DUAL-ATTENTION NETWORK FOR FEW-SHOT SEGMENTATION

Zhikui Chen, Han Wang, Suhua Zhang, Fangming Zhong**

School of Software Technology, Dalian University of Technology, Dalian, Liaoning 116620, China
zkchen, fmzhong@dlut.edu.cn

ABSTRACT

Few-shot segmentation aims at segmenting target object areas with only a few labeled samples. Previous methods extract class-specific prototypes to guide segmentation. However, using one or more prototypes to represent the whole object inevitably drops vital spatial information, ignoring many details in original images. To address the issue, we propose a Dual-Attention Network (DANet) for few-shot segmentation. Firstly, a light-dense attention module is proposed to set up pixel-wise relations between feature pairs at different levels to activate object regions, which can leverage semantic information in a coarse-to-fine manner. Secondly, in contrast to the previous prototype-based methods that offer a holistic representation for each object class, we propose a prototypical channel attention module which incorporates channel interdependencies to enhance the discriminative capacity of features. The extensive experiments on two benchmarks show that our approach outperforms the state-of-the-arts in most cases.

***Index Terms*—** Few-shot learning, Semantic segmentation, Attention mechanism

1. INTRODUCTION

Driven by sufficient image datasets and pixel-level mask annotations, semantic segmentation has made substantial progress with deep neural networks. However, precise mask annotations for large-scale datasets, such as biomedical and land use domains, are costly to obtain. Besides, conventional segmentation approaches mostly perform well on seen object classes but could not generalize to novel classes.

Few-shot segmentation aims at segmenting novel classes based on the transferable knowledge from a limited number of seen class samples. The main challenging problem is how to incorporate sufficient class-specific semantic cues to facilitate segmentation tasks. Current few-shot segmentation methods [1–4] usually employ a two-branch encoder-decoder architecture to extract features from both query and support images, and then adopt different methods to generate segmentation probability maps. Class-based methods [3, 5, 6] squeeze all foreground regions to produce class representations and guide pixel-level classifications by computing cosine similarity or concatenating features to aggregate semantic information. Cluster-based methods [7–9] adopt expectation maximization algorithm, superpixel algorithm or K-means clustering to generate part-aware representations. However, due to the large variations in the same category, using one or several prototypes to stand for the whole objects inevitably drops some vital spatial information, poses or textures. In addition, these methods ignore the relations between original support-query feature pairs, thus leading to suboptimal results.

To tackle the aforementioned limitations, in this paper, we propose a Dual-Attention Network (DANet) for few-shot semantic segmentation, which integrates attention mechanisms with prototype learning. Specifically, we propose a Light-dense Attention module, which establishes region correspondence at different levels. Through constructing attention maps in horizontal and vertical directions, the light unit makes preliminary comparison for larger regions. Simultaneously, the dense one makes elaborate comparison by setting up pixel-level correspondence. Such duplex comparisons can explore the coarse-to-fine semantic correlations to activate object category area of query set more accurately. In addition, a Prototypical Channel Attention module is introduced to exploit the inter-channel relations to activate channels with higher response. Finally, the outputs of these two attention modules are integrated together for mask generation. Extensive experiments are conducted on Pascal-5^i [1] and FSS-1000 [10] datasets, and the results indicate that DANet outperforms the baselines in most cases, especially under 1-shot setting with the mIoUs of **61.88%** and **82.6%**, respectively.

The contributions of our work are:

- A Light-dense Attention module is proposed for few-shot segmentation that explores the semantic correlations in a coarse-to-fine manner to activate object category area.
- We propose a Prototypical Channel Attention module to further activate limited channels with higher response, which can provide a superior class representation.

*Corresponding author. This work is supported by the National Natural Science Foundation of China (62076047,62006035), the Dalian Science and Technology Innovation Foundation (2021JJ12SN44), and the Fundamental Research Funds for the Central Universities (DUT20LAB136).

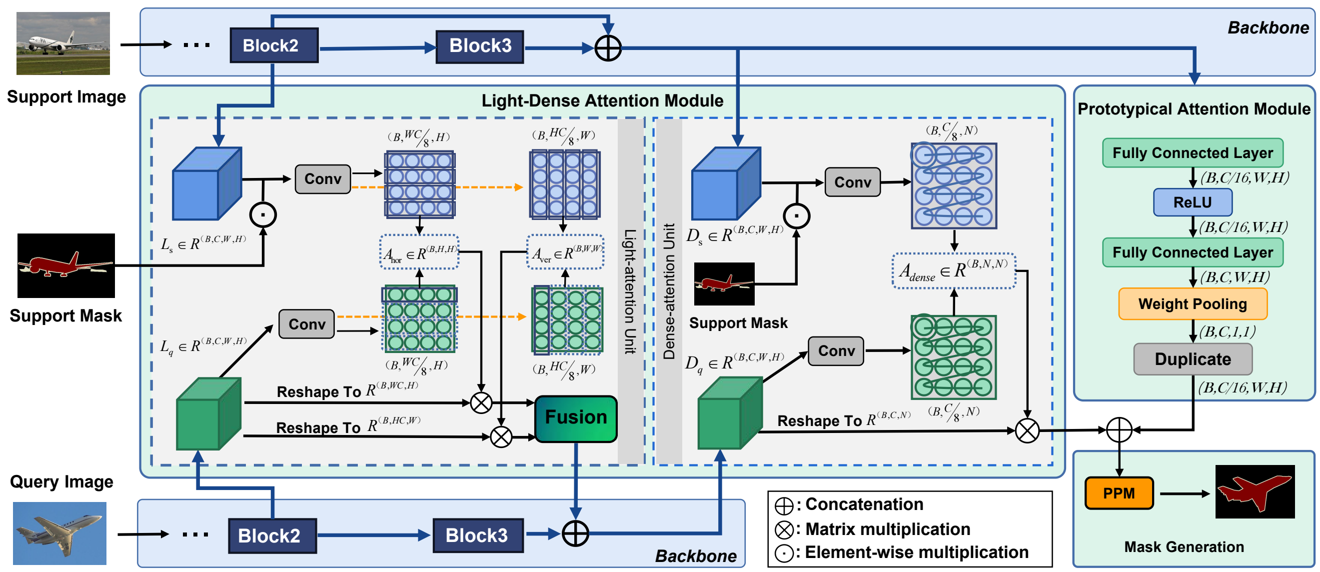

Fig. 1. Framework of the proposed DANet which consists of two attention modules.

2. THE PROPOSED METHOD

2.1. Problem Setting

The aim of few-shot segmentation is to segment objects with a few annotated images, where two image sets are usually provided, i.e., D_{train} and D_{test}. We train the segmentation model on D_{train} and evaluate the model on D_{test}, and we have $D_{train} \cap D_{test} = \varnothing$. Following [1], we align the training and testing for few-shot segmentation with the episodic paradigm. Training set D_{train} and testing set D_{test} both consist of a certain number of episodes, each of which contains a support set S and a query set Q. The goal of our approach is to learn from S and Q of D_{train} and then generalize to D_{test} given a few labeled S samples of it.

2.2. Overview

The overall framework of our DANet is shown in Fig. 1. DANet integrates attention mechanisms with prototype learning to facilitate few-shot segmentation, where two attention modules are designed. Specifically, we first extract query and support feature maps simultaneously. Then, the light-dense attention module takes middle-level features (block2 and block3) as input and builds regions correspondence at different levels. This significantly enhances the local query features which contribute more to support features. In terms of the prototypical attention module, we introduce a weighted gating mechanism which amplifies the prototype representation by exploiting inter-channel relations. Finally, we integrate the output of two modules and feed it into the Pyramid Pooling Module (PPM) [11] to further generate the probability map, i.e, predicted mask.

2.3. Light-dense Attention

As can be seen in Fig. 1, the light-dense attention module consists of a light-attention unit and a dense-attention unit. The light-attention unit collects local region relations information in horizontal and vertical directions to reduce the computation load of attention map while the dense-attention unit focuses on establishing pixel-wise correspondence.

Light Attention unit. As previous few-shot segmentation works, we take a ResNet-50 [9] as the feature extractor. The images I_s and I_q are fed into the extractor to obtain the feature maps L_s and L_q, where $\{L_s, L_q\} \in \mathbb{R}^{B\times C\times H\times W}$. The background regions of L_s are filtered out with support masks. We select the middle-level features from the extractor as the initial input, i.e., block2 of ResNet-50. The process can be formulated as

$$L_s = \mathcal{F}(I_s) \odot M, L_q = \mathcal{F}(I_q) \qquad (1)$$

where $\mathcal{F}$ indicates the extractor, M is the binary mask and $\odot$ indicates the operation of element-wise multiplication.

To build attention coefficients between support and query set, we feed L_s and L_q into two 1×1 convolutional layers to generate the local features F_s and F_q, respectively, where $\{F_s, F_q\} \in \mathbb{R}^{B\times \frac{C}{8}\times H\times W}$. In horizontal direction, we treat each row of local features as part of objects and then measure similarity of the same row elements from F_s and F_q. To split features into H groups, we directly reshape F_s and F_q to $\mathbb{R}^{B\times \frac{WC}{8}\times H}$. The similarity attention map $A_{hor} \in \mathbb{R}^{B\times H\times H}$ can be denoted as:

$$A_{hor}(ij) = exp(F_q^{\mathrm{T}}(i)\cdot F_s(j)) / \sum_{j=1}^{H} exp(F_q^{\mathrm{T}}(i)\cdot F_s(j)) \quad (2)$$

where $F_s(i)$ and $F_q(j)$ indicate the i^{th} and j^{th} deep pixel of local features. In addition, we further reshape F_s and F_q to

$\mathbb{R}^{B\times \frac{HC}{8}\times W}$ and obtain the attention map $A_{ver} \in \mathbb{R}^{B\times W\times W}$ according to Eq. (2). Correlation attention A_{hor}, A_{ver} reflect the degree of influence of different support regions. To propagate the support semantic cues, we put the query feature L_q in another 1×1 convolutional layer to achieve its local feature $F_v \in \mathbb{R}^{B\times C\times H\times W}$. Then, we reshape F_v to $F_{v1} \in \mathbb{R}^{B\times WC\times H}$ and $F_{v2} \in \mathbb{R}^{B\times HC\times W}$. Finally the output of light attention unit can be computed as:

$$F_{light} = \phi(\alpha F_{v1} A_{ver}^{\mathrm{T}}) + \psi(\beta F_{v2} A_{hor}^{\mathrm{T}}) + F_v \qquad (3)$$

where F_{light} is the output feature of light attention unit, ϕ and ψ are the matrix alignment transformations, α and β are scalar values.

Dense Attention unit. We take another intermediate layer output $\{L_s^h, L_q^h\} \in \mathbb{R}^{B\times C\times H\times W}$ of the extractor as input. i.e., block3 of ResNet-50. The Light Attention unit output F_{light} and query feature L_q^h are concatenated and then fed into a convolutional layer for dimension reduction, the process can be formulated as:

$$D_s = \mathcal{F}_{1\times1}(L_s \oplus (L_s^h \odot M)), D_q = \mathcal{F}_{1\times1}(F_{light} \oplus L_q^h) \quad (4)$$

where $\mathcal{F}_{1\times1}$ indicates the 1×1 convolutional layer, M is the binary mask, $\oplus$ is the concatenation operation and $\odot$ is the element-wise multiplication operation. Note that L_s^h and L_q^h are expanded to the same size with L_s following [6].

Different from light attention unit, here we leverage attention mechanism inspired by a standard non-local block [12] to establish pixel-level correspondence. Specifically, we feed D_s and D_q into two 1×1 convolutional layers to generate the local features F_s and F_q, where $\{F_s, F_q\} \in \mathbb{R}^{B\times \frac{C}{8}\times H\times W}$. Then we reshape F_s, F_q to $\mathbb{R}^{B\times \frac{C}{8}\times N}$, where $N = W \times H$ is much larger than that in light attenion unit. The dense attention map is obtained according to Eq. (2). We put D_q into another 1×1 convolutional layer to obtain $F_v \in \mathbb{R}^{B\times C\times H\times W}$, which can be reshaped to $F_{v1} \in \mathbb{R}^{B\times C\times N}$. The output feature of dense attention unit F_{dense} can be computed as:

$$F_{dense} = \varphi(\theta F_{v1} A_{dense}^{\mathrm{T}}) + F_v \qquad (5)$$

where φ indicates the matrix alignment transformation, A_{dense}^{T} is the transpose of A_{dense}, and θ is a scalar value.

2.4. Prototypical Channel Attention

Since each channel of features can be regarded as a class-specific element, we attempt to explore the channel interdependencies to emphasize the 'better' ones. To reduce model complexity and compute efficiently, we design a squeeze block consisted of fully connected layers with a reduction ratio (equal to 16) and a ReLU function. The feature map $D_s \in \mathbb{R}^{B\times C\times H\times W}$, as state in Eq. (4), is firstly passed through the block to obtain the channel attention $F_s^{'}$:

$$F_s^{'} = FCN(Relu(FCN(D_s))) \qquad (6)$$

Inspired by [13], we calculate the weights of each local region of $F_s^{'}$ to obtain the prototype vector F_{pro} as follow,

$$F_{pro} = (\sum_{i=0}^{HW} e^{F_{s(i)}^{'}} \cdot F_{s(i)}^{'}) / \sum_{j=0}^{HW} e^{F_{s(j)}^{'}} \qquad (7)$$

where H and W are the width and height of $F_s^{'}$. We expand F_{pro} to the same size with F_{dense}, and concatenate them along channel dimension. The concatenation is then fed into a Pyramid Pooling Module [11] to yield the probability map.

3. EXPERIMENTS

3.1. Dataset and Evaluation Metric

We evaluate our approach on Pascal-5^i [1] and FSS-1000 [10] datasets. Pascal-5^i is composed of the PASCAL VOC 2012 [17] and SBD dataset [18]. 20 classes are divided into 4 splits and the model is trained in a cross-validation manner. Three splits are used for training while the remaining one is for test. Following the same settings in [1], 1000 support-query image pairs are randomly sampled from testing set to evaluate the model. The few-shot segmentation dataset FSS-1000 [10] consists of 1000 object classes, each of which only has 10 images with binary masks. We adopt the same settings in [10] in our experiments. Concretly, the training, validating, and testing splits are composed of 520, 240, and 240 object classes, respectively. We employ mean Intersection-over-Union (mIoU) and Foreground-Background Intersection-over-Union (FB-IoU) as the evaluation metrics. Note that mIoU computes the average of all class IoUs in each fold and FB-IoU computes the average of all foreground-background IoUs, neglecting the classes information.

3.2. Implementation Details

To evaluate the performance of our approach, all experimental settings are the same as that in PFENet [6]. We take PFENet as the baseline and choose ResNet-50 as the backbone of our DANet for fair comparison. For Pascal-5^i and FSS-1000, we set batch size as 4 and adopt a SGD optimizer with the learning rate of 2.5×10^{-3} to train our model for 200 epochs. All images of two datasets are resized and cropped to 473×473. We run all the experiments on an NVIDIA TITAN Xp GPU.

3.3. Comparisons with State-of-the-arts

Table 1 reports the comparison of our proposed DANet with the state-of-the-arts on PASCAL-5^i. With a ResNet-50 backbone, our DANet outperforms other approaches with a mIoU increase of 1.08% under 1-shot setting, while being comparable with other methods under 5-shot setting. This demonstrates that our proposed dual attention modules are effective in improving the performance of few-shot segmentation.

Table 1. Comparison with the state-of-the-arts using class mIoU (%) and FB-IoU (%) on Pascal-5^i for 1-shot and 5-shot segmentation. FB-IoU is the average across 4 splits. The best is marked in bold.

Method	1-shot						5-shot					
	Fold0	Fold1	Fold2	Fold3	Mean	FB-IoU	Fold0	Fold1	Fold2	Fold3	Mean	FB-IoU
ResNet-101												
FWB [14]	51.30	64.49	56.71	52.24	56.19	-	54.84	67.38	62.16	55.30	59.92	-
DAN [15]	54.70	68.60	**57.80**	51.60	58.20	71.90	57.90	69.00	60.10	54.90	60.50	72.30
ResNet-50												
PGNet [16]	56.00	66.90	50.60	50.40	56.00	69.90	57.70	68.70	52.90	54.60	58.50	70.70
RPMMs [7]	55.15	66.91	52.61	50.68	56.34	-	56.28	67.34	54.52	51.00	57.30	-
PPNet [8]	47.83	58.75	53.80	45.63	51.50	-	58.39	67.83	**64.88**	56.73	61.96	-
ASGNet [9]	58.84	67.86	56.79	53.66	59.29	69.20	63.66	70.55	64.17	57.38	**63.94**	**74.20**
PFENet [6]	61.70	69.50	55.40	56.30	60.80	**73.30**	63.10	70.70	55.80	57.90	61.90	73.90
DANet	**63.28**	**70.77**	56.83	**56.64**	**61.88**	71.78	**65.31**	**71.81**	56.97	**58.55**	63.16	73.11

Table 2. Comparison with other state-of-the-arts using Mean-IoU (%) on FSS-1000 under 1-shot and 5-shot settings. The best is marked in bold.

Method	Backbone	Mean-IoU (%)	
		1-shot	5-shot
OSLSM [1]	VGG16	70.3	73.0
GNet [20]	VGG16	71.9	74.3
FSS [10]	VGG16	73.5	80.1
DoG-LSTM [21]	VGG16	80.8	83.4
DANet	VGG16	**82.0**	**84.3**
DANet	Resnet50	**83.6**	**86.3**

Moreover, we also plot the qualitative segmentation results in Fig. 2.

Table 2 reports the results of our proposed DANet and other state-of-the-arts on FSS-1000. With a VGG16 [19] backbone, the proposed DANet achieves superior performance by significant margins of 1.2% and 0.9%, respectively.

Overall, our proposed DANet outperforms the others for few-shot segmentation on both two datasets in most cases.

3.4. Ablation study

To verify and validate the effectiveness of our proposed approach, we conduct extensive ablation studies with a ResNet-50 backbone on Pascal-5^i. We use Mean-IoU as evaluation metric and average the scores across all splits. As shown in Table 3, compared with the baseline, only using the Light-dense Attention directly increases the mIoU score by 0.57% and 0.81%, respectively. We can see that both of the dual-attention modules achieve promotions individually. Besides, using two attention modules together obtains a 63.16% mIoU score for 5-shot setting, which is 1.26% higher than the baseline. It's worth noting that we don't have to retrain the model for k-shot tasks. The attention maps and prototype vector are both averaged over the support samples, to provide a more balanced guidance.

Table 3. Ablation study of the proposed Light-Dense Attention (LDA) and Prototypical Channel Attention (PCA) on Pascal-5^i under 1-shot and 5-shot settings.

Backbone	base	PCA	LDA	Mean-IoU (%)	
				1-shot	5-shot
ResNet50	✓			60.80	61.90
ResNet50	✓	✓		61.03	62.34
ResNet50	✓		✓	61.37	62.71
ResNet50	✓	✓	✓	**61.88**	**63.16**

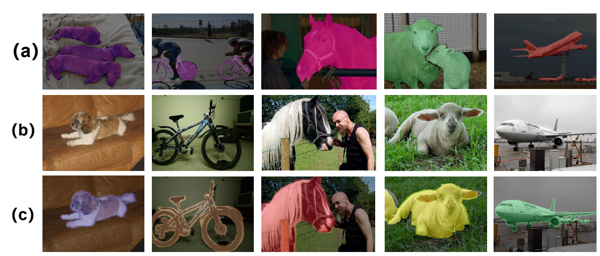

Fig. 2. Qualitative results of 1-shot segmentation on dataset Pascal-5^i. From top to bottom: (a) Support images and their ground-truth, (b) Query images, (c) Results of our DANet.

In summary, the combination of two modules achieves the best results, which indicates the effectiveness of our DANet.

4. CONCLUSION

In this work, we have proposed a dual-attention network for few-shot segmentation. DANet is able to establish region correspondence at different levels to activate target object area by the Light-dense Attention module and exploit the inter-channel relations of prototypes to provide a superior class representation. Extensive experiments have shown the effectiveness of our method.

5. REFERENCES

[1] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots, "One-shot learning for semantic segmentation," *arXiv preprint arXiv:1709.03410*, 2017.

[2] Nanqing Dong and Eric P. Xing, "Few-shot semantic segmentation with prototype learning," in *BMVC*. 2018, p. 79, BMVA Press.

[3] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9197–9206.

[4] Binghao Liu, Jianbin Jiao, and Qixiang Ye, "Harmonic feature activation for few-shot semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 3142–3153, 2021.

[5] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen, "Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5217–5226.

[6] 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*, 2020.

[7] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye, "Prototype mixture models for few-shot semantic segmentation," in *ECCV (8)*. 2020, vol. 12353 of *Lecture Notes in Computer Science*, pp. 763–778, Springer.

[8] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He, "Part-aware prototype network for few-shot semantic segmentation," in *ECCV (9)*. 2020, vol. 12354 of *Lecture Notes in Computer Science*, pp. 142–158, Springer.

[9] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim, "Adaptive prototype learning and allocation for few-shot segmentation," in *CVPR*. 2021, pp. 8334–8343, Computer Vision Foundation / IEEE.

[10] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang, "Fss-1000: A 1000-class dataset for few-shot segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2869–2878.

[11] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *CVPR*. 2017, pp. 6230–6239, IEEE Computer Society.

[12] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu, "Dual attention network for scene segmentation," in *CVPR*. 2019, pp. 3146–3154, Computer Vision Foundation / IEEE.

[13] Alexandros Stergiou, Ronald Poppe, and Grigorios Kalliatakis, "Refining activation downsampling with softpool," *CoRR*, vol. abs/2101.00440, 2021.

[14] Khoi Nguyen and Sinisa Todorovic, "Feature weighting and boosting for few-shot segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 622–631.

[15] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen, "Few-shot semantic segmentation with democratic attention networks," in *Proceedings of the European Conference on Computer Vision*, 2020.

[16] 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 *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9587–9595.

[17] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[18] Bharath Hariharan, Pablo Andrés Arbeláez, Ross B. Girshick, and Jitendra Malik, "Simultaneous detection and segmentation," in *ECCV (7)*. 2014, vol. 8695 of *Lecture Notes in Computer Science*, pp. 297–312, Springer.

[19] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[20] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alexei A. Efros, and Sergey Levine, "Few-shot segmentation propagation with guided networks," *CoRR*, vol. abs/1806.07373, 2018.

[21] 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 *WACV*. 2021, pp. 2673–2682, IEEE.