Multiscale Attention-Based Prototypical Network For Few-Shot Semantic Segmentation

Yifei Zhang, Désiré Sidibé, Olivier Morel, Fabrice Meriaudeau

Multiscale Attention-Based Prototypical Network For Few-Shot Semantic Segmentation

Yifei Zhang*‡, Désiré Sidibé†, Olivier Morel*, Fabrice Meriaudeau*
*ERL VIBOT CNRS 6000, ImViA, Université Bourgogne Franche Comté, 71200, Le Creusot, France
†Université Paris-Saclay, Univ Evry, IBISC, 91020, Evry, France
‡Email: yifei.zhang@u-bourgogne.fr

***Abstract*—Deep learning-based image understanding techniques require a large number of labeled images for training. Few-shot semantic segmentation, on the contrary, aims at generalizing the segmentation ability of the model to new categories given only a few labeled samples. To tackle this problem, we propose a novel prototypical network (MAPnet) with multiscale feature attention. To fully exploit the representative features of target classes, we firstly extract rich contextual information of labeled support images via a multiscale feature enhancement module. The learned prototypes from support features provide further semantic guidance on the query image. Then we adaptively integrate multiple similarity-guided probability maps by attention mechanism, yielding an optimal pixel-wise prediction. Furthermore, the proposed method was validated on the PASCAL-5^i dataset in terms of 1-way N-shot evaluation. We also test the model with weak annotations, including scribble and bounding box annotations. Both the qualitative and quantitative results demonstrate the advantages of our approach over other state-of-the-art methods.**

I. Introduction

Despite the undeniable success of deep learning-based methods in various application domains, much research dedicates to exploring advanced technologies in limited-data and challenging scenarios, such as robotics [1], natural language processing [2, 3], and drug discovery in medical applications [4]. Recently, semi-supervised learning [5, 6, 7] has emerged as a hot topic in the computer vision community. Contrary to leveraging a large amount of data, few-shot learning aims to recognize new categories under limited supervision. Especially for few-shot segmentation task, the trained model predicts pixel-level mask of new categories on the query image, given only a few labeled support images. The semantic guidance ability of support images and the generalizability to unseen class may significantly affect the segmentation performance.

Existing methods generally address this problem by learning a set of parameters or prototypes from *support* images and guiding the pixel-wise segmentation on the *query* image. However, most of the previous studies do not explore the support information sufficiently, which is not taking enough advantage of potential semantic information of support images. Usually, they only consider a simple connection between the support set and query set (e.g., cosine similarity), which is adverse to the generalizability. For the above reasons, we propose a novel few-shot segmentation network called multiscale attention-based prototypical network (MAPnet). To fully exploit the representative features from labeled support images, our method extracts rich contextual information via a multiscale feature enhancement module. This module consists of three elaborated branches that aggregate multiscale features of target classes. Multiple learned prototypes provide further similarity-based guidance on the query feature, containing multiscale feature attention. Then we employ the attention mechanism to adaptively weight the probability maps for the final mask prediction. We find that this method effectively strengthens the segmentation model's generalizability, especially for the 5-shot setting. Moreover, the use of attention-based gating accelerates the convergence to a lower loss. The network was trained in an end-to-end manner without any post-processing steps. Figure 1 illustrates the overall workflow of our MAPnet.

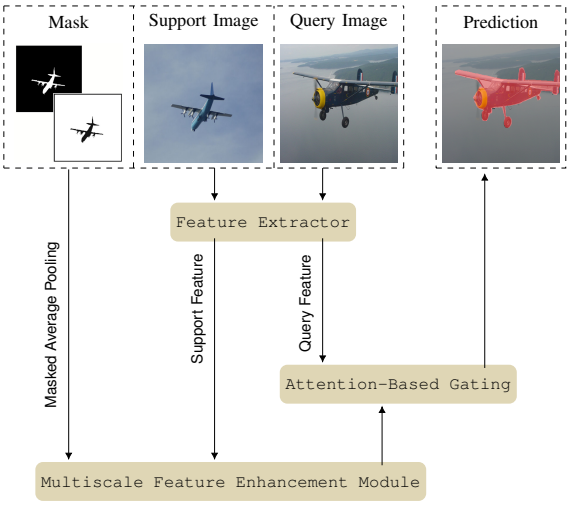

Fig. 1. An overview of the proposed method (MAPnet). Given a query image of a new category, e.g., aeroplane, the goal of few-shot segmentation is to predict a mask of this category regarding only a few labeled samples.

Overall, this paper makes the following contributions:

- We propose a novel few-shot segmentation method based on the prototypical network.
- We develop a multiscale feature enhancement module to fully exploit the support features. The learned prototypes provide further semantic guidance on the query features.
- We apply the attention mechanism to fuse multiple probability maps for an optimal pixel-wise prediction.

- Extensive experiments demonstrate the effectiveness of the proposed method. Notably, our model achieves the mean IoU score of 56.0% on PASCAL-5^i for the 1-way 5-shot setting, which gains a remarkable improvement of 5.1% in the binary IoU comparing to 1-way 1-shot.

The remainder of this paper is organized as follows. Section II reviews the state of the art for few-shot segmentation. Section III describes the problem definition and the proposed architecture in detail. Section IV reports the extensive experimental results on the PASCAL-5^i dataset as well as the evaluations with weak annotations. Conclusions are drawn in Section V.

II. Related work

a) Semantic segmentation: In the early stage, Fully Convolutional Network (FCN) [8], as the CNN-based network, was firstly applied to tackle the semantic segmentation challenge. Then plenty of classical segmentation networks have appeared, such as SegNet [9], U-Net [10], RefineNet [11], the series of DeepLab [12, 13, 14]. Dilated convolutions are introduced to extract larger contextual information without reducing the image resolution. So far, the computer vision community has published a large number of segmentation algorithms with high performance. As one of the most powerful concepts in the deep learning techniques, the attention mechanism has been broadly used in image segmentation tasks, such as [15, 16, 17]. The attention distribution can enable the model to selectively pick valuable information [18, 19]. We adopt the attention mechanism in our work, aiming to adaptively fuse the multiscale attention of query image in semi-supervised segmentation.

b) Few-shot learning for segmentation: Many approaches for few-shot learning are proposed to generalize prior knowledge to new tasks using only a few examples [5, 7]. Some research [2, 20] introduced the metric learning-based matching network for the few-shot classification task. The non-parametric structure facilitates the generalization of models to new training sets. Several studies, such as [21], have focused on the graph-based methods for few-shot learning. Moreover, Snell et al. [22] presented a method to represent the prototypes per class in a representation space, known as prototypical networks.

Few-shot semantic segmentation refers to the pixel-level prediction of new categories on the query set, given only a few labeled support images. Shaban et al. [23] first presents a dual branch parallel network for semi-supervised semantic segmentation, known as OSLSM, including a conditioning branch and a segmentation branch. The conditioning branch extracts representative high-level features from the supporting image-label pair, while the segmentation branch integrates the parameters learned from the conditioning branch and produces a segmentation mask on the query image. Other variants of OSLSM include Co-FCN [24], PL+SEG [25] and MDL [26]. All of which extend such a dual branch structure to achieve substantial performance improvement. In the AMP model, Siam et al. [27] replaces the guidance branch with a multi-resolution weight. Hu et al. [28] elaborated an attention-based multi-context guiding network, which sequentially integrates the support features to elevate the segmentation accuracy. Besides, SG-One [29] presented a Masked Average Pooling block (MAP) to extract the representative vectors of support objects. Then the segmentation mask was predicted via a similarity guidance network. More recently, Wang et al. [30] introduced a novel prototype alignment network, called PANet. A prototype alignment regularization was employed to encourage the model to perform the few-shot segmentation in a reverse direction. In this paper, we extend the work of prototype learning with multiscale feature attention. The proposed architecture effectively strengthens the generalizability and discriminating ability of the segmentation model.

III. Methodology

A. Problem setting

The primary motivation of the few-shot segmentation is to develop a segmentation model with high generalizability. Given only one or a few examples, the model can produce pixel-level prediction with sufficient accuracy on a new category. Usually, few-shot learning is considered as a N-way-K-shot classification task that discriminates between N classes with K examples per category.

In this paper, we adopt the problem definition proposed in [23, 27, 30]. Suppose there are two semantic class sets L_{train} and L_{test}, the few-shot segmentation model deals with a dataset $D = \{D^{train}, D^{test}\}$ where D^{train} and D^{test} are composed of image samples including at least one pixel belonging to L_{train} and L_{test}, respectively. The training set, which contains N_{train} image samples, can be defined as $D^{train} = (X^i, Y(l)^i)_i^{N_{train}}$ where X^i denotes the i^{th} training image and $Y(l)^i$ is the corresponding segmentation mask of class l. The test set is given as $D^{test} = (X^i, Y(l)^i)_i^{N_{test}}$. It is important to note that the model is tested on new semantic classes that do not belong to the training set, i.e. $L_{train} \cap L_{test} = \emptyset$.

Both the training and test sets contain several *episodes* that consist of a set of labeled support images $S = (x_s^i, y(l)_s^i)_{i=1}^K$ and a set of query images $Q = (x_q, y_q(l))$ where l is the semantic class. The support set comprises K labeled examples for each of the N classes, which defines a N-way-K-shot segmentation. During training, $episodes = (S, Q)$ randomly sampled from D^{train} are used to perform segmentation on the query set, namely $\hat{y_q} = (S, x_q)$. The performance is measured by a loss function $loss(\hat{y_q}, y_q)$ where y_q is the corresponding segmentation mask. Therefore, the optimal parameters of few-shot segmentation model are $\theta^* = \text{argmin}_\theta \, loss(\hat{y_q}, y_q)$. While testing, the model is given a set of labeled support images sampled from D^{test}. Then the few-shot segmentation model is expected to predict the segmentation mask on the query image for the relative class. By taking advantage of prior knowledge, the model can rapidly generalize for a new class of limited supervised information.

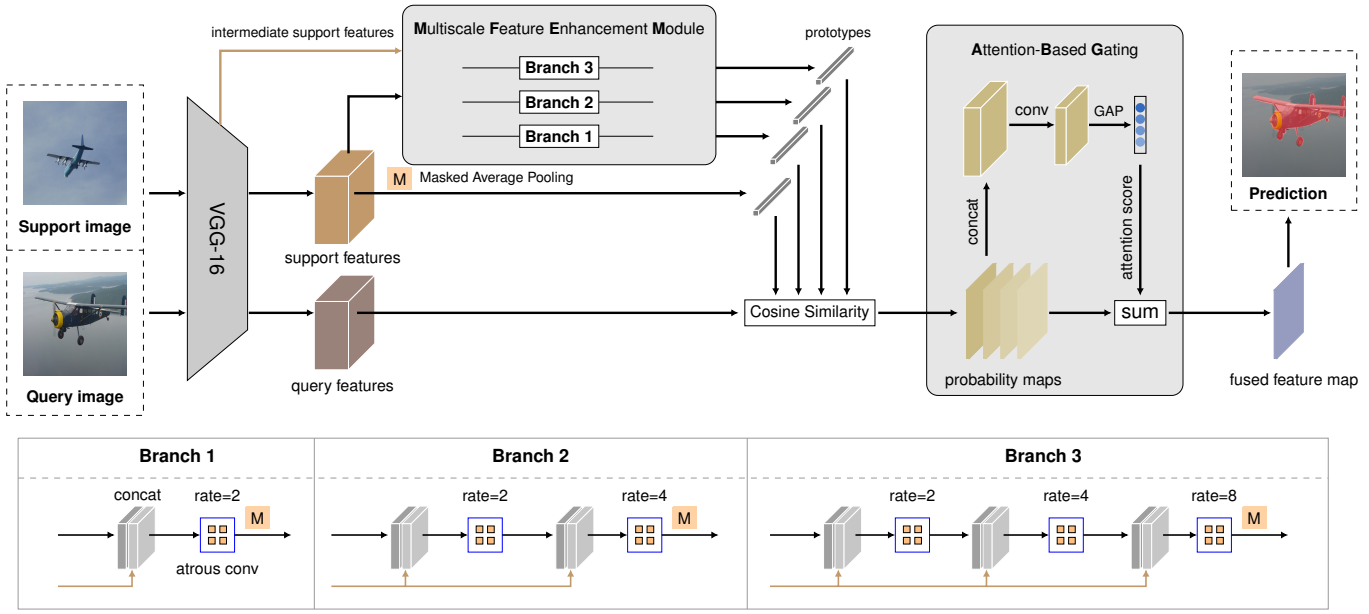

Fig. 2. Illustration of the proposed method (MAPnet) for few-shot semantic segmentation.

B. Proposed method

In this subsection, we present our few-shot segmentation method (MAPnet) in detail. The overall structure of the proposed method is shown in Figure 2.

1) Overview: The proposed MAPnet is based on the prototypical network. The prototype learning-based methods enable the network to learn a set of feature vectors with adequate discriminative information. In the early work of [22], researchers proposed a prototypical network that learns a common metric space. Few-shot classification can be achieved by computing distances to prototype representations of each class. However, such methods do not explore potential semantic information of support images in sufficient depth. The learned prototypes provide limited semantic guidance on the query feature, constraining the segmentation model's generalizability. Therefore we adopt the idea of prototype learning and introduce a novel few-shot segmentation method with multiscale feature attention.

In general, our method contains a feature extractor, a multiscale feature enhancement module, and an attention-based gating (see Figure 2). The support images and query images are embedded into a high-level feature space via a shared feature extractor. Concerning the practical implementation, we employ the first five convolutional blocks of VGG-16 [31] as the backbone network. The convolutions in the fifth convolutional blocks are replaced by atrous convolutions with a rate of 2. Besides, we retain the third convolutional block's output as intermediate features for further multiscale feature enhancement.

2) Multiscale Feature Enhancement Module: Generally, each category that appeared in the input images differs in shape and size. The uniscale filters learned by the neural network may lead to many restrictions on the similarity-guided semantic guidance. Thus, we define a multiscale feature enhancement module (MFE module) to provide multiscale feature supervision. In consideration of the trade-off between high performance and computational cost, we elaborate three branches in MFE module. Each branch takes the intermediate support features and high-level support features as input. The support features are resized and concatenated for providing effective feature aggregation. This module enables the few-shot segmentation network more expressive as the model becomes deeper and wider. Empirically, we employ multiscale atrous convolution with rates $r = 2, 4, 8$ to preserve more spatial and contextual information.

In order to enhance the discriminative power of the model, we leverage both foreground and background information of support images, known as $y(l)(+,-)$, to extract the representative prototypes of target classes l. The background information provides complementary clues for semantic understanding. Masked Average Pooling [29] is used in the process to compute these feature vectors. The set of feature vectors can be defined as $V = \{v_0(+,-), v_1(+,-), ..., v_n(+,-)\}$ where $v_n(+,-)$ is the n_{th} pair of support feature vectors. Each pair of prototypes is used to generate the corresponding probability map with multiscale feature attention.

3) Attention-Based Gating: Different from existing few-shot segmentation methods, our model produces a set of similarity-guided probability maps by estimating the distance between a series of representative prototypes and the high-level query features. Following the work in [29, 30], we employ the cosine similarity as the non-parametric nearest neighbor classifier. Thus we employ the attention-based gating block as a combination strategy to generate an optimal mask prediction. Suppose that the concatenation of probability maps is P, the attention score g can be defined as $g = softmax(w * P)$,

TABLE I
RESULTS OF 1-WAY 1-SHOT AND 1-WAY 5-SHOT SEMANTIC SEGMENTATION ON PASCAL-5^i USING MEAN-IOU(%) METRIC. THE RESULTS OF 1-NN AND LOGREG ARE REPORTED BY [23].

Methods	1-shot					5-shot				
	Pascal-5^0	Pascal-5^1	Pascal-5^2	Pascal-5^3	Mean	Pascal-5^0	Pascal-5^1	Pascal-5^2	Pascal-5^3	Mean
1-NN	25.3	44.9	41.7	18.4	32.6	34.5	53.0	46.9	25.6	40.0
LogReg	26.9	42.9	37.1	18.4	31.4	35.9	51.6	44.5	25.6	39.3
OSLSM [23]	33.6	55.3	40.9	33.5	40.8	35.9	58.1	42.7	39.1	43.9
co-FCN [24]	36.7	50.6	44.9	32.4	41.1	37.5	50.0	44.1	33.9	41.4
SG-One [29]	40.2	**58.4**	48.4	38.4	46.3	41.9	58.6	48.6	39.4	47.1
PANet [30]	42.3	58.0	**51.1**	41.2	48.1	**51.8**	64.6	**59.8**	46.5	55.7
MAPnet	**42.9**	58.3	48.8	**42.6**	**48.2**	51.6	**65.1**	58.4	**48.8**	**56.0**

TABLE II
RESULTS OF 1-WAY 1-SHOT AND 1-WAY 5-SHOT SEGMENTATION ON PASCAL-5^i USING BINARY-IOU(%) METRIC. Δ DENOTES THE DIFFERENCE BETWEEN 1-SHOT AND 5-SHOT.

Mehtods	1-shot	5-shot	Δ
co-FCN [24]	60.1	60.2	0.1
OSLSM [23]	61.3	61.5	0.2
MDL [26]	63.2	63.7	0.5
PL+SEG [25]	61.2	62.3	1.1
AMP-2+FT [27]	62.2	63.8	1.6
SG-One [29]	63.1	65.9	2.8
PANet [30]	66.5	70.7	4.2
MAPnet	**66.7**	**71.8**	**5.1**

TABLE III
TRAINING AND EVALUATION ON PASCAL-5^i DATASET USING 4-FOLD CROSS-VALIDATION, WHERE i DENOTES THE NUMBER OF SUBSETS.

Dataset	Test classes
Pascal-5^0	aeroplane, bicycle, bird, boat, bottle
Pascal-5^1	bus, car, cat, chair, cow
Pascal-5^2	diningtable, dog, horse, motorbike, person
Pascal-5^3	potted plant, sheep, sofa, train, tv/monitor

where w denotes a convolutional layer and global average pooling layer (GAP) [32]. Then our network learns the convolutional kernels ρ over the fused probability map. More formally,

$$\hat{y_q} = softmax[\rho * \sum_{i}^{I}(g_i \cdot p_i)] \quad (1)$$

where g_i and p_i denote the i_{th} attention score and probability map, respectively.

IV. EXPERIMENTS

A. Setup

a) Dataset: We evaluate the proposed method on the PASCAL-5^i dataset, which derives from PASCAL VOC 2012 [33] with SBD [34] augmentation. This dataset was firstly created by Shaban et al. [23], then widely used in the few-shot segmentation task. Similar to the setup of OSLSM [23], we sample 5 classes out of all 20 categories as test label-set $L_{test} = \{5i + 1, ..., 5i + 5\}$ with i being the folder number. The remaining 15 classes form the train label-set L_{train}. As shown in Table III, our model is trained on three splits, then tested on the rest one in a cross-validation manner. In this work, we evaluate the performance of our model on 1,000 randomly sampled episodes for each folder.

b) Implementation details: We conduct the experiments with implementations in PyTorch [35]. The backbone network (i.e., VGG-16) was initialized with pre-trained weights on ImageNet [36]. We resized the input images to 320×320 with random horizontal flipping. All the few-shot segmentation models were trained on a single Nvidia TITAN Xp GPU with 12GB memory, using stochastic gradient descent (SGD) with a batch size of 1, a momentum of 0.9, and weight decay of 0.0005 for a maximum of 40,000 iterations. The initial learning rate was set to 1e-3 and reduced by 0.1 every 10,000 iterations.

c) Evaluation metrics: Following the previous works on the few-shot segmentation [23, 29, 30], we apply two standard metrics to evaluate the performance of learning models: mean-IoU and binary-IoU. Generally, the mean Intersection-over-Union (mean-IoU) is used to measure each foreground class's accuracy and average over all the categories. Binary-IoU deals uniformly with all object categories as one foreground class and averages the IoU of both foreground and background. Based on these two metrics, we can fairly compare the accuracy in terms of 1-way N-shot semantic segmentation.

B. Comparison

Table I shows the comparison result of our method MAPnet and other previous methods in terms of 1-way 1-shot and 1-way 5-shot segmentation. We observe that our model achieves 48.2% on the whole for the 1-way 1-shot setting, which substantially outperforms the baseline network OSLSM by +7.4%. Also, the performance of MAPnet is competitive to the state-of-the-art method PANet. Our model earns the largest gain of +1.4% on PASCAL-5^3 compared to PANet, where the test classes are *potted plant*, *sheep*, *sofa*, *train* and *tv/monitor*. Our method yields a mean IoU of 56.0% overall with five support images, which achieves significant improvement over other baseline networks. Compared to SG-One, which has a simple but effective prototypical structure, we can see that the proposed method leads to a relative improvement of +9.7% on PASCAL-5^0, +6.5% on PASCAL-5^1, +9.8% on PASCAL-5^2, and +9.4% on PASCAL-5^3.

Besides, we report the averaged binary-IoU on the four-fold cross-validation in Table II. Our method shows remarkable improvement in 1-way 5-shot, which gains an increment of

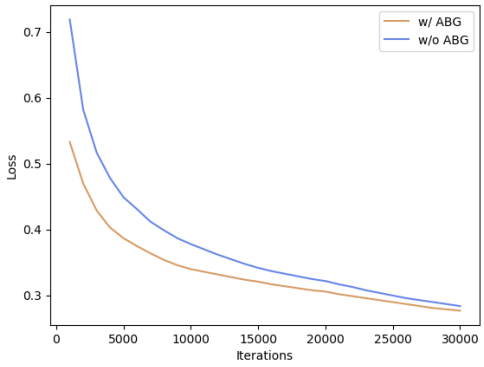

Fig. 3. Training loss of models with and without attention-based gating (ABG) for 1-way 1-shot segmentation on PASCAL-5^0.

TABLE IV
EVALUATION RESULTS OF USING DIFFERENT TYPES OF ANNOTATIONS IN MEAN-IOU(%) METRIC.

Methods	1-shot			5-shot		
	Dense	Scribble	Bbox	Dense	Scribble	Bbox
PANet [30]	48.1	44.8	45.1	55.7	54.6	52.8
MAPnet	48.2	44.1	45.7	56.0	53.5	53.7

5.1% comparing to 1-way 1-shot. The main reason behind the increase of accuracy is that our multiscale feature enhancement module provides richer contextual information for the semantic guidance of target classes. Namely, the attention-based multiscale feature aggregation becomes more prominent as the number of support images increases. Moreover, we replace the attention-based gating with element-wise addition, aiming to compare the convergence speed of our method trained with and without multiscale attention. As shown in Figure 3, we observe that the attention-based gating speeds up the convergence and reaches a lower loss, which also earns a 3.5% gain in accuracy.

Furthermore, we demonstrate qualitative results on PASCAL-5^i in Figure 4. We present different cases involving outdoor scenes and indoor scenes. These examples show the high discriminatory power and generalizability of our method. It is capable of extracting sufficient contextual information of the target class, and a challenging case is shown in Figure 4 row 4. We also present two typical failure cases in our experiments. Based on the observation of the first failure case, we find that it is not easy to distinguish the objects with similar characteristics, especially when these objects are placed in an overlapping manner. Another failure case shows that the model has a limited capacity to recognize irregular objects and their boundary delineation. These failure cases may be challenging issues in future work.

C. Test with weak annotations

We report the experimental results on different weak annotations in Table IV, including scribbles and bounding box annotations. Different from tedious and inefficient per-pixel annotating, these weak annotations are frequently used in interactive image segmentation [37]. In our experiments, the pixel-wise masks of support images are replaced by the corresponding weak annotations at the test time. In general, our model's performance using weak annotations is comparable to the result with pixel-wise annotations, indicating the robustness of MAPnet. We also observe that using bounding box annotations achieves higher accuracy than using scribble annotations. A potential reason could be that our method learns more representative prototypes within the valid region of the bounding box. Figure 5 shows some qualitative examples of the segmentation results.

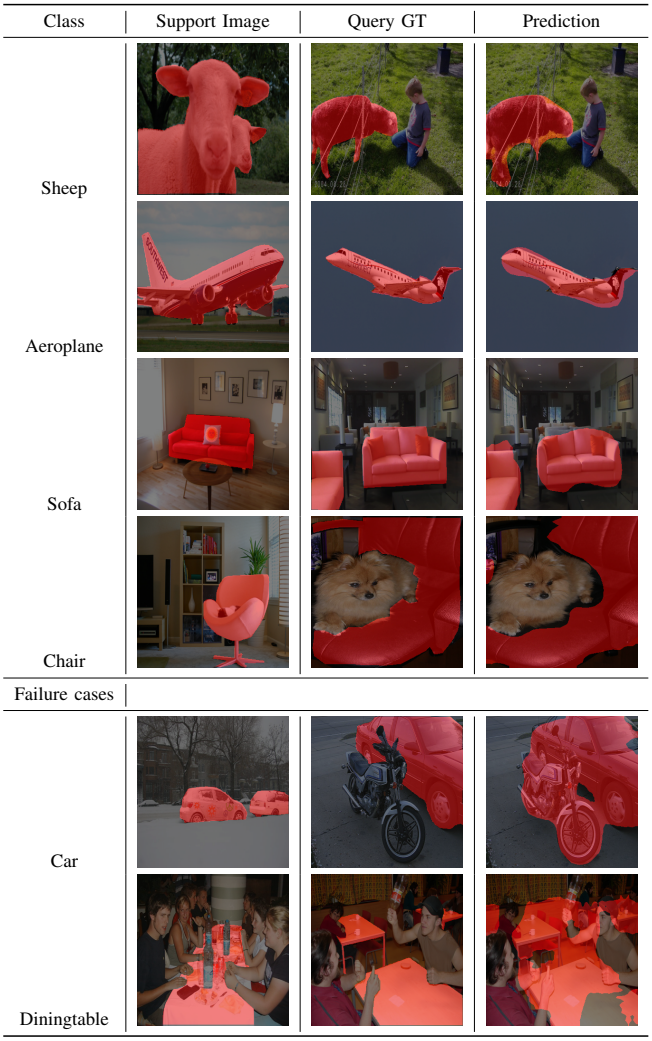

Fig. 4. Qualitative results of our method for 1-way 1-shot segmentation on PASCAL-5^i.

V. CONCLUSION

This work has presented MAPnet, a novel few-shot segmentation method based on the prototypical network. The proposed method provides effective semantic guidance on the query feature by a multiscale feature enhancement module. We elaborate the branches in this module to fully exploit the support information. Moreover, we employ the attention mechanism on the similarity-guided probability maps to produce

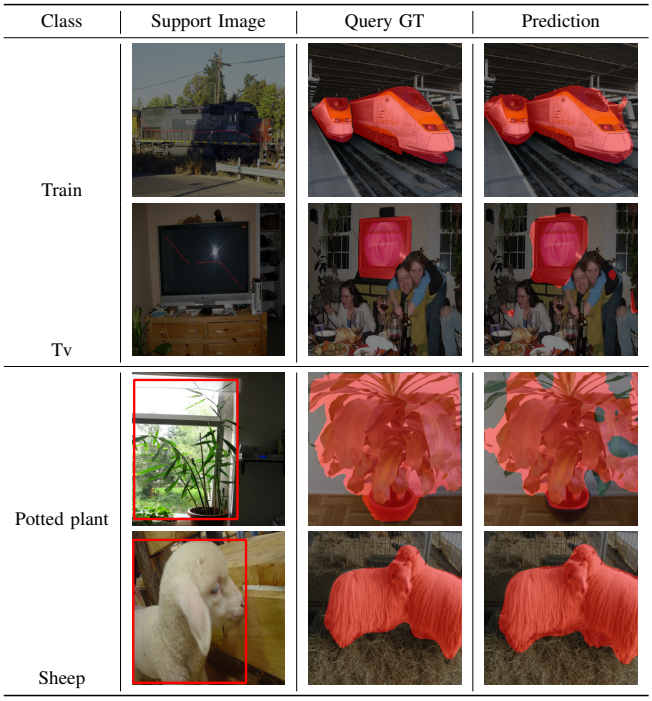

Fig. 5. Qualitative results of our model using scribble and bounding box annotations for 1-way 5-shot setting. The chosen example in support images shows the annotation types.

an optimal pixel-wise prediction, which also speeds up the convergence. Extensive experiments demonstrate the improved generalizability and discriminating ability of the proposed method. Our model achieves a comparable accuracy with the state-of-the-art, outperforming most of the previous methods.

Acknowledgment

The authors would like to acknowledge the French ANR project ICUB (ANR-17-CE22-0011) for its financial support as well as a hardware grant from NVIDIA.

References

[1] Y. Duan, M. Andrychowicz, B. C. Stadie, J. Ho, J. Schneider, I. Sutskever, P. Abbeel, and W. Zaremba, "One-shot imitation learning," 2017.

[2] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," in *Advances in neural information processing systems*, 2016, pp. 3630–3638.

[3] ukasz Kaiser, O. Nachum, A. Roy, and S. Bengio, "Learning to remember rare events," 2017.

[4] H. Altae-Tran, B. Ramsundar, A. S. Pappu, and V. Pande, "Low data drug discovery with one-shot learning," *ACS central science*, vol. 3, no. 4, pp. 283–293, 2017.

[5] Li Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[6] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1126–1135.

[7] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *In International Conference on Learning Representations (ICLR)*, 2017.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 12, pp. 2481–2495, 2017.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[11] G. Lin, A. Milan, C. Shen, and I. D. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," *CoRR*, vol. abs/1611.06612, 2016. [Online]. Available: http://arxiv.org/abs/1611.06612

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[13] 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," *CoRR*, vol. abs/1606.00915, 2016. [Online]. Available: http://arxiv.org/abs/1606.00915

[14] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *arXiv preprint arXiv:1802.02611*, 2018.

[15] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.

[16] M. Ren and R. S. Zemel, "End-to-end instance segmentation and counting with recurrent attention," *CoRR*, vol. abs/1605.09410, 2016. [Online]. Available: http://arxiv.org/abs/1605.09410

[17] Y. Li, X. Chen, Z. Zhu, L. Xie, G. Huang, D. Du, and X. Wang, "Attention-guided unified network for panoptic segmentation," *CoRR*, vol. abs/1812.03904, 2018. [Online]. Available: http://arxiv.org/abs/1812.03904

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *CoRR*, vol. abs/1709.01507, 2017. [Online]. Available: http://arxiv.org/abs/1709.01507

[20] X. Li, L. Yu, C.-W. Fu, M. Fang, and P.-A. Heng, "Revisiting metric learning for few-shot image classification," *ArXiv*, vol. abs/1907.03123, 2019.

[21] V. Garcia and J. Bruna, "Few-shot learning with graph neural networks," *arXiv preprint arXiv:1711.04043*, 2017.

[22] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, 2017, pp. 4077–4087.

[23] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," *arXiv preprint arXiv:1709.03410*, 2017.

[24] K. Rakelly, E. Shelhamer, T. Darrell, A. Efros, and S. Levine, "Conditional networks for few-shot semantic segmentation," 2018.

[25] N. Dong and E. Xing, "Few-shot semantic segmentation with prototype learning." in *BMVC*, vol. 3, no. 4, 2018.

[26] Z. Dong, R. Zhang, X. Shao, and H. Zhou, "Multi-scale discriminative location-aware network for few-shot semantic segmentation," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2. IEEE, 2019, pp. 42–47.

[27] M. Siam, B. Oreshkin, and M. Jagersand, "Adaptive masked proxies for few-shot segmentation," *arXiv preprint arXiv:1902.11123*, 2019.

[28] T. Hu, P. Yang, C. Zhang, G. Yu, Y. Mu, and C. G. Snoek, "Attention-based multi-context guiding for few-shot semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8441–8448.

[29] X. Zhang, Y. Wei, Y. Yang, and T. Huang, "Sg-one: Similarity guidance network for one-shot semantic segmentation," *arXiv preprint arXiv:1810.09091*, 2018.

[30] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9197–9206.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[32] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[33] 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.

[34] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 991–998.

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[37] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," *CoRR*, vol. abs/1604.05144, 2016. [Online]. Available: http://arxiv.org/abs/1604.05144