Unlocking Multimodal Potential for Few-Shot Semantic Segmentation with Vision-Enriched Text

Siyu Chen[1], Jiaxiang Fang[1], Shiqiang Ma[2]✉, and Fei Guo[1]✉*

[1] Central South University, School of Computer Science and Engineering, Changsha, China
{csy619,224712227,guofei}@csu.edu.cn
[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China sq.ma@siat.ac.cn

Abstract. Few-Shot Semantic Segmentation (FSS), as an emerging technology, aims to transfer knowledge from base classes to novel classes using a limited number of support images. However, current FSS methods often struggle due to the inherent sparsity of data and the variability of features within and across classes, which limits their ability to effectivelygeneralize to novel classes. To uncover the latent commonalities between base and novel class objects, multimodal fusion methods have been introduced to FSS, providing richer semantic information. However, effectively matching text-based global semantic information with image-based local features remains a significant challenge. In this paper, we attempt to adapt multimodal fusion techniques to alleviate the problem of insufficient effective information in FSS. Specifically, we use Vision Enriched Prompts to perceive context and utilize the results of vision-language fusion to guide the correlation calculation between support and query images. By doing this, our method alleviates the problem of insufficient support information, base classes bias, and most importantly unlocks the potential of multimodal in FSS, moreover, extensive experiments on COCO-20^i datasets demonstrate that our model achieves 11.2% and 10.6% increase on 1-shot and 5-shot compared to the previously renowned BAM method.

Keywords: computer vision · few-shot semantic segmentation · vision-language fusion.

1 Introduction

Sematic segmentation is an important task within computer vision that classifies each pixel of an image into different classes. With the development of Deep Neural Network, semantic segmentation has gained great progress in the past

[*] Siyu Chen and Jiaxiang Fang contribute equally to this work.
Corresponding author: Shiqiang Ma and Fei Guo.

years [1, 2, 36, 37]. However, traditional semantic segmentation relies on pixel-level annotation on training data, which proved to be very time-consuming and labor-intensive. In addition, in some cases, such as the analysis of medical images for rare diseases, we are unable to obtain adequate data [25, 26], these problems make it difficult for the data-dependent semantic segmentation task. Based on the above questions, Few-shot semantic segmentation was proposed that transfers the knowledge learned from the limited support images to the query image segment and is emerging as a rapidly developing technology [3–5].

In FSS setting, the base classes and novel classes may have some similarities, for example, dogs and cats may all have furry tails, similar colors and so on. Due to these similarity, the model may wrongly segment some pixels of novel classes to base classes. The reason is that the image feature could not provide enough discriminative information about different classes [19,35]. Inspired by this, some researches started to fuse more discriminative textual information with images to facilitate FSS [13–15]. They found that integrating textual information with images can greatly enrich the model's understanding of novel classes.

Previously, some methods tried to use word embedding to provide textual information, for example, MIANet [13] fuses pre-trained word embedding from Word2Vec with prototypes extracted from images. However, the word embedding only learns the distribution of text in semantic space and has less relevance to image space which may lead to a poor performance of language-vision fusion. The emergence of CLIP [12] provides a novel approach for the fusion of information between vision and language. It aligns text with images through contrastive learning on a large amount of data. By doing this, the images and texts are mapped into the same embedding space, so that an image and its corresponding text description are close to each other. Inspired by CLIP, some methods attempted to utilize pre-trained CLIP encoder to achieve better vision-language fusion performance in semantic segmentation [14, 15]. However, there are still some problems when applying CLIP to FSS. For example, CLIP is image-level supervised which is not suitable for pixel-level task like semantic segmentation, and it only use simple predefined text prompt which has less correlation with the context of image and limits the ability to adapt to downstream work.

To address aforementioned issues we design a model named VTSeg. We first propose a Vision-Language Guided Correlation Module which is composed of two components. The first one named Image-Image Correlation, it follows previous work to calculate the correlation between support and query features. The second one, Image-Text Correlation module will further calculate the similarity between text feature and image feature at each pixel to achieve dense prediction. Besides, we propose a Vision Enriched Prompt Module to enrich text prompt with support feature which enable the text prompt to be more aware of the context in image. Our model can maximizes the application of vision-language in FSS and works as a new baseline.

In summary, our contributions are as follows:

- We propose a Vision-Language Guided Correlation Module which utilizes the result of vision-language fusion to optimize the correlation calculation between support and query images.
- We propose a Vision Enriched Prompt Module which introduces support feature and learnable context to further enrich text prompt.
- We unlock the potential of multimodal methods in FSS and our method demonstrates better performance on commonly used datasets PASCAL-5^i and COCO-20^i for FSS.

2 Related Work

2.1 Few-shot Semantic Segmentation

Few-shot segmentation methods [6, 7] can generally be divided into two categories: parameter-based models and prototype-based models. Parameter-based models use convolutional operations to enhance the model's receptive field, aiming to capture spatial contextual features [8,9,13,17,22]. Prototype-based models utilize metric tools without parameters to measure the distance between global feature prototypes which represent the classes and the features of the query image [10, 11, 27, 28].

Furthermore, parameter-based models and prototype-based models may encounter new issues when integrating with textual information: (1) Parameter-based models are prone to overfitting when trained with a small amount of data. (2) Prototype-based models can severely interfere with pixel-level classification tasks. Due to the limited training data, the model may become overly reliant on coincidental correspondences between text and images, leading to poor generalization on new data. In addition, prototype models typically rely on global feature prototypes to represent entire classes, which may not be sufficient to capture the diversity within the classes, especially when the textual descriptions pertain to specific subclasses or instances.

To address the aforementioned issues, our method combines visual and linguistic modalities, allowing the model to rely not only on image data but also on textual information to enhance the understanding of categories. This multimodal fusion provides an additional data source, which helps to improve the model's generalization ability in scenarios where data is scarce. Specifically, we designed the Vision-Language Guided Correlation Module, which enhances the model's ability to recognize new categories by calculating the correlation between support images and query images, as well as the similarity between query features and text features. This approach reduces the model's over-reliance on limited support images, thereby reducing the risk of overfitting. Furthermore, the Vision Enriched Prompt Module enables the model to adjust text prompts according to each input image, which helps to capture the diversity within categories, especially when text descriptions pertain to specific subclasses or instances. Through the Vision Enriched Prompt Module, we can effectively overcome the problem of prototype-based models making incorrect segmentations of targets when faced with diverse visual features.

2.2 Multimodal Fusion

Multimodal fusion refers to the combination of information from different modalities, such as vision and language, to enhance the performance and generalization capabilities of models [20]. In the context of Few-Shot Semantic Segmentation (FSS), multimodal fusion [12, 21] is particularly important as it can help models better understand and recognize new classes with a limited number of support samples. Despite the immense potential of multimodal fusion, effectively matching text-based global semantic information with image-based local features remains a significant challenge. To address this challenge, researchers have proposed various methods to enhance the model's understanding of new categories.

Guo et al. [33] proposed a multi-view distillation-based multimodal fusion framework for few-shot action recognition. The framework combines text and visual modalities, and through the introduction of a Probability Prompt Selector, compensates for the information inconsistency between prototypes and queries, thereby effectively utilizing labels. Azeem et al. [34] proposed a Unified Multimodal Fusion Transformer (UMFT), which extracts visual features from ViT and text embeddings from BERT, aligning multimodal representations in an end-to-end manner. The proposed Affinity-Guided Fusion (AFM) captures semantically related image-text pairs by simulating the affinity relationships between them, selectively combining the most informative pairs.

However, these methods often overlook the differences between language and vision, leading to poor performance in language-vision fusion. To address the aforementioned issues and maximize the application of vision-language in FSS, we propose a vision-language guided correlation module to calculate the correlation between support and query features, and we design a vision-enriched prompt module to enrich text prompts with support features. This approach not only leverages predefined classes but also adjusts according to each input image, thereby alleviating the base-class bias issue in few-shot semantic segmentation.

3 Methods

3.1 Problem Definition

Few-shot semantic segmentation aims to learn a model that studies how to segment target class with limited images on base classes, and transfers the knowledge to novel classes. There are mainly two imagesets D_{base} and D_{novel} and class sets C_{base} and C_{novel}, where $D_{base} \cap D_{novel} = \emptyset$ and $C_{base} \cap C_{novel} = \emptyset$. Following previous work [22, 27, 29], both D_{base} and D_{novel} contain several episodes. Each episode contains a support set $I_s = \{I_i^s, M_i^s\}_{i=1}^{K}$ and a query set $I_q = \{I^q, M^q\}$, where K, I_i^s, I^q, M_i^s, M^q represent K-shot, support image, query image, support mask and query mask respectively. This model is trained to segment each pixels of images in D_{base} to C_{base} during training and evaluated the performance on D_{novel} with class label C_{novel} during test.

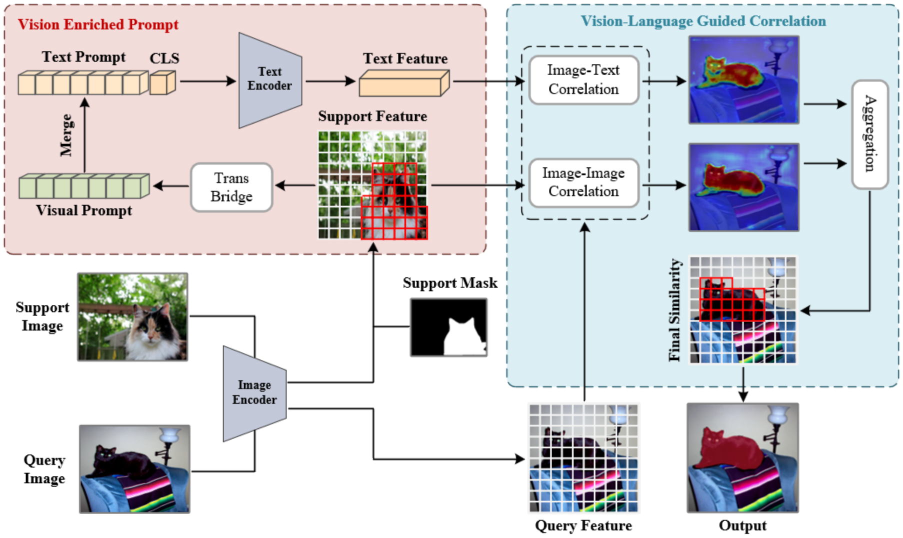

Fig. 1. Overview of our model VTSeg. Vision-Language Guided Correlation module calculate the correlation between image features and text features to transfer information. Vision Enriched Prompt module utilize support features and learnable parameters to enrich text prompt with context.

3.2 Vision-Language Guided Correlation

Following previous work, we use Image-Image Correlation(IIC) to establish the correspondence between support and query images, however, IIC that only use limited images is not sensitive to details in image features, showing serious shortcomings when dealing with complex scenes that have similar classes. To alleviate this problem, we add another Image-Text Correlation module(ITC). Text information can not only provide additional semantic information but also help the model understand the semantic relationships between different classes, reducing the probability of misclassification of classes with similar local features (such as cats and dogs) and improving the model's segmentation accuracy.

To utilize text information, we first send the support and query features to IIC which compute the similarity between them to establish the correspondence, transfer the knowledge of target class learned from support image to query image, and generate a coarse activation diagram.

$$S_1 = \cos\left(F^s, F^q\right), \tag{1}$$

$$\cos\left(\mathrm{F}^s,\ \mathrm{F}^q\right) = \frac{F^s \cdot F^q}{\|F^s\|\,\|F^q\|}, \tag{2}$$

where F^s and F^q means support feature and query feature respectively. *cos* means calculate the cosine similarity between F^s and F^q through Equation 2.

Secondly, we calculate the similarity between query feature and text feature through ITC in the same way.

$$S_1 = \cos\left(F^s, t\right), \tag{3}$$

where t represents text feature.

$$S = \left[S_1, S_2\right], \tag{4}$$

Then, we concatenate S_1 and S_2 to form the final activation results and decode them through a light transformer-based decoder to obtain the final output.

3.3 Vision Enriched Prompt

In the aforementioned methods, we extract text features through fixed text prompt, such as "a photo of [cls]", however, this method is limited. On the one hand, fixed text prompts are not universally applicable to all datasets and different text prompts need to be designed for different datasets, which requires the operation of professionals and is very time-consuming. On the other hand, fixed text prompts cannot perceive the context of images, leading to poor integration with the images and even disrupting the extracted image features.

To solve this problem, inspired by CoCoOp [16], we propose using support features to enrich text prompts, enabling it to automatically perceive context information in different images during training. To do this, we firstly utilize learnable parameters to represent context of text prompt which is trainable during training:

$$t = [\overbrace{p_1 p_2 \ldots p_n}^{context}, cls], \tag{5}$$

where $\overbrace{p_1 p_2 \ldots p_n}^{context}$ means learnable parameters related to context, cls represents class tokens of different classes and *context* means the settig of context length. Then, to enrich the context of text prompt with support feature, we use extracted support feature F_s through pretrained image encoder, and merge it with text prompt as follow:

$$v = TransBridge(F^s), \tag{6}$$

$$t' = t + v, \tag{7}$$

where F^s means extracted support feature, $TransBridge$ represents a light network composed of linear and relu layers that tokenize support feature and transfer it to visual prompt. We regard the obtained visual prompt v as a bias and utilize it to optimize the text prompt during training. This method enables the text prompt to be aware of context. Vision-enriched text prompts not only rely on predefined classes, but also adjust according to each input image, which can alleviate the base-class bias issue in few-shot semantic segmentation.

Table 1. Performance comparison with state-of-the-art on PASCAL-5^i and COCO-20^i in terms of mIoU(%). Results in bold denote the best performance.

Methods	**Setting**	**PASCAL-5^i**					**COCO-20^i**				
		5^0	5^1	5^2	5^3	**Mean**	20^0	20^1	20^2	20^3	**Mean**
NTRENet [18]	1-shot	65.40	72.30	59.40	59.80	64.23	36.80	42.60	39.90	37.90	39.30
BAM [19]		68.97	73.59	67.55	61.13	67.81	38.96	47.04	46.41	41.57	43.50
DPCN [20]		65.70	71.60	69.10	60.60	66.75	42.00	47.00	43.20	39.70	42.98
FECANet [21]		69.20	72.30	62.40	**65.70**	67.40	38.50	44.60	42.60	40.70	41.60
MIANet [15]		68.51	75.76	67.46	63.15	68.72	**42.49**	52.95	47.77	47.42	47.66
QPENet [30]		65.20	71.90	64.10	59.50	65.18	41.50	47.30	40.90	39.40	42.28
DCP [31]		67.20	72.90	65.20	59.40	66.18	43.00	48.60	45.40	44.80	45.45
RiFeNet [32]		68.40	73.50	67.10	59.40	67.10	39.10	47.20	44.60	45.40	44.08
ours		**71.15**	**76.49**	**69.36**	62.72	**69.93**	40.97	**55.20**	**47.79**	**49.57**	**48.38**
NTRENet [18]	5-shot	66.20	72.80	61.70	62.20	65.73	38.20	44.10	40.40	38.40	40.28
BAM [19]		70.59	75.05	70.79	67.20	70.91	47.02	52.62	48.59	49.11	49.34
DPCN [20]		70.00	73.20	70.90	65.50	69.90	46.00	54.90	50.80	47.40	49.78
FECANet [21]		**72.90**	74.00	65.20	67.80	69.98	44.60	51.50	48.40	45.80	47.58
MIANet [13]		70.20	77.38	70.02	**68.77**	71.59	45.84	58.18	51.29	51.90	51.65
QPENet [30]		68.40	74.00	67.40	65.20	68.75	47.30	52.40	44.30	44.90	47.22
DCP [31]		70.50	75.30	68.00	67.70	70.38	47.00	54.70	51.70	50.00	50.85
RiFeNet [32]		70.00	74.70	69.40	64.20	69.58	44.30	52.40	49.30	48.40	48.60
ours		72.60	**77.52**	**72.73**	65.75	**72.15**	**49.02**	**59.56**	**55.21**	**54.52**	**54.58**

4 Experiment

4.1 Experimental Setting

Datasets. The experiments are conducted on two generalized datasets for FSS: PASCAL VOC and MS COCO. For PASCAL VOC which contains 20 classes, we augment it with SBD and divide it into four non-overlapping splits. Each split has five classes and is denoted as PASCAL-5^i. Following the setting of PASCAL-5^i, we construct COCO-20^i from MS COCO which contains 80 classes and each split has 20 non-overlapping classes. We use cross-validation to assess the performance of our model, with one split used as the test set and the other three splits used for training.

Evaluation Metric. We adopt the mean Intersection over Union (mIoU) as our primary evaluation metric. mIoU is denoted as $mIoU = 1/C \sum_{i=1}^{C} IoU_i$, where C means the number of classes in each split, and IoU_i indicates intersection-over-union for class i.

4.2 Implementation Details

Our method is built on the Pytorch framework. Our model is trained for 200 epoches on the PASCAL-5^i with a batchsize of 4, and for 50 epoches on the COCO-20^i with a batchsize of 6. During training, we use AdamW as our optimizer and the learning rate is set to 0.0001. We use Resnet50 [24] as the image

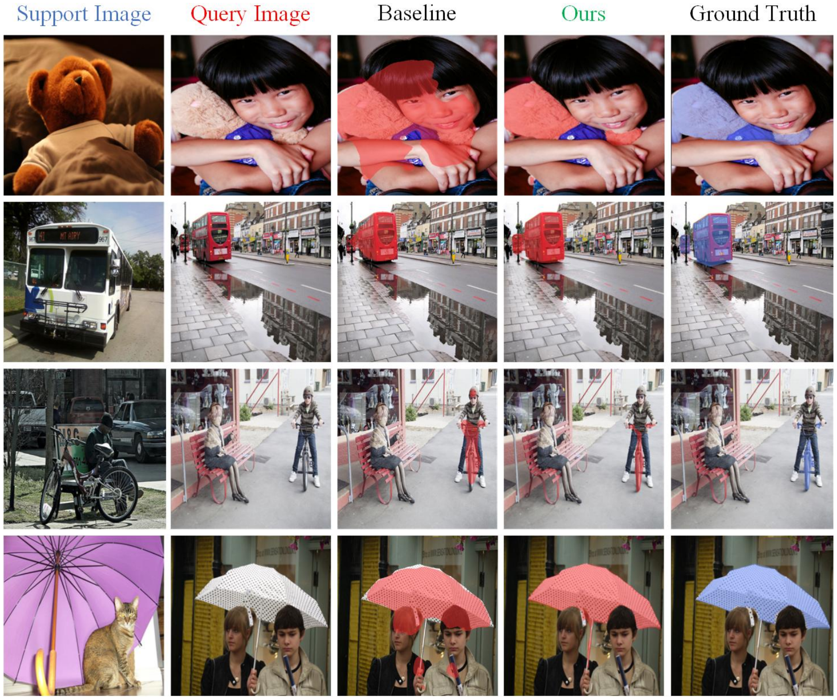

Fig. 2. Qualitative comparison on COCO-20^i. From left to right: support image, query image, prediction of baseline, prediction of ours.

encoder and transformer as the text encoder to extract image features and text features. PSPNet [17] serves as our base learner in all experiments. To ensure stability and fairness of performance for comparison, we randomly sample 1, 000 and 10, 000 test samples for PASCAL-5^i and COCO-20^i.

4.3 Segmentation Performance

Comparison with State-of-the-Art Methods. Our method achieves a new state-of-the-art performance in PASCAL-5^i and COCO-20^i as shown in Table 1. For PASCAL-5^i, our model outperforms prior arts MIANet 1.21% under 1-shot setting. Comparing with PASCAL-5^i, COCO-20^i is a more challenging dataset. It contains more diverse scenes, more complex background, multiple classes may exist within one image and different classes may overlap. It is obviously that our method shows great performance with improvement of 0.72% and 2.93% compared with MIANet on 1-shot and 5-shot setting respectively. Especially under 5-shot setting, our model outperform MIANet at all splits.

Previous methods, whether utilizing convolution or prototypes, relied solely on image information. However, this approach lacks the discriminative information needed to distinguish between similar classes, leading to the misclassification

Table 2. Ablation studies of different components on PASCAL-5^i.

Corr	VEP	5^0	5^1	5^2	5^3	mIoU	Δ
		61.87	72.78	64.10	55.17	63.48	0
✓		67.35	73.39	65.29	59.92	66.49	+3.01
✓	✓	71.15	76.49	69.36	62.72	69.93	+6.45

of novel classes as base classes and resulting in poor model performance. In our approach, this deficiency is addressed by incorporating text information, which provides unique discriminative features for each class.

Qualitative Comparison. We report qualitative results generated by our method on the COCO-20^i dataset in Fig.2. From the first row, we can see that given a support image containing only a teddy bear, our model VTSeg can correctly segment the query image even in the presence of interference from the "person" class. This demonstrates that incorporating textual information enhances the discriminative power of our model. The second row shows that our model VTSeg can segment the target class even if it only occupies a small part of the image. Sometimes, there may be perspective distortion [12] among different images within the same class, as shown in the third and fourth rows of Fig. 2. We found that our model VTSeg performs well in the presence of this issue. The reason is that VTSeg does not solely rely on the image features limited by the support image; it also leverages global and discriminative information from text.

4.4 Ablation Studies

To validate the effectiveness of our proposed modules, we conducted ablation experiments on PASCAL-5^i 1-shot using ResNet50 backbone. The results are shown in Table 2 where *Corr* and *VEP* represent Vision-Language Guided Correlation Module and Vision-Enriched Prompt Module, respectively. Following [22,23], we build our baseline using ResNet50 as the image feature extractor. In the second row, we first show the improvement over baseline with *Corr*, which leverages additional discriminative information provided by text prompt. We then enrich the text prompt with support features, enabling the text to incorporate class-specific information to better distinguish between different classes. With the combined effect of these two modules, our model's performance has improved by 6.45% over the baseline.

5 Conclusion

In this paper, we explore the feasibility of multimodal approaches in Few-Shot Segmentation (FSS). By employing the Vision-Enriched Prompt Module, we enable text prompts to perceive contextual information and enhance their generalization across different datasets. Additionally, we propose a Vision-Language

Guided Correlation Module to activate segmentation targets in query images through the fusion of enriched text prompts and image features. Extensive experiments validate the feasibility of multimodal approaches in FSS and achieve significant performance improvements on the PASCAL-5^i and COCO-20^i datasets. **Limitation and Future Work.** Although our approach outperforms existing methods on various datasets, there are still some limitations that need to be addressed. The performance of our model is highly dependent on the quality and relevance of the text prompts used for fusion, which may not always be optimal. For future work, we plan to explore more efficient ways to integrate textual and visual information. Additionally, we aim to improve the model's ability to generalize to novel classes by enhancing the diversity of the training data and by incorporating data augmentation techniques.

6 Acknowledgement

This work is supported by grants from the National Natural Science Foundation of China (NSFC 62322215, 62402488), the Excellent Young Scientists Fund in Hunan Province (2022JJ20077), and the Scientific Research Fund of Hunan Provincial Education Department (22A0007).

References

1. AA. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., "Segment anything," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015-4026, (2023).
2. L. Ke, M. Ye, M. Danelljan, Y.-W. Tai, C.-K. Tang, F. Yu, et al., "Segment anything in high quality," Advances in Neural Information Processing Systems, vol. 36, 2024.
3. L. Jiang, S. Shi, Z. Tian, X. Lai, S. Liu, C.-W. Fu, and J. Jia, "Guided point contrastive learning for semi-supervised point cloud semantic segmentation. in Proceedings of the IEEE/CVF international conference on computer vision, pp. 6423-6432, 2021.
4. X. Lai, Z. Tian, L. Jiang, S. Liu, H. Zhao, L. Wang, and J. Jia, "Semi-supervised semantic segmentation with directional context-aware consistency," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1205-1214, 2021.
5. J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," Advances in neural information processing systems, vol. 30, 2017.
6. F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1199-1208, 2018.
7. Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," ACM computing surveys (csur), vol. 53, no. 3, pp. 1–34, 2020.
8. N. Dong and E. P. Xing, "Few-shot semantic segmentation with prototype learning.," in BMVC, vol. 3, 4, 2018.
9. 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, pp. 9197-9206, 2019.

10. J. Min, D. Kang, and M. Cho, "Hypercorrelation squeeze for few-shot segmentation," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 6941-6952, 2021.
11. K. Rakelly, E. Shelhamer, T. Darrell, A. A. Efros, and S. Levine, "Conditional networks for few-shot semantic segmentation," in International Conference on Learning Representations, 2018.
12. Y. Yang, Q. Chen, Y. Feng, and T. Huang, "Mianet: Aggregating unbiased instance and general information for few-shot semantic segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7131-7140, 2023.
13. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2881-2890, 2017.
14. Z. Tian, H. Zhao, M. Shu, Z. Yang, R. Li, and J. Jia, "Prior guided feature enrichment network for few-shot segmentation," IEEE transactions on pattern analysis and machine intelligence, vol. 44, no. 2, pp. 1050-1065, 2020.
15. B. Yang, C. Liu, B. Li, J. Jiao, and Q. Ye, "Prototype mixture models for few-shot semantic segmentation," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VIII 16, pp. 763-778, Springer, 2020.
16. 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8334-8343, June 2021.
17. 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, pp. 9197-9206, 2019.
18. 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/CVF conference on computer vision and pattern recognition, pp. 8334-8343, 2021.
19. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in International conference on machine learning, pp. 8748-8763, PMLR, 2021.
20. Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu, "Denseclip: Language-guided dense prediction with context-aware prompting," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
21. C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in European Conference on Computer Vision, pp. 696-712, Springer, 2022.
22. Y. Liu, N. Liu, Q. Cao, X. Yao, J. Han, and L. Shao, "Learning nontarget knowledge for few-shot semantic segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11573-11582, June 2022.
23. C. Lang, G. Cheng, B. Tu, and J. Han, "Learning what not to segment: A new perspective on few-shot segmentation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8057-8067, 2022.
24. J. Liu, Y. Bao, G.-S. Xie, H. Xiong, J.-J. Sonke, and E. Gavves, "Dynamic prototype convolution network for few-shot semantic segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11553-11562, 2022.

25. H. Liu, P. Peng, T. Chen, Q. Wang, Y. Yao, and X.-S. Hua, “Fecanet: Boosting few-shot semantic segmentation with feature-enhanced context-aware network,” IEEE Transactions on Multimedia, vol. 25, pp. 8580-8592, 2023.
26. R. Cong, H. Xiong, J. Chen, W. Zhang, Q. Huang, and Y. Zhao, “Query guided prototype evolution network for few-shot segmentation,” IEEE Transactions on Multimedia, 2024.
27. C. Lang, G. Cheng, B. Tu, and J. Han, “Few-shot segmentation via divide-and-conquer proxies,” International Journal of Computer Vision, vol. 132, no. 1, pp. 261-283, 2024.
28. X. Bao, J. Qin, S. Sun, X. Wang, and Y. Zheng, “Relevant intrinsic feature enhancement network for few-shot semantic segmentation,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 765-773, 2024.
29. A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, “One-shot learning for semantic segmentation,” arXiv preprint arXiv:1709.03410, 2017.
30. K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Conditional prompt learning for vision-language models,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16816-16825, 2022.
31. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.
32. G. Zhang, G. Kang, Y. Yang, and Y. Wei, “Few-shot segmentation via cycle-consistent transformer,” Advances in Neural Information Processing Systems, vol. 34, pp. 21984-21996, 2021.
33. F. Guo, YK. Wang, H. Qi, W. Jin, L. Zhu, J. Sun, "Multi-view distillation based on multi-modal fusion for few-shot action recognition (CLIP-MDMF)," Knowledge-Based Systems, vol. 304, 112539, 2024.
34. A. Azeem, Z. Li, A. Siddique, Y. Zhang, S. Zhou, "Unified multimodal fusion transformer for few shot object detection for remote sensing images," Information Fusion, vol. 111, 102508, 2024.
35. T. Lüddecke, A. Ecker. “Image Segmentation Using Text and Image Prompts.” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7086-7096, 2022.
36. A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, et al. “A Review on Deep Learning Techniques Applied to Semantic Segmentation.” arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition, Apr. 2017.
37. Y. Guo, Y. Liu, T. Georgiou, MS. Lew. “A Review of Semantic Segmentation Using Deep Neural Networks.” International Journal of Multimedia Information Retrieval, Vol. 7, pp. 87–93, 2018.