



RESEARCH ARTICLE

A few-shot semantic segmentation method based on adaptively mining correlation network

Zhifu Huang, Bin Jiang  and Yu Liu* 

School of Automation Science and Engineering, South China University of Technology, Guangzhou, China

*Corresponding author. E-mail: ayylau@scut.edu.cn

Received: 28 June 2022; **Revised:** 16 January 2023; **Accepted:** 24 January 2023; **First published online:** 13 March 2023

Keywords: computer vision, deep learning, convolutional neural network, few-shot semantic segmentation, intelligent system

Abstract

The goal of few-shot semantic segmentation is to learn a segmentation model that can segment novel classes in queries when only a few annotated support examples are available. Due to large intra-class variations, the building of accurate semantic correlation remains a challenging job. Current methods typically use 4D kernels to learn the semantic correlation of feature maps. However, they still face the challenge of reducing the consumption of computation and memory while keeping the availability of correlations mined by their methods. In this paper, we propose the adaptively mining correlation network (AMCNet) to alleviate the aforementioned issues. The key points of AMCNet are the proposed adaptive separable 4D kernel and the learnable pyramid correlation module, which form the basic block for correlation encoder and provide a learnable concatenation operation over pyramid correlation tensors, respectively. Experiments on the PASCAL VOC 2012 dataset show that our AMCNet surpasses the state-of-the-art method by 0.7% and 2.2% on 1-shot and 5-shot segmentation scenarios, respectively.

1. Introduction

Recently, the development of deep convolutional neural networks [1, 2] contributes to some significant breakthroughs in many traditional vision tasks, for example, object detection [3–5], robot vision [6], and semantic segmentation [7, 8]. For example, in data annotation, the manual labeling costs much time and money if a large training set needs to be established. The automated labeling helps reduce costs and improve efficiency if robots or machines can be trained to label data as humans. Few-shot learning is exactly proposed to train machines or robots to work like humans. Specifically, humans can easily learn a novel concept after seeing several examples from the same class. However, for machines or robots, the shortage of annotated samples [9] always restricts the generalization ability of algorithms in the few-shot scenario. Current works [10, 11] suggest that the key point is whether there exist reliable correlations established by machines between supports and queries.

We propose a novel convolutional neural network architecture, named adaptive mining correlation network (AMCNet), to alleviate the aforementioned issues. As done in previous works [10, 12], we attach importance to middle-layer features due to their effectiveness for accurate correlations capture. More specifically, we utilize a weight-shared feature extractor to generate these middle-layer feature maps for 4D correlation tensors generation. For obtaining different levels of receptive fields over the support-related region, we introduce adaptive separable 4D kernel (AS-Conv4d) to adaptively learn generated correlation representations. AS-Conv4d consists of three separable 2D kernels. Due to the variable receptive field in support-related subspace, AS-Conv4d allows query-related subspace to take a more flexible strategy to integrate the information in support-related subspace.

Furthermore, we design a learnable pyramid correlation module (PCM) to squeeze and concatenate pyramid correlation tensors adaptively. It propagates the target-related information across different levels of feature via the top-down form. Based on the proposed AS-Conv4d and PCM, we build

AMCNet. We confirm its efficacy in 1-shot and 5-shot scenarios with comprehensive experiments on the PASCAL-5ⁱ [13].

The main contributions of this paper are summarized as follows:

- (i) We develop a 4d kernel called AS-Conv4d. It encourages encoder in query-related subspace to take a more flexible strategy to absorb the information in support-related subspace.
- (ii) Based on AS-Conv4d, we build PCM. It is conducive to automatically building the squeezed semantic feature for query segmentation by concatenating pyramid correlation with learnable mixing operation.
- (iii) This work on the PASCAL-5ⁱ [13] shows that our AMCNet achieves a mean Intersection-over-Union score of 63.5% for 1-shot scenario and 68.8% for 5-shot scenario, surpassing the state-of-the-art method by 0.7% and 2.2%.

The rest of this paper is organized as follows. In Section 2, the task of the few-shot semantic segmentation is briefly described. In Section 3, the presented modules including AS-Conv4d and PCM are clearly explained. We report the experimental results and corresponding analyses in Section 4. A conclusion of this work is in Section 5.

2. Task Description

We follow OSLSM [13] to partition the PASCAL VOC 2012 dataset [14] into fourfold $\{F_i\}_{i=1}^4$ with category set $\{C_i\}_{i=1}^4$, in which $C_i \cap C_j = \emptyset, j = 1, 2, 3, 4$ and $i \neq j$. We sample three of them to form the training set D_{train} and the remaining one for the test set D_{test} . Our network will be trained on D_{train} and evaluated on D_{test} . For the few-shot setting, both D_{train} and D_{test} are arranged with the episodic paradigm [15], which suggests that for either D_{train} or D_{test} , each episode is comprised of a support and a query set. For example, we sample k image-mask pairs of class c to form the support set $S(c) = \{I_s^i(c), M_s^i(c)\}_{i=1}^k$, where for the episode of class c $I_s^i(c)$ and $M_s^i(c)$ are the i th support image and the corresponding mask, respectively; and then we random sample an example of class c but different from those supports to form the query set $Q(c) = \{I_q(c), M_q(c)\}$ of this episode, where $I_q(c)$ and $M_q(c)$ are the input query image and the ground-truth binary mask, respectively. Each batch of input data to the model is formulated by $I_q(c)$ and $S(c)$. The ground-truth mask $M_q(c)$ serves as supervision to force the network to generate the predicted mask $\hat{M}_q(c)$ during the training, while during the test it just plays a role in evaluating the performance of our network.

3. Method

3.1. Semantic correlation generation

Most traditional semantic correlation learning methods [10, 12, 16] pay attention to the pairwise similarity between the support and the query images. Following their works, we provide generated correlations formed by feature maps for later semantic encoding.

Suppose that $I_s \in \mathbb{R}^{H \times W \times 3}$ and $I_q \in \mathbb{R}^{H \times W \times 3}$ are a support RGB image and a query RGB image in the same episode, we get features $F_s \in \mathbb{R}^{h' \times w' \times c}$ and $F_q \in \mathbb{R}^{h' \times w' \times c}$ via the backbone model which is pre-trained on Imagenet [9] as done in the previous few-shot segmentation works. We subsequently mask the extracted support feature $F_s \in \mathbb{R}^{h' \times w' \times c}$ with the scaled-down mask $M_s \in \{0, 1\}^{h' \times w'}$ to only retain the foreground region for accurate object localization as in ref. [17]:

$$F_s = F_s \odot M_s \in \mathbb{R}^{h' \times w' \times c} \tag{1}$$

where \odot is Hadamard product. We here flatten F_s and F_q to $F'_s \in \mathbb{R}^{h'w' \times c}$ and $F'_q \in \mathbb{R}^{h'w' \times c}$ for the sake of convenience. Subsequently, the semantic correlation representation is established by cosine similarity:

$$C' = \frac{F'_q \cdot F'^T_s}{\|F'_q\| \|F'_s\|} \in \mathbb{R}^{h'w' \times h'w'} \tag{2}$$

For each entry $c_i \in C'$, the irrelevant matching scores ranging from -1.0 to 0 are mapped to 0 as:

$$c_i = \max(0, c_i) \tag{3}$$

Subsequently for the following correlation learning, $C' \in \mathbb{R}^{h' \times w' \times h' \times w'}$ is reshaped to $C \in \mathbb{R}^{h' \times w' \times h' \times w'}$.

3.2. Adaptive separable 4D kernel

Full 4D convolution implementation scheme is revisited in this section, and then we introduce our AS-Conv4d for comparison. The formulation of full 4D convolution is

$$(K * C)(x, y) = \sum_{u,v} K(u, v)C(x - u, y - v) \tag{4}$$

where $C(x, y) \in \mathbb{R}^{H_q \times W_q \times H_s \times W_s}$ represents correlation tensor which is established by cosine similarity, and $K \in \mathbb{R}^{d \times d \times d \times d}$ is 4D convolution kernel. Although some works [10, 12] in terms of semantic correlation learning have verified its efficacy, it is so difficult to form an encoder by full 4D convolution kernel on few-shot semantic segmentation because of quadratic complexity [16].

Furthermore, we propose a novel 4D kernel called AS-Conv4d to make the query feature flexibly absorb the relevant information of the support feature. AS-Conv4d fixes the search window size of query-related subspace and changes the search window size of support-related subspace. Specifically, we factorize a 4D filter $K(x, y) \in \mathbb{R}^{d \times d \times d \times d}$ into three 2D filters $K_1(x), K_2(y), K_3(y) \in \mathbb{R}^{d \times d}$ as:

$$(K * C)(x, y) = K_1(x) * \{[K_2(y) + K_3(y)] * C(x, y)\} \tag{5}$$

where $x \in \mathbb{R}^2$ and $y \in \mathbb{R}^2$ are the position of the query-related subspace and the support-related subspace in semantic correlation, respectively. Note that K_2 and K_3 are different shapes of 2D filter. For instance, in our work, we set K_2 to the size of 3×3 and set K_3 to the size of 5×5 . In comparison with previous 4D kernels [12, 16], AS-Conv4d not only reduces computational complexity $O(d^4)$ to $O(d^2)$ but also keeps a better balance between receptive field and spatial resolution and thus builds a closer connection between query and support subspace.

3.3. Model architecture

An encoder–decoder architecture is implemented to learn different levels of semantic correlations $\{C_i\}_{i=3}^5$ which are cast over all intermediate convolutional layers, that is, the third to the fifth convolution layers in ResNet50. In our encoder, we utilize three parallel sequences, which are informed by a series of 4D convolutions, group normalizations [18], and ReLU activations, to learn different levels of semantic correlations $\{C_i\}_{i=3}^5$. And then with the top-down form of PCM, we mix the compressed pyramid correlations $\{C_i\}_{i=3}^5$ to spread relevant information to lower layers, that is, from C_5 to C_3 .

Specifically for encoding as illustrated in Fig. 1, our AMCNNet learns correlations by squeezing the shape of the support-related 2D subspace (H_s, W_s) while maintaining the query-related subspace (H_q, W_q), and then PCM concatenates adjacent pyramid layers. After two PCMs, we propagate C_5 to C_4 and C_{54} to C_3 for the mixed squeezed correlation C_{543} , then we utilize global average pooling over (H_s, W_s) to produce the encoding result $Z \in \mathbb{R}^{H_q \times W_q \times c}$ which denotes abstract semantic correspondence learned by our model in original correlations for the following decoding.

We implement a simple decoder network as illustrated in Fig. 1 (bottom), which is built by 2D convolutions and ReLU activations. The condensed representation Z is input to it, and then we can get the predicted segmentation mask $\hat{M}_q \in \{0, 1\}^{H \times W}$. We utilize cross-entropy loss to optimize the learnable parameters in our model as follows:

$$CELoss = \sum_{x,y} [M_q(x, y) \times \log(\hat{M}_q(x, y)) + (1 - M_q(x, y)) \times \log(1 - \hat{M}_q(x, y))] \tag{6}$$

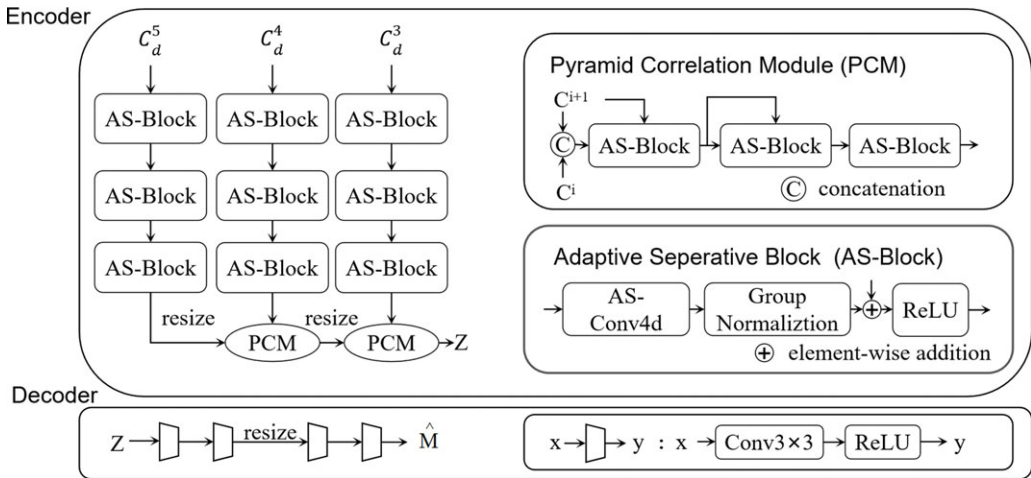


Fig. 1. The frameworks of semantic correlation encoder and decoder in our AMCNet.

where \hat{M}_q and M_q represent the prediction result and the ground-truth over all pixel locations (x, y) . While test, \hat{M}_q is compared with M_q by the Intersection-over-Union score for the evolution of our model.

4. Experiment

4.1. Implementation details

ResNet50 [19, 20] and ResNet101 [17] with the weights pretrained on ImageNet [9] are utilized as the backbone in our network. We use Adam to train the whole model on a GeForce RTX 3080 GPU. During training, the learning rate, batch size, and image size are 0.001, 8, and 400×400 for PASCAL-5ⁱ and COCO-20ⁱ datasets, and the epoch is 300 for PASCAL-5ⁱ dataset and the epoch is 40 for COCO-20ⁱ dataset.

4.2. Evaluation metrics

The mean Intersection-over-Union (mIoU) and the foreground-background Intersection-over-Union (FB-IoU) are utilized for evaluation in this work. The mIoU is formulated by $mIoU = \frac{1}{m} \sum_{i=1}^m IoU_i$ where IoU_i is the Intersection-over-Union score of class i and m is the number of categories in the test set. The FB-IoU is formulated by $FB-IoU = \frac{1}{2} \sum_{i=0}^1 IoU_i$, where the foreground class 1 represents all object categories included in the test set, while the background class 0 includes all pixels outside the foreground area.

4.3. Experiments on PASCAL-5ⁱ

We extend our work to K -shot scenario ($K > 1$). A query image I_q and the corresponding K support image-mask pairs $S = \{(I_s^k, M_s^k)\}_{k=1}^K$ are input into the proposed AMCNet, and then our network outputs K query mask predictions $\{\hat{M}_q^k\}_{k=1}^K$ in a forward way. These predictions $\{\hat{M}_q^k\}_{k=1}^K$ play an important role in the pixel-wise voting. Specifically, if at least half of K voters at this location (x, y) is 0, the prediction result at this location (x, y) is labeled as a background pixel; otherwise, this location is labeled as a foreground pixel. In this work, following most of works [13, 17, 19–24], we take $K = 5$ for comprehensively evaluating our model performance.

We report the performance comparison with state-of-the-arts on PASCAL-5ⁱ [13] in Tables I and II. We can actually see either for the mIoU or the FB-IoU evaluation, and our AMCNet records new state-of-the-art in both 1-shot and 5-shot scenarios. Specifically, for the ResNet50-based methods in Table I,

Table I. Comparison with state-of-the-arts on PASCAL-5ⁱ [13] in mIoU.

Backbone	Methods	1-shot (%)					5-shot (%)				
		5 ⁰	5 ¹	5 ²	5 ³	Mean	5 ⁰	5 ¹	5 ²	5 ³	Mean
VGG16	OSLSM [13]	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
	PANet [21]	42.3	58.0	51.1	41.2	48.1	51.8	64.6	59.8	46.5	55.7
	PFENet [17]	56.9	68.2	54.4	52.4	58.0	59.0	69.1	54.8	52.9	59.0
ResNet50	PANet [21]	44.0	57.5	50.8	44.0	49.1	55.3	67.2	61.3	53.2	59.3
	PGNet [20]	56.0	66.9	50.6	50.4	56.0	57.7	68.7	52.9	54.6	58.5
	PFENet [17]	61.7	69.5	55.4	56.3	60.8	63.1	70.7	55.8	57.9	61.9
	SAGNN [25]	64.7	69.6	57.0	57.2	62.1	64.9	70.0	57.0	59.3	62.8
	CMN [26]	64.3	70.0	57.4	59.4	62.8	65.8	70.4	57.6	60.8	63.7
	ASGNet [22]	58.8	67.9	56.8	53.7	59.3	63.7	70.6	64.2	57.4	63.9
	RePRI [23]	59.8	68.3	62.1	48.5	59.7	64.6	71.4	71.1	59.3	66.6
	AMCNet (Ours)	63.0	70.6	60.0	60.4	63.5	68.7	72.8	66.0	67.6	68.8

Best results in bold.

Table II. Comparison with state-of-the-arts on PASCAL-5ⁱ [13] in FB-IoU and Params.

Backbone	Methods	FB-IoU		
		1-shot (%)	5-shot (%)	Params
VGG16	OSLSM [13]	61.3	61.5	272.6M
	co-FCN [24]	60.1	60.2	34.2M
	SG-One [27]	63.9	65.9	19.0M
	PANet [21]	66.5	70.7	14.7M
	PFENet [17]	72.0	72.3	10.4M
ResNet50	CANet [19]	66.2	69.6	19.0M
	PGNet [20]	69.9	70.5	17.2M
	PFENet [17]	73.3	73.9	10.8M
	ASGNet [22]	69.2	74.2	10.4M
	AMCNet (Ours)	76.4	80.0	6.5M

Params: the number of learnable parameters.

AMCNet achieves 63.5% and 68.8% in terms of mIoU for the 1-shot and 5-shot settings, respectively, which surpasses the state-of-the-art by 0.7% and 2.2%.

Furthermore, as shown in Table II, AMCNet also achieves the best performance in FB-IoU, that is, 76.4% and 80.0% for the ResNet50-based methods, while only requiring the slightest number of learning parameters, which confirms the effectiveness of AMCNet on the topic of few-shot semantic segmentation.

Finally, in comparison with the state-of-the-art methods, AMCNet has the slightest learnable parameters, which means that it can effectively reduce memory consumption while achieving the best segmentation performance. Especially in industry, it is very important to cut down computation for pursuing time efficiency.

4.4. Experiments on COCO-20ⁱ

We also extend our experiments on COCO-20ⁱ [28], a more challenging dataset including total 80 object classes. As shown in Table III, Our AMCNet also outperforms the state-of-the-art method in both the 1-shot scenario and the 5-shot scenario. For instance, AMCNet surpasses the state-of-the-art method CMN

Table III. Comparison with state-of-the-arts on COCO-20ⁱ [28] in mIoU.

Backbone	Methods	1-shot (%)					5-shot (%)				
		20 ⁰	20 ¹	20 ²	20 ³	Mean	20 ⁰	20 ¹	20 ²	20 ³	Mean
ResNet50	PPNet [29]	28.1	30.8	29.5	27.7	29.0	39.0	40.8	37.1	37.3	38.5
	PMMs [30]	29.5	36.8	28.9	27.0	30.6	33.8	42.0	33.0	33.3	35.5
	PFENet [17]	36.5	38.6	34.5	33.8	35.8	36.5	43.3	37.8	38.4	39.0
	RePRI [23]	32.0	38.7	32.7	33.1	34.1	39.3	45.4	39.7	41.8	41.6
	CMN [26]	37.9	44.8	38.7	35.6	39.3	42.0	50.5	41.0	38.9	43.1
ResNet101	FWB [31]	19.9	18.0	21.0	28.9	21.2	19.1	21.5	23.9	30.1	23.7
	PFENet [17]	36.8	41.8	38.7	36.7	38.5	40.4	46.8	43.2	40.5	42.7
	SAGNN [25]	36.1	41.0	38.2	33.5	37.2	40.9	48.3	42.6	38.9	42.7
	AMCNet (Ours)	37.6	44.2	41.1	40.6	40.9	42.3	50.8	45.6	44.5	45.8

Best results in bold.

[26] by 1.6% and 2.7% with the mIoU scores in the 1-shot and the 5-shot scenarios. The significant performance improvement on COCO-20ⁱ [28] denotes the remarkable capability of our AMCNet to handle complex scenes.

4.5. Results analyses

Experiments show that our AMCNet achieves the best performance on both PASCAL-5ⁱ [13] and COCO-20ⁱ [28] datasets. For the PASCAL-5ⁱ [13] dataset, we improve the best mIoU score to 63.5% in 1-shot scenario and 68.8% in 5-shot scenario. For the COCO-20ⁱ [28] dataset, we improve the best mIoU score to 40.9% in 1-shot scenario and 45.8% in 5-shot scenario. Some predicted results for 5-shot semantic segmentation are shown in Fig. 2. The images of Fig. 2(a) are the densely annotated support samples, while the first column images of Fig. 2(b) are the query images with the predicted mask and the second column images of Fig. 2(b) are the query images with the ground-truth mask. We change the transparency level of binary masks and then integrate them with corresponding RGB images for the convenience of comparison. Note that for simulating 5-shot setting, the support and the query samples are different instances although they are from the same category.

As shown in Fig. 2, we can see that although there exist large intra-class variations in the object category of dining table, dog, horse, motorbike, and person, our model still has a remarkable ability to segment a novel concept after only seeing a few examples. This confirms the remarkable ability of our AMCNet to segment novel concepts although large intra-class variations exist in few-shot scenario. Furthermore, it is worth noting that our AMCNet achieves the best performance with the fewest learnable parameters (6.5M for ResNet-based models). More qualitative examples of the proposed AMCNet can be seen in Fig. 3. As illustrated in Fig. 3, due to the more flexible receptive field of AS-Conv4d and the appropriate mixing operation of PCM, AMCNet has a remarkable ability to retain essential semantic information across different scales and thus has a good performance for the capture of both the large and small objects.

5. Conclusion

In this paper, a fully convolutional network-based upon pseudo-dense 4D convolutions is proposed to handle complex few-shot semantic segmentation. Despite under limited supervision, experiment on benchmarks has verified the superiority of the proposed adaptive separable 4D convolutional kernel (AS-Conv4d) and PCM in fine-grained segmentation. We comprehensively incorporate them into our AMCNet on the PASCAL-5ⁱ and COCO-20ⁱ datasets and update the state-of-the-art records. Possible future work includes extending our work from few-shot to zero-shot scenario.



Fig. 2. Images with binary mask: images of (a) are the support images, the first column of (b) images are the predicted results, and the second column images of (b) are the ground-truths.



Fig. 3. More qualitative examples of our models. The first, third, fifth, and seventh columns are the predicted results, and the second, fourth, sixth, and eighth columns are the ground truths.

Financial support. This work was supported by the Key R&D Program of Guangdong Province (2021B0101200001) and by the Guangdong Basic and Applied Basic Research Foundation (2020B1515120071, 2021B1515120017).

Competing interests. No conflicts.

Author Contributions. Zhifu Huang designed and implemented the research and wrote the manuscript. Bin Jiang assisted in the research and edited the manuscript. Yu Liu directed the research and reviewed the manuscript.

References

- [1] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition” (2014), arXiv: 1409.1556.
- [2] G. Huang, Z. Liu, L. van der Maaten and K. Q. Weinberger, “Densely connected convolutional networks,” **In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, (2017) pp. 4700–4708.
- [3] Q. Wang, L. Zhang, L. Bertinetto, W. Hu and P. H. S. Torr, “Fast online object tracking and segmentation: A unifying approach,” **In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, (2019) pp. 1328–1338.
- [4] J. Zhi, D. Luo, K. Li, Y. Liu and H. Liu, “A novel method of shuttlecock trajectory tracking and prediction for a badminton robot,” *Robotica* **40**(6), 1682–1694 (2022).
- [5] S. Zare, M. R. H. Yazdi, M. T. Masouleh, D. Zhang, S. Ajami and A. A. Ardekani, “Experimental study on the control of a suspended cable-driven parallel robot for object tracking purpose,” *Robotica* **40**(11), 3863–3877 (2022).
- [6] 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. Pattern Anal. Mach. Intell.* **40**(4), 834–848 (2019).
- [7] E. Shelhamer, J. Long and T. Darrell, “Fully convolutional networks for semantic segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 640–651 (2017).
- [8] L. Kenye and R. Kala, “Improving RGB-D SLAM in dynamic environments using semantic aided segmentation,” *Robotica* **40**(6), 2065–2090 (2022).
- [9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and F.-F. Li, “Imagenet: A large-scale hierarchical image database,” **In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, (2009) pp. 248–255.
- [10] S. Li, K. Han, T. W. Costain, H. Howard-Jenkins and V. Prisacariu, “Correspondence networks with adaptive neighbourhood consensus,” **In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, (2020) pp. 10193–10202.
- [11] C. B. Choy, J. Y. Gwak, S. Savarese and M. Chandraker, “Universal correspondence network,” **In: Proceedings of the International Conference on Neural Information Processing Systems**, (2016) pp. 2414–2422.
- [12] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla and J. Sivic, “Neighbourhood consensus networks,” **In: Proceedings of the Advances in Neural Information Processing Systems**, (2018) pp. 1651–1662.
- [13] A. Shaban, S. Bansal, Z. Liu, I. Essa and B. Bootstittle, “One-shot learning for semantic segmentation,” (2017), arXiv: 1709.03410.
- [14] M. Everingham, S. M. A. Eslami, L. van Gool, C. K. I. Williams, J. Winn and A. Zisserman, “The pascal visual object classes challenge: A retrospective,” *Int. J. Comput. Vis.* **111**(1), 98–136 (2015).
- [15] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu and D. Wierstra, “Matching networks for one shot learning,” **In: Proceedings of the Advances in Neural Information Processing Systems**, (2016) pp. 3630–3638.
- [16] G. Yang and D. Ramanan, “Volumetric correspondence networks for optical flow,” **In: Proceedings of the Advances in Neural Information Processing Systems**, (2019) pp. 794–805.
- [17] Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li and J. Jia, “Prior guided feature enrichment network for few-shot segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(2), 1050–1065 (2020).
- [18] Y. Wu and K. He, “Group normalization,” *Int. J. Comput. Vis.* **128**(3), 742–755 (2020).
- [19] 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 the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, (2019) pp. 5217–5226.
- [20] 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 the IEEE/CVF International Conference on Computer Vision**, (2019) pp. 9587–9595.
- [21] 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/CVF International Conference on Computer Vision**, (2019) pp. 622–631.
- [22] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim and J. Kim, “Adaptive prototype learning and allocation for few-shot segmentation,” **In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, (2021) pp. 8334–8343.
- [23] M. Boudiaf, H. Kervadec, Z. I. Masud, P. Piantanida, I. B. Ayed and J. Dolz, “Few-shot segmentation without meta-learning: A good transductive inference is all you need?,” **In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, (2021) pp. 13979–13988.
- [24] K. Rakelly, E. Shelhamer, T. Darrell, A. Efros and S. Levine, “Conditional networks for few-shot semantic segmentation,” **In: Proceedings of the International Conference on Learning Representations Workshop**, (2018).
- [25] G.-S. Xie, J. Liu, H. Xiong and L. Shao, “Scale-aware graph neural network for few-shot semantic segmentation,” **In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, (2021) pp. 5475–5484.
- [26] G.-S. Xie, H. Xiong, J. Liu, Y. Yao and L. Shao, “Few-shot semantic segmentation with cyclic memory network,” **In: Proceedings of the IEEE/CVF International Conference on Computer Vision**, (2021) pp. 7293–7302.
- [27] X. Zhang, Y. Wei, Y. Yang and T. S. Huang, “SG-One: Similarity guidance network for one-shot semantic segmentation,” *IEEE Trans. Cybern.* **50**(9), 3855–3865 (2020).
- [28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, “Microsoft COCO: common objects in context,” **In: Proceedings of the European Conference on Computer Vision**, (2014) pp. 740–755.

- [29] Y. Liu, X. Zhang, S. Zhang and X. He, “Part-aware prototype network for few-shot semantic segmentation,” *In: Proceedings of the European Conference on Computer Vision*, (2020) pp. 142–158.
- [30] B. Yang, C. Liu, B. Li, J. Jiao and Q. Ye, “Prototype mixture models for few-shot semantic segmentation,” *In: Proceedings of the European Conference on Computer Vision*, (2020) pp. 763–778.
- [31] K. Nguyen and S. Todorovic, “Feature weighting and boosting for few-shot segmentation,” *In: Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2019) pp. 622–631.