, appearance, layout, and so on. To tackle these problems, we present an active-reference network (ARNet) for few-shot segmentation. The proposed active-reference mechanism not only supports accurately co-occurrent objects in either support or query images, but also relaxes high constraint on pixel-level labeling, allowing for weakly boundary labeling. To extract more intrinsic feature representation, a category-modulation module (CMM) is further applied to fuse features extracted from multiple support images, thus forgetting useless and enhancing contributive information. Experiments on PASCAL-5^i dataset show the proposed method achieves a m-IOU score of 56.5% for 1-shot and 59.8% for 5-shot segmentation, being 0.5% and 1.3% higher than current state-of-the-art method.

Index Terms— Few-shot segmentation, weekly-labeled supported, active-reference mechanism, few-shot learning

1. INTRODUCTION

Deep learning has significantly contributed to the development of multimedia domain. However, they generally require a large-scale dataset to train, where labeling is extremely time-consuming and annoying. Essentially, humans could understand a new concept with a few samples, which inspires researchers to transfer knowledge from known to unknown.

In this paper, we follow the idea of few-shot learning to tackle the problem of few-shot segmentation, which aims to locate foreground pixels of unseen class with clues from a few labeled samples. As shown in Fig. 1(a), most current methods [1, 2] follow the design of two-branch structure, which extracts segmentation related knowledge from support branch to perform tasks in query branch. To better find the co-occurrence objects, CRNet [3] proposes a cross-referencer mechanism to concurrently make predictions for both the support and query images, where we show its workflow in Fig. 1(b). However, their network couldn't support weakly labelling, since their designed cross-reference procedures require reinforced feature representation to provide an auxiliary loss in the training phase.

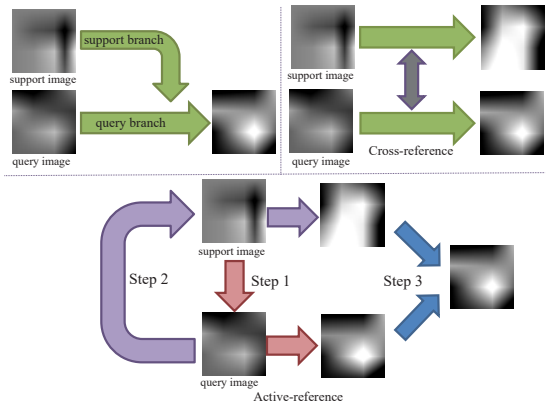


Fig. 1. Comparison among previous two-branch structure (a), CRNet with cross-reference mechanism (b), and the proposed active-reference network (c).

To model co-occurrence context information with weakly-labeled support images, we propose an active-reference mechanism, which is shown in Fig. 1(c). We describe the proposed mechanism with three steps. In the first step, support image guides the segmentation of query image. Afterwards, query image with segmentation result is fed back

978-1-6654-3864-3/21/$31.00

to guide the precise segmentation of support image. Finally, such co-occurrence context information is utilized to refine the query mask. Essentially, the proposed active-reference mechanism not only allows for weak boundary labels due to re-prediction for eliminating wrong predictions, but also offers a description on category-level context information for better co-occurrence segmentation by designing mutual segmentation strategy.

Most previous methods achieve k-shot segmentation results by applying non-learnable methods to fuse segmentation results obtained by 1-shot model, such as averaging and feature concatenating. To successfully fuse and align features extracted from multiple support images, we design a category-modulation module(CMM) to totally forget extrinsic information and enhance intrinsic characteristics. The proposed CMM consists of several forget and enhance blocks (FEBs), which contain two parts, i.e., a forget and an enhance block. Specifically, its former block equipped with a forget-gated activation could forget or optimize existing information, meanwhile the latter block equipped with an enhance-gated activation could generate new information to establish dense connections.

The main contributions of this paper are as follows:

- We propose an active-reference mechanism to mutually segment support and query images, which not only better locates co-occurrence objects by involving category-level context information, but also allows for weak boundary labels.
- For *k*-shot segmentation, we propose a category-modulation module to fuse features extracted from support images, which removes differences in feature representation of same-category objects by forgetting extrinsic and enhancing intrinsic information.

2. RELATEDWORK

2.1. Few-shot Learning

We roughly divide current few-shot learning methods into two categories, i.e, initialization based and metric learning based methods. The former methods generally define few-shot learning as "learning to fine-tune", which aims to learn proper model initialization or predict network parameters. For example, Lee et al. [4] believes training linear classifier in the few-shot regime provides for better generalization performance, where they successfully learn feature embedding that generalizes well under a linear classification rule for novel categories.

Metric learning based methods aim to close the gap among samples of the same category, while widening the gap among samples of different categories in the embedding space. For example, Wei et al. [5] propose a simple and interpretable universal weighting framework to estimate informativeness of heterogeneous features, which provides a tool to analyze the interpretability of various loss functions. Most related to our work, Relation Network (RN) [6] learn a deep distance metric to compare a small number of images after transforming them into embedding space for feature representation.

2.2. Few-shot Semantical Segmentation

Few-shot semantic segmentation extends segmentation to any new category with only a few annotated examples. Many works formulate the few-shot segmentation task as a guided segmentation task with a two-branch structure. For example, Shaban et al. [1] first applies few-shot learning on semantic segmentation with a two-branch structure, where the support branch directly predicts weights of the last layer in the query branch for segmentation. and Michaelis et al. [7] combine a siamese embedding with a U-net to segment an unseen object in a cluttered scene guided by a single example. Recently, Zhang et al. [8] present CANet, where their proposed two-branch dense comparison module performs multi-level feature comparison between the support image and the query image, and an iterative optimization module iteratively refines the predicted results.

For *k*-shot segmentation question, most existing methods firstly use the 1-shot model to predict each support image and then fuse these segmentation results for final output. Zhang et al. [8] tries to model such process as a learnable structure by utilizing an attention mechanism for result fusion.

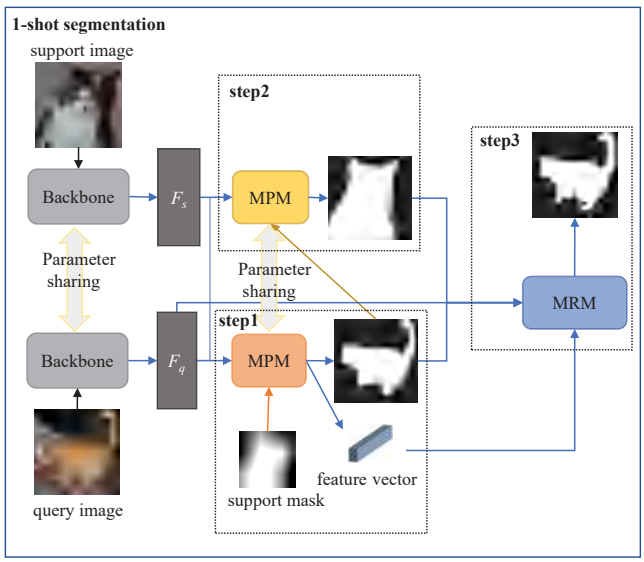
3. THE PROPOSED METHOD

3.1. Overall Architecture

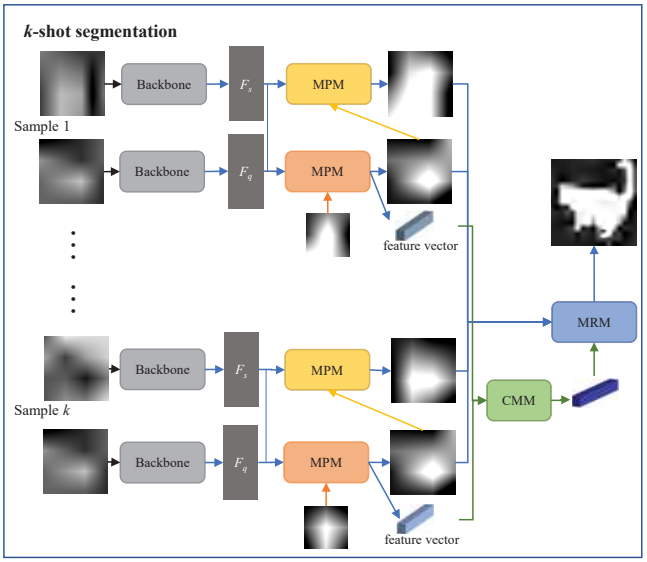
As shown in Fig. 2, we design 4 modules, i.e., backbone, mask predict (MPM), category-modulation (CMM), and mask refine module (MRM). It's noted CMM aligns feature representation among different support images in k-shot learning.

Guided by an active-reference mechanism, we design structure of 1-shot segmentation in Fig. 2(a), where we use different steps and colors to describe the workflow. At first, two parallel backbone networks extract high-level semantic feature maps from input images, represented as F_s and F_q, respectively. Specifically, we use the first four layers of Resnet-50 as backbone network. Afterwards, two MPMs are constructed, where the first MPM predicts query mask q_1 guided by support image as step1, and the second MPM predicts support mask s guided by query mask q_1 as step2. Finally, MRM refines the query mask to output q_2 based on q_1 and s as step3.

ARNet for 1-shot segmentation needs to predict three masks, i.e., q_1, s, and q_2, where the corresponding loss values can be represented as $\mathcal{L}_{q_1}$, $\mathcal{L}_s$ and $\mathcal{L}_{q_2}$. Utilizing cross-entropy loss function, the total loss for ARNet is:



(a) ARNet architecture for 1-shot segmentation



(b) ARNet architecture for 5-shot segmentation

Fig. 2. The pipeline of the proposed ARNet architecture.

$$\mathcal{L} = \alpha\mathcal{L}_{q_1} + \beta\mathcal{L}_s + \gamma\mathcal{L}_{q_2} \tag{1}$$

where α, β, γ are weights of each loss.

The overall architecture of ARNet for k-shot segmentation is shown in Fig. 2 (b). In step3 of k-shot segmentation, we propose CMM to fuse feature vectors extracted from multiple support images, where extrinsic and intrinsic part of original feature is forgotten and enhanced, respectively. In fact, fused and enhanced features would lead to accurate few-shot segmentation results of query images. We extend loss function of 1-shot segmentation to k-shot segmentation as:

$$\mathcal{L} = \alpha\sum_{i=1}^{k}\mathcal{L}_{q_1}^{i} + \beta\sum_{i=1}^{k}\mathcal{L}_{s}^{i} + \gamma\mathcal{L}_{q_2} \tag{2}$$

where $\mathcal{L}^i$ is the loss of the i-th mask.

Since the support image mask is re-predicted in the active-reference mechanism, ARNet can utilize weak labels on support images for few-shot segmentation, such as bounding boxes or inaccurate masks. Guided by the core idea of the proposed active-reference mechanism, ARNet firstly utilizes the entire bounding box area as the support mask to achieve an inaccurate query mask, which is further re-predicted to achieve an accurate support mask. Finally, MRM obtains an accurate target object mask by refining the query mask based on multiple mask images. Compared with pixel-level masks in most previous works, we argue the cost of weakly labeling is much lower in time and labor-consuming.

3.2. Design of Mask Predict Module

We design the proposed mask predict module as Fig. 3. Specifically, feature map of support image F_s computed by backbone is firstly multiplied by the binary support mask M_s to remove background, which outputs feature map containing target objects F_1 as:

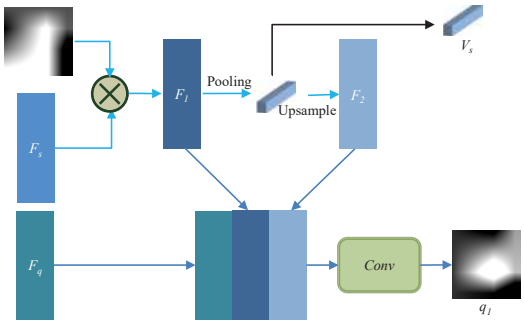


Fig. 3. Structure design of the proposed mask predict module.

$$F_1 = M_s \otimes F_s \tag{3}$$

where $\otimes$ is element-wise multiplication. Then, F_1 is processed by global average pooling $f_p(\cdot)$ to output feature of support image V_s with $V_s = f_p(F_1)$. Based on V_s, we perform up-sampling operation $f_u(\cdot)$ to obtain feature map F_2.

Finally, we concatenate original feature map of query image F_q, F_1, and F_2. Afterwards, we process the concatenated feature with 3×3 convolutional layers to compute the predicted query mask q_1:

$$q_1 = conv(conv(conv(F_q + F_1 + F_2))) \tag{4}$$

where operator + represents concatenate operation, and function $conv(\cdot)$ refers to convolution operation. Unlike most of previous few-shot segmentation methods, we import F_1 for

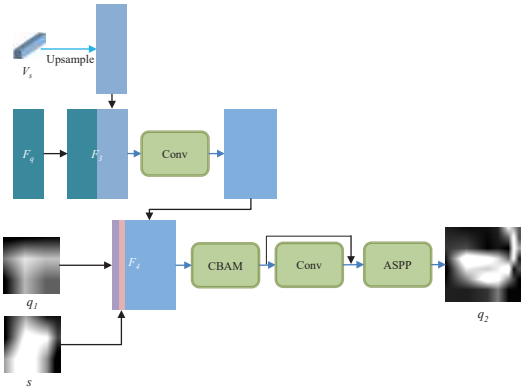


Fig. 4. Structure design of the mask refine module.

feature concatenating, where we argue down-sampling operation to achieve F_2 could lose information of F_1.

3.3. Design of Mask Refine Module

We design the proposed mask refine module as Fig. 4. Specifically, we firstly up-sample V_s given by MPM to the same size of feature map of query image F_q, and then cascades both feature map to output intermediate feature map F_3:

$$F_3 = F_q + f_u(V_s) \tag{5}$$

Afterwards, F_3 is processed by 3×3 convolutions and cascaded with the predicted support mask s and query mask q_1:

$$F_4 = q_1 + s + conv(conv(F_3)) \tag{6}$$

To better describe the internal relationship among different feature channels, we construct Convolutional Block Attention Module (CBAM) to involve spatial and channel-wise attention information. After feature enhancement via attention module, we process feature map via two 3×3 convolution blocks with feature residuals. Finally, we adopt Atrous Spatial Pyramid Pooling module (ASPP) [11] to extract informative information in different scales, thus obtaining the final segmentation results:

$$q_2 = f_A(f_\omega(F_4) \oplus conv(f_\omega(F_4))) \tag{7}$$

where $\oplus$ represents element-wise addition, $f_\omega(\cdot)$ and $f_A(\cdot)$ represent CBAM attention and ASPP module. In k-shot segmentation, MRM collects feature vectors from CMM instead of MPM, where multiple predicted masks of support and query image are paired and cascaded for computing.

3.4. Category-Modulation Module for *k*-shot Segmentation

We propose Category-Modulation Module (CMM), which not only fuses feature vectors extracted from multiple support images, but also builds more intrinsic feature representation by forgetting useless and enhancing contributive information.

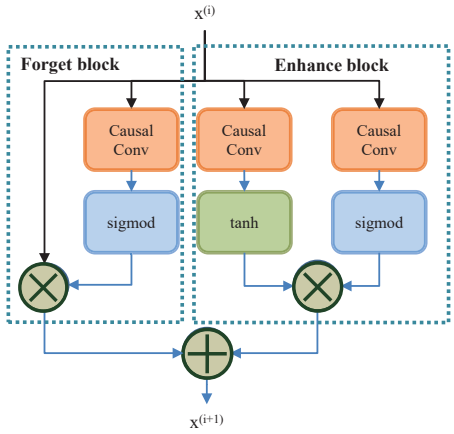


Fig. 5. Structure design of the Forget and Enhance Block.

Specifically, CMM is composed of serially-connected multiple forget and enhance block (FEB), where we show the structure design of FEB in Fig. 5. The FEB consists of forget block and enhance block, where the former forget block forgets category-level extrinsic features, and the latter enhance block strengthens the representation of contributive and intrinsic features. The whole process can be expressed:

$$x^{(i+1)} = (x^{(i)} \otimes \sigma(C(x^{(i)}))) \oplus (tanh(C(x^{(i)})) \otimes \sigma(C(x^{(i)}))) \tag{8}$$

where functions $C(\cdot)$ and $\sigma(\cdot)$ are causal convolution and sigmoid function respectively, $x^{(i)}$ is the input of the i-th FEB, and $\otimes$ stands for element-wise multiplication. It's noted that causal convolution is the core component of FEB. Essentially, the current output only depends on the current and previous input information in causal convolution. When feature vectors of multiple support images are sequentially sent to CMM, causal convolution functions in FEB could make full use of the previous information for feature enhancement.

4. EXPERIMENTS

4.1. Dataset and Evaluation Metric

PASCAL-5^i [1] is a few-shot segmentation dataset , which consists of PASCAL VOC 2012 and additional annotations in SDS with 20 categories. Meanwhile, ***COCO 2014*** [12] is a challenging large-scale dataset containing 80 categories, which is mainly acquired from complex daily scenes. Following [8], we use the mean Intersection-over-Union (mIoU) of all classes and the average of the foreground IoU and background IoU (binary-IoU) to measure model performance.

4.2. Comparison with State-of-the-arts

PASCAL-5^i. Table 1 and Table 2 show 1-shot and 5-shot results, respectively. It can be seen from these two tables that our network is superior to the previous method and achieves a new state-of-the-art performance.

Table 1. Results of 1-shot and 5-shot segmentation on PASCAL-5^i using mean-IOU metric.

Method	1-shot					5-shot					diff
	split-1	split-2	split-3	split-4	Mean	split-1	split-2	split-3	split-4	Mean	Mean
OSLSM[1]	33.6	55.3	40.9	33.5	40.8	35.9	58.1	42.7	39.1	43.9	3.1
SG-One[9]	40.2	58.4	48.4	38.4	46.3	41.9	58.6	48.6	39.4	47.1	0.8
CANet[8]	52.5	65.9	51.3	51.9	55.4	55.5	67.8	51.9	53.2	57.1	1.7
PGNet[10]	**56.0**	**66.9**	50.6	50.4	56.0	57.7	**68.7**	52.9	54.6	58.5	2.5
Ours	54.0	63.9	**55.6**	**52.3**	**56.5**	**57.9**	68.3	**57.2**	**55.7**	**59.8**	**3.3**

Table 2. Results of 1-shot and 5-shot segmentation on PASCAL-5^i using binary-IOU metric.

Method	1-shot	5-shot
OSLSM[1]	61.3	61.5
SG-One[9]	63.9	65.9
CANet[8]	66.2	69.6
PGNet[10]	69.9	70.5
Ours	**72.1**	**73.0**

Table 3. Results of 1-shot and 5-shot segmentation on MS COCO dataset.

Method	mIOU		Binary-IOU	
	1-shot	5-shot	1-shot	5-shot
PANet[2]	20.9	29.7	59.2	63.5
CANet[8]	**49.9**	51.6	-	-
CRNet[3]	45.8	47.2	-	-
Ours	44.5	**52.3**	**68.4**	**70.7**

As shown in Table 1, our approach outperforms previous methods in both 1-shot and 5-shot segmentation. The performance of CANet is improved by 1.7% compared with 1-shot segmentation by fusing multiple feature maps of support images with a learnable attention module. Thanks to the active-reference mechanism, the proposed method has the largest performance gap between 1-shot and 5-shot, which means that the proposed method can greatly utilize multiple supporting information to guide query image segmentation.

MS COCO. The evaluation results of our method on the MS COCO dataset are shown in Table 3. ARNet is at the leading level on the MS COCO dataset. Regarding that PASCAL VOC and MS COCO contain 20 and 80 object categories, the increase in categories leads to higher requirements of models. Therefore, the evaluation result of mean-IOU has been dropped significantly.

Qualitative results. Some qualitative results are shown in Fig. 6. When annotations of support images are of different categories, our method could achieve different segmentation results for the same query image.

4.3. Experiments with Weak Annotations

We use the bounding boxes annotation as weak labels to evaluate the performance of our model further. In the test, we use the area marked by the bounding box in PASCAL-5^i as the mask of the target object in support image. As shown in Table 4, given support images with pixel-wise labels, the results of MPM and MRM both perform well. However, when using a weakly labeled support image, the query mask predicted by MPM is quite different from the ground truth, but the MRM still has an outstanding performance, which means that our active-reference mechanism can withstand the noise introduced by the background area. It removes the wrong guidance information by re-predicting the support mask, and enhances the robustness of support information.

Table 4. Results of MPM and MRM with annotations.

Annotations	MPM	MRM
Pixel-wise labels	51.2	56.5
Bounding box	32.7	54.3

Table 5. Results of our method with different modules.

MPM1	MPM2	MRM	mIOU
√			51.2
√	√		50.1
		√	53.9
√		√	54.7
√	√	√	**56.5**

4.4. Ablation Study

Network Design. We compare by removing certain modules. When removing MRM, we adopt the query mask predicted by MPM as the final result. When removing MPM, we no longer cascade the query mask predicted by MPM with feature map produced by MRM. The feature fusion could be obtained by multiplying support mask and support feature map. As shown in Table 5, where MPM1 is the MPM in step 1 and MPM2 is in step 2, elimination of each module would reduce the performance of the proposed method, representing each module plays a positive role in segmentation.

CMM vs. Attention vs. Mask Fusion. we compared CMM with other k-shot solutions. OSLSM uses the mask fusion method to fuse predicted masks. CANet uses an attention mechanism to fuse support feature map. To quantify the performance of each solution, we apply attention mechanism and mask fusion method to our method respectively, and compare them with the 1-shot result of our method. The experiment results are shown in Table 6, which shows CMM can maximize model improvement.

Fig. 6. Qualitative results produced by ARNet.

Table 6. Comparison of different 5-shot solutions.

Method	mIOU	Increment
1-shot result	56.5	0
Mask Fusion (OSLSM [1])	56.8	0.3
Attention (CANet [8])	57.7	1.2
CMM (ours)	**59.8**	**3.3**

5. CONCLUSION

In this paper, we offer an active-reference mechanism, which not only extracts co-occurrent information from support and query images, but also allows for weakly boundary labeling. Furthermore, we fuse features by category-modulation module to forget useless and enhance contributive information.

acknowledge

This work was supported by National Key R&D Program of China under Grant 2018YFC0407901, the Fundamental Research Funds for the Central Universities under Grant B200202177.

6. REFERENCES

[1] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," in *Proceedings of BMVCs*, 2017.

[2] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *Proceedings of ICCV*, 2019, pp. 9196–9205.

[3] W. Liu, C. Zhang, G. Lin, and F. Liu, "Crnet: Cross-reference networks for few-shot segmentation," in *Proceedings of CVPR*, 2020, pp. 4164–4172.

[4] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *Proceedings of CVPR*, 2019, pp. 10657–10665.

[5] J. Wei, X. Xu, Y. Yang, Y. Ji, Z. Wang, and H. T. Shen, "Universal weighting metric learning for cross-modal matching," in *Proceedings of CVPR*, 2020, pp. 13002–13011.

[6] F. Sung, Y. Yang, L. Z., T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of CVPR*, 2018, pp. 1199–1208.

[7] C. Michaelis, M. Bethge, and A. S. Ecker, "One-shot segmentation in clutter," in *Proceedings of ICML*, 2018, pp. 3546–3555.

[8] C. Zhang, G. Lin, F. Liu, R. Yao, and C. Shen, "Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *Proceedings of CVPR*, 2019, pp. 5217–5226.

[9] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "Sg-one: Similarity guidance network for one-shot semantic segmentation," *CoRR*, vol. abs/1810.09091, 2018.

[10] C. Zhang, G. Lin, F. Liu, J. Guo, Q. Wu, and R. Yao, "Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation," in *Proceedings of ICCV*, 2019, pp. 9586–9594.

[11] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. PAMI*, vol. 40, no. 4, pp. 834–848, 2018.

[12] T. L., M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *Proceedings of ECCV*, 2014, pp. 740–755.