Self-guided Few-Shot Semantic Segmentation for Remote Sensing Imagery Based on Large Vision Models

Xiyu Qi[1,2,4], Yifan Wu[1,3,4], Yongqiang Mao[1,2,4], Wenhui Zhang[1,2,4], and Yidan Zhang[1,2(✉)]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
zhangyd@aircas.ac.cn

[2] Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China

[3] Key Laboratory of Technology in Geo-spatial Information Processing and Application System, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China

[4] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China

Abstract. The Segment Anything Model (SAM) exhibits remarkable versatility and zero-shot learning abilities, owing largely to its extensive training data (SA-1B). Recognizing SAM's dependency on manual guidance given its category-agnostic nature, we identified unexplored potential within few-shot semantic segmentation tasks for remote sensing imagery. This research introduces a structured framework designed for the automation of few-shot semantic segmentation. It utilizes the SAM model and facilitates a more efficient generation of semantically discernible segmentation outcomes. Central to our methodology is a novel automatic prompt learning approach, leveraging prior guided mask to produce coarse pixel-wise prompts for SAM. Extensive experiments on the DLRSD datasets underlines the superiority of our approach, outperforming other available few-shot methodologies.

Keywords: Remote sensing images · visual foundation model · semantic segmentation · prompt learning

1 Introduction

Recent advancements in remote sensing technologies [18,21,34] have revolutionized the way we collect and analyze data relevant to our understanding of the

X. Qi and Y. Wu—Contributed equally to this work and should be considered co-first authors.

W. Yu and L. Xuan (Eds.): MLICOM 2023, LNICST 518, pp. 69–80, 2024.
https://doi.org/10.1007/978-3-031-71716-1_6

Earth's surface. The data collected from these methods serve central to a multitude of applications such as weather forecasting [9], environmental monitoring [15], urban planning [32], and defense intelligence [20]. Despite the wealth of data made available through remote sensing, effective utilization of this information poses a significant challenge. One bottleneck lies in the semantic segmentation of the remote sensing imagery [22,35]. Traditional methods rely heavily on manual guidance, which is both time-inefficient and susceptible to human error.

The Segment Anything Model (SAM) [14], an example of large vision models, demonstrates considerable versatility and zero-shot learning [24] abilities, powered by its rich training data. However, its ability to perform one-shot or few-shot semantic segmentation tasks for remote sensing imagery [4,30] remains largely underexplored. The SAM model's design is category-agnostic, which inevitably forces reliance on manual guidance. Recognizing the potential and the need to optimize this aspect, we ventured into the integration of few-shot learning in this regard.

In this research, we introduce a novel few-shot semantic segmentation algorithm designed to automate the semantic segmentation process termed Self-guided Large Vision Model (Few-shot SLVM). Our approach enables the use of the SAM model with the intent to achieve efficient generation of semantically rich segmentation outcomes. The cornerstone of this method is an innovative automatic prompt learning technique that leverages prior guided masks to produce coarse pixel-wise prompts for SAM, bypassing the need for intensive manual guidance.

The framework we propose for few-shot semantic segmentation provides a promising avenue for the efficient parsing of remote sensing imagery. The system's capacity to produce high-quality segmented images with limited supervision is bound to drive advancements in various applications dependent on remote sensing. We focused our experiments on the DLRSD [26] datasets with the goal of validating the superiority of this approach over other existing methodologies for few-shot semantic segmentation, particularly in the context of remote sensing.

To summarize, our major contributions are:

1) We introduce the Self-guided Large Vision Model (Few-shot SLVM), a novel few-shot semantic segmentation framework, that significantly automates the segmentation process for remote sensing imagery without heavy reliance on manual guidance.
2) We propose an innovative 'automatic prompt learning' technique using the Segment Anything Model (SAM) for rendering coarse pixel-wise prompts, bringing a novel solution to semantic segmentation of remote sensing imagery.
3) We carry out extensive benchmarking on the DLRSD datasets, showcasing the superiority of our methodology against existing few-shot segmentation techniques within the domain of remote sensing.

2 Related Work

In this section, we initially introduce recent work in the field of semantic segmentation, laying the foundation for essential context with regards to our proposed method. Subsequently, we steer our focus towards the few-shot semantic segmentation techniques that underpin our proposed method while concluding this section with a discussion on the few-shot learning approaches that are explicitly designed for semantic segmentation of aerial imagery.

2.1 Semantic Segmentation for Visual Scenes

Semantic segmentation forms a critical research area within computer vision, carrying substantial impact on interpreting visual scenes. Several works executed on traditional datasets [14,19] have relied heavily on conventional Convolutional Neural Networks (CNNs), yielding valuable but large-dataset-dependent methods [7,36]. FCN (Fully Convolutional Network) [17], as one of the pioneering CNN-based segmentation methods, has played a crucial role in shaping the development of Encoder-Decoder networks. Building upon the rich foundation of FCN and integrating skip connections, U-Net [25] has emerged as an effective solution for various biomedical segmentation tasks, including cell and nucleus segmentation, tumor detection, and organ segmentation. DeepLab [1–3], an improvement over FCN, exploits atrous convolution to enlarge the receptive field of convolutional neural networks without inflating the number of parameters. These aforementioned CNN network architectures have exhibited notable performance in semantic segmentation tasks. Due to their remarkable achievements, CNNs have attracted significant attention and extensive research efforts. Densely Connected Convolutional Network (DenseNet) [11] and ShuffleNet [37] are among the CNN-based models that have been explored for image semantic segmentation. While these methods offer valuable insights, they often encounter challenges concerning time-efficiency, primarily due to their heavy reliance on large-scale annotated datasets.

2.2 Few-Shot Learning and Large Vision Models

Few-shot learning has emerged as an effective approach for semantic segmentation, as it can generalize from limited data [28,31]. The success of few-shot learning heavily depends on the transferability of the trained neural networks. In terms of distance measurement, certain metric learning methods aim to learn a metric space by computing distances between instances and new categories [27,30]. Additionally, meta-learning has been proposed to enhance the few-shot adaptation ability of models by discovering a set of initialized parameters that can rapidly adapt to novel domains [5]. In recent years, Large Vision Models, like GPT-3 [6] and CLIP [23], have sparked interest in the field of few-shot learning due to their impressive performance in numerous visual and text-based tasks, including SAM [14] due to its strong zero-shot adaptation performance. CoOp [39] introduces prompt learning concept, borrowed from natural language processing (NLP), into the vision domain to adapt pre-trained vision-language models for better performance. CoCoOp [38] takes this a step further by learning a lightweight neural network that generates an input-conditional token (vector) for each image. This dynamic prompt adapts to each instance, making the model less sensitive to class shift. The Clip-Adapter [8] is proposed to adapt the Adapter to specific downstream tasks by fine-tuning it while keeping the CLIP [23] model pre-training weights unchanged. Furthermore, several subsequent works have been proposed [12,16,29] to further tailor large visual models for diverse visual tasks. In contrast to these approaches, our primary focus in this study is on the few-shot semantic segmentation task of the visual language model for remote sensing imagery.

2.3 Few-Shot Learning in Semantic Segmentation of Remote Sensing Imagery

Efforts focused on the application of few-shot learning for semantic segmentation of remote sensing imagery are limited [19,30]. Jiang et al. [13] propose a deep metric learning-based method that maps images into a feature space where the distance reflects semantic similarity. This method enables the model to effectively identify and segment new classes with only a few examples. DMML-Net [30] addresses the issue by formulating segmentation as a metric-based pixel classification task. The network architecture includes a deep feature pyramid comparison network that supports multiscale metric learning, allowing the model to learn a robust feature representation at multiple scales. Chen et al. [4] introduce a contrastive learning framework for few-shot learning scenarios, aimed at learning generalizable and discriminative features from both labeled and unlabeled data. While studies by [10,33] have shown promising gains in using SAM for semantic segmentation tasks in medical imaging, the application of large vision models remains largely limited to medical imaging, despite its high potential applicability and relevance to remote sensing imagery.

3 Methodology

Our methodology outlines the systematic integration of the Segment Anything Model (SAM), Prior Guided Metric Learning, and an innovative few-shot learning setup to effectively automate semantic segmentation within the setting of remote sensing imagery. To provide more insight, we denote the training dataset as $D = \{I, T\}$, support set as $D_s = \{I_s, T_s\}$, and query set as $D_q = \{I_q, T_q\}$, where $I = \{i^1, ..., i^n\}$ represents images, and $T = \{t^1, ..., t^n\}$ corresponds to their segmentation ground truth. The primary objective is to design a plug-and-play self-prompting module, enabling SAM to obtain the location and size information of the segmentation target, only necessitating a few labeled data, for instance k images (Fig. 1).

3.1 Prior Guided Metric Learning

Assuming an input image X and the output segmentation mask Y, with the encoder function $E(\cdot)$ and decoder function $D(\cdot)$ of the Segment Anything Model (SAM), we can define the process of generating the mask for SAM as follows:

$$Y = D(E(X)) \tag{1}$$

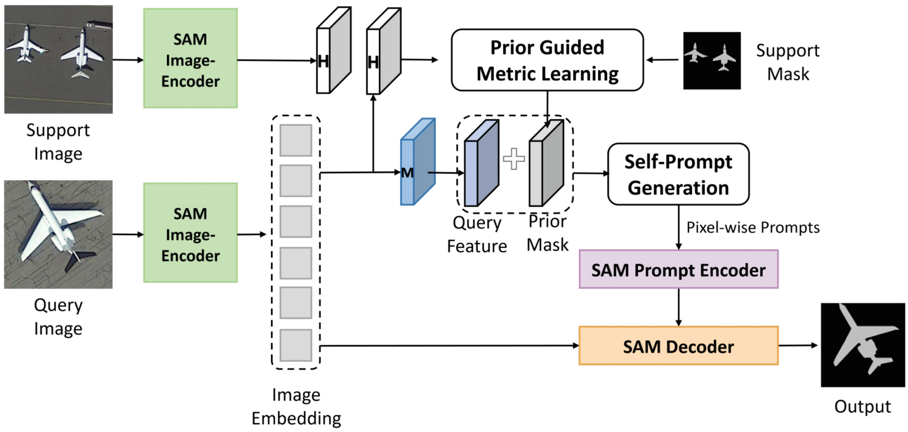

Fig. 1. The overview of our proposed Few-shot Semantic Segmentation framework based on Self-guided Large Vision Model (Few-shot SLVM). We utilize the pretrained Large Vision Model, SAM, to extract both high-level (H) and intermediate (M) semantic features from support and query images. The pre-trained high-level features with a support mask are transformed into prior mask utilizing cosine similarity measures. We take the query features and generated prior mask as input to produce coarse pixel-wise prompts for SAM. During the training process, the model is trained using the image embeddings from the SAM encoder and the resized ground truth label, while the cumbersome encoder, prompt encoder, and decoder parts of the SAM structure are kept frozen.

where E represents a pre-trained image embedding encoder and D is a learned mask decoder, the Prior Guided Metric Learning Module is introduced to incorporate prior information with the prompt. Specifically, after passing through the powerful encoder of the large vision model SAM, we first perform the Hadamard product between the high-level support features $E_H(I_S)$ and the mask M. Subsequently, we use cosine similarity to calculate the pixel-wise association between the high-level query features $E_H(I_Q)$ and the mask-weighted support features, defined as:

$$P = cosine(E_H(I_Q), E_H(I_S) \odot M) \tag{2}$$

By concatenating the intermediate query features $E_M(I_Q)$ with the pixel-level prior-guided information P, new query features are generated to effectively integrate the support information with the prior information, resulting in enhanced segmentation results.

3.2 Self-prompt Generation

In this step, we design a novel automatic prompt learning method that generates coarse pixel-wise prompts for SAM from the prior mask Y_P to guide the segmentation prediction. To ensure the spatial alignment between the dense mask prompt and the image, the mask is initially fed to the network at a resolution 4 times lower than the original image. Further downsampling is done through two 2×2 convolutions with a stride of 2, yielding output channels of 4 and 16, respectively. Following [14], each layer is separated by GELU activations and layer normalization. Then, the mask and image embeddings are combined through element-wise addition.

To effectively map image embeddings and dense mask prompt embeddings to the output mask, a learned output token embedding is initially incorporated into the prompt embeddings set. In order to provide the decoder with essential geometric information, the dense mask prompt embeddings are utilized as position encodings. These position encodings are then added to the image embeddings as guiding cues whenever the attention layer incorporates the position encoding. Following the execution of the decoder, the updated image embedding undergoes unsampling by a factor of $4\times$ using two transposed convolutional layers. Subsequently, the tokens re-attend to the image embedding, and the modified output token embedding is fed into a compact MLP consisting of three layers. This MLP generates a vector that aligns with the channel dimension of the upscaled image embedding. Finally, the model predicts a mask by performing a spatially point-wise product between the upscaled image embedding and the output of the MLP. In this step, we introduce a prompt indicator W to assign weights to the guided embeddings. The prompt indicator W is calculated based on the confidence of the prior mask Y_P. As a result, the reformulated output mask Y is obtained using the following formulation:

$$Y = D(W \odot E(X), (1 - W) \odot E(Y_P)) \tag{3}$$

During the training process, both the cumbersome encoder and decoder of SAM are kept frozen, guiding it to focus on the area of interest through the continuous optimization of self-guided prompt embeddings.

3.3 Few-Shot Learning Adaptation

Our proposed Few-shot Self-guided Large Vision Model (SLVM) functions by learning from a limited set of support examples and extrapolating this learning to the query set. For this, we employ a cosine similarity loss function, defined as:

$$L = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{Y_i \cdot T_i}{\|Y_i\| \cdot \|T_i\|}\right) \tag{4}$$

Lastly, to enhance performance, we introduce a fine-tuning strategy. The training objective comprises both the Self-guidance loss L_s and the Fine-tuning loss L_f, represented as:

$$L_{total} = \alpha L_s + \beta L_f \tag{5}$$

The strategy operates with a two-fold phase that initially trains with self-guidance loss only and then fine-tunes using the total loss. Through this detailed walkthrough of our method, we lay out the blueprint of our Few-shot SLVM

Table 1. Comparisons of few-shot segmentation performance between our proposed Few-shot SLVM and other methods under different splits on the DLRSD dataset.

Methods	1-Shot					5-Shot				
	Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
CANet	25.31	12.55	18.41	26.66	20.73	28.29	17.10	21.36	29.45	24.05
PANet	36.15	20.55	26.98	38.41	30.52	40.85	23.61	35.87	45.67	36.50
DMML-Net	45.03	31.23	47.38	**47.17**	42.70	57.23	39.86	56.62	**62.60**	54.08
Ours	**50.70**	**34.15**	**50.47**	43.64	**44.74**	**61.90**	**52.20**	**58.72**	60.05	**58.22**

model's capability to combine the power of SAM, few-shot learning, and prior metric learning for semantic segmentation in remote sensing images.

4 Experiments

4.1 Datasets

DLRSD dataset [26] employed in the experiments comprises 2100 high-resolution aerial images, with each image having dimensions of 256×256 pixels. These images encompass 17 distinct object classes, including airplanes, bare soil, buildings, cars, and various others. Each sample in the dataset is labeled with pixel-level annotations, providing detailed ground truth information for precise object segmentation. The dataset poses several challenges encountered in real-world scenarios, such as occlusion, shadows, and variations in terrain scales, making it a valuable resource for evaluating algorithms robustness. Similar to the methodology of Wang et al. [30], the DLRSD dataset is partitioned into four separate folds. The first three folds consist of four categories each, while the fourth fold includes five categories, namely sea, ship, tank, tree, and water. This partitioning allows for a more comprehensive evaluation and analysis of the proposed method's performance on different object classes and scenarios.

4.2 Implement Details

For the following experiments, we employ the SAM ViT-Huge model as our backbone. During the training phase, we utilize the PyTorch framework and trained end-to-end with the AdamW optimizer. We use a mini-batch size of 8 and set the initial learning rate to 0.00025. To decay the learning rate, we employ a Cosine Annealing scheduler and the momentum is set to 0.9. For data augmentation, we randomly perform flipping (vertically or horizontally) and rotation operations on the input images, with a resulting size of 256×256 pixels. All experiments are conducted on 2 NVIDIA GeForce RTX 3090 Ti GPUs, and the training process lasts for 1000 epochs. By utilizing appropriate pretrained backbones and carefully setting hyperparameters, we achieve optimal performance in terms of accuracy and efficiency in our experiments.

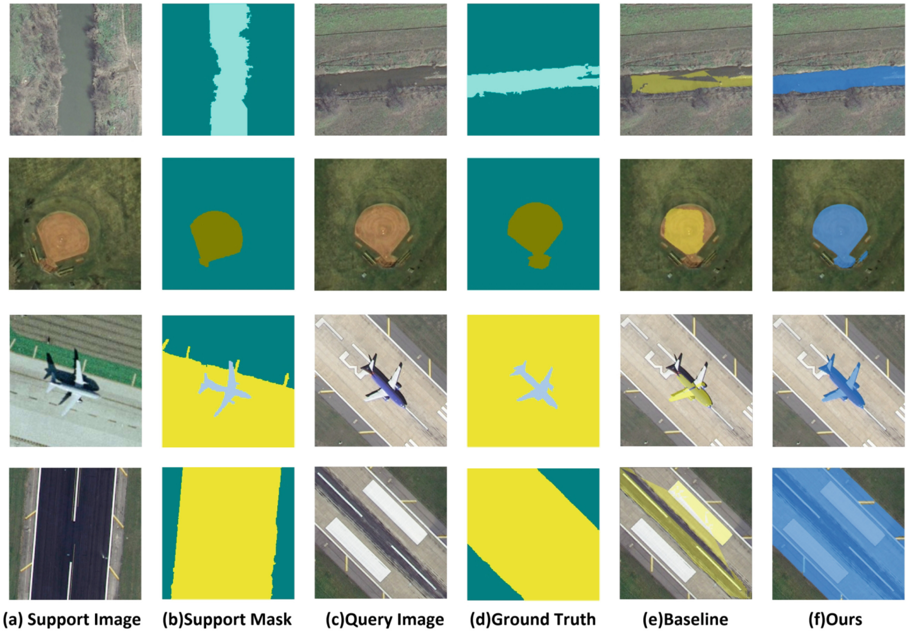

Fig. 2. Qualitative results of the baseline and our proposed Few-shot Semantic Segmentation framework based on Self-guided Large Vision Model (Few-shot SLVM). The samples are from iSAID-5^i dataset. (a) Support images. (b) Support masks. (c) Query images. (d) Ground truth of query images. (e) Predictions of baseline. (f) Predictions of ours.

4.3 Results

To provide a comprehensive evaluation, we compared our proposed Few-shot SLVM with three other state-of-the-art few-shot segmentation methods evaluated on the DLRSD dataset, taking into account the 1-shot and 5-shot settings, and utilizing mean intersection over union (mIoU) as the evaluation metric.

Comparative Analysis: As Table 1 depicts, Few-shot SLVM surpasses competing methodologies, underscoring its robustness and accuracy. The mean IoU at the 1-shot and 5-shot settings improved by 2.04% and 4.14% over the closest competing models, which demonstrates the model's superior segmentation quality. Particularly, in the case of chaparral, court, dock, and field categories within fold-1, which encompass diverse objects and background regions with indistinct boundaries and complex scale variations, our proposed method exhibited a 12.34% performance boost in comparison to the second-ranked method. This substantiates the notion that our method effectively incorporates more intricate details and features, enhancing its perceptual capabilities to tackle challenges such as blurred boundaries and complex scale changes.

Qualitative Analysis: As shown in Fig. 2, we also provide a qualitative comparison of the results between the proposed Few-shot SLVM and the baseline for a more thorough understanding of the few-shot learning representation. The visualized results reveal notable disparities. Specifically, the segmentation outcomes obtained by the baseline model exhibit significant issues of semantic boundary ambiguity in a wide range of complex object regions. In contrast, the proposed method achieves more comprehensive and accurate segmentation results, particularly in terms of boundaries. Moreover, our approach effectively mitigates the problem of missing target areas observed in the baseline model for certain ground objects.

4.4 Ablation Study

We performed an ablation study to understand the contribution of each component in our model. We focused on three main modules: (1) Automatic Prompt Learning (APL), (2) Prior Guided Metric Learning (PGML), and (3) Few-Shot Learning Adaptation (FSLA). Results from the ablation study:

Table 2. Ablation study on the proposed components of our method on the DLRSD dataset under one-shot setting.

Methods			Fold-0	Fold-1	Fold-2	Fold-3	Mean
APL	PGML	FSLA					
–	–	–	39.02	21.19	38.70	36.04	33.74
✓			46.77	27.30	46.44	41.85	40.59
✓	✓		50.66	31.35	45.03	42.16	42.30
✓	✓	✓	**50.70**	**34.15**	**50.47**	**43.64**	**44.74**

Ablation Study Insights: The stark performance drop in the absence of 'Automatic Prompt Learning' (Table 2) confirms its centrality in generating precise segmentation prompts, contributing to an overall 6.85% increase in Mean IoU. Similarly, disabling 'Few-Shot Learning Adaptation' leads to performance degradation, highlighting its role in enhancing the model's adaptability to limited data scenarios.

Module Effectiveness: 1. Automatic Prompt Learning: The value of the APL component is evident from the substantial uptick in the overall accuracy. Through the automation of prompt generation, it diminishes reliance on manual interventions. This hastens the segmentation workflow and either maintains or heightens the quality of the results. The difference in the mean IoU between the model with no components and the one with APL indicates a marked improvement of 6.85%, underscoring the importance of APL. 2. Prior Guided Metric Learning (PGML): Delving into the performance with and without the PGML, the significance of this module comes to light. When we compare the model with only APL to the one equipped with both APL and PGML, there's a modest increase in Mean IoU from 40.59% to 42.30%. This suggests that PGML refines the feature representations and contributes to better distance metrics. By incorporating prior knowledge, PGML offers a more informed and directed approach to metric learning, thus making the model more resilient and effective in differentiating between classes. 3. Few-Shot Learning Adaptation: The FSLA component bolsters the model's ability to learn from limited data. This trait is paramount, especially in remote sensing scenarios where access to abundant labeled data might be constrained. Its role is apparent in the enhancement of the model's generalization capabilities, especially on classes that are either unseen or minimally represented. The final row of the table demonstrates that integrating FSLA boosts the mean IoU to 44.74%, the best among the configurations.

5 Conclusion

In conclusion, this research emphasizes the advancements brought forth by the Few-Shot Self-guided Large Vision Model (Few-shot SLVM) in few-shot semantic segmentation for remote sensing imagery. The Few-shot SLVM integrates three pivotal modules: Automatic Prompt Learning (APL), Prior Guided Metric Learning (PGML), and Few-Shot Learning Adaptation (FSLA). APL streamlines the segmentation process, reducing manual input and enhancing accuracy. PGML optimizes feature representation, refining class differentiation and subsequently boosting the mean IoU. FSLA excels in limited data scenarios, improving model generalization for rare or unseen classes. Collectively, these components elevate the Few-shot SLVM to deliver state-of-the-art results in few-shot remote sensing segmentation. Future endeavors will focus on refining these modules and extending their applicability across broader datasets for robust model validation.

References

1. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv preprint arXiv:1412.7062 (2014)
2. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
4. Chen, Y., Wei, C., Wang, D., Ji, C., Li, B.: Semi-supervised contrastive learning for few-shot segmentation of remote sensing images. Remote Sens. **14**(17), 4254 (2022)
5. Chen, Y., Liu, Z., Xu, H., Darrell, T., Wang, X.: Meta-baseline: exploring simple meta-learning for few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9062–9071 (2021)
6. Dale, R.: GPT-3: what's it good for? Nat. Lang. Eng. **27**(1), 113–118 (2021)
7. Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C.: ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS J. Photogramm. Remote. Sens. **162**, 94–114 (2020)

8. Gao, P., et al.: CLIP-Adapter: better vision-language models with feature adapters. Int. J. Comput. Vis., 1–15 (2023)
9. Gherboudj, I., Ghedira, H.: Assessment of solar energy potential over the United Arab Emirates using remote sensing and weather forecast data. Renew. Sustain. Energy Rev. **55**, 1210–1224 (2016)
10. He, S., Bao, R., Li, J., Grant, P.E., Ou, Y.: Accuracy of segment-anything model (SAM) in medical image segmentation tasks. arXiv preprint arXiv:2304.09324 (2023)
11. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
12. Huang, T., Chu, J., Wei, F.: Unsupervised prompt learning for vision-language models. arXiv preprint arXiv:2204.03649 (2022)
13. Jiang, X., Zhou, N., Li, X.: Few-shot segmentation of remote sensing images using deep metric learning. IEEE Geosci. Remote Sens. Lett. **19**, 1–5 (2022)
14. Kirillov, A., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
15. Li, J., Pei, Y., Zhao, S., Xiao, R., Sang, X., Zhang, C.: A review of remote sensing for environmental monitoring in China. Remote Sens. **12**(7), 1130 (2020)
16. Lin, Z., et al.: Frozen clip models are efficient video learners. In: Avidan, S., Brostow, G., Cisse, M., Farinella, G.M., Hassner, T. (eds.) European Conference on Computer Vision, pp. 388–404. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19833-5_23
17. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
18. Mao, Y., et al.: Beyond single receptive field: a receptive field fusion-and-stratification network for airborne laser scanning point cloud classification. ISPRS J. Photogramm. Remote. Sens. **188**, 45–61 (2022)
19. Mao, Y., Guo, Z., Xiaonan, L., Yuan, Z., Guo, H.: Bidirectional feature globalization for few-shot semantic segmentation of 3D point cloud scenes. In: 2022 International Conference on 3D Vision (3DV), pp. 505–514. IEEE (2022)
20. Michael, K., Masters, A.: Realized applications of positioning technologies in defense intelligence. In: Applications of Information Systems to Homeland Security and Defense, pp. 167–195. IGI Global (2006)
21. Moselhi, O., Bardareh, H., Zhu, Z.: Automated data acquisition in construction with remote sensing technologies. Appl. Sci. **10**(8), 2846 (2020)
22. Qi, X., Mao, Y., Zhang, Y., Deng, Y., Wei, H., Wang, L.: PICS: paradigms integration and contrastive selection for semisupervised remote sensing images semantic segmentation. IEEE Trans. Geosci. Remote Sens. **61**, 1–19 (2023)
23. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
24. Romera-Paredes, B., Torr, P.: An embarrassingly simple approach to zero-shot learning. In: International Conference on Machine Learning, pp. 2152–2161. PMLR (2015)
25. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds.) Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015, Proceedings, Part III 18, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

26. Shao, Z., Yang, K., Zhou, W.: A benchmark dataset for performance evaluation of multi-label remote sensing image retrieval. Remote Sens. **10**(6) (2018)
27. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
28. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: 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)
29. Udandarao, V., Gupta, A., Albanie, S.: SuS-X: training-free name-only transfer of vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2725–2736 (2023)
30. Wang, B., Wang, Z., Sun, X., Wang, H., Fu, K.: DMML-Net: deep metametric learning for few-shot geographic object segmentation in remote sensing imagery. IEEE Trans. Geosci. Remote Sens. **60**, 1–18 (2021)
31. Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: a survey on few-shot learning. ACM Comput. Surv. (CSUR) **53**(3), 1–34 (2020)
32. Wellmann, T., et al.: Remote sensing in urban planning: contributions towards ecologically sound policies? Landsc. Urban Plan. **204**, 103921 (2020)
33. Wu, J., et al.: Medical SAM adapter: adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
34. Wu, Y., Huang, L., Zhang, B., Wang, X., Qi, X.: An accurate and efficient BP algorithm based on precise slant range model and rapid range history construction method for GEO SAR. Remote Sens. **15**(21), 5191 (2023)
35. Yi, Y., Zhang, Z., Zhang, W., Zhang, C., Li, W., Zhao, T.: Semantic segmentation of urban buildings from VHR remote sensing imagery using a deep convolutional neural network. Remote Sens. **11**(15), 1774 (2019)
36. Yuan, X., Shi, J., Gu, L.: A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Syst. Appl. **169**, 114417 (2021)
37. Zhang, X., Zhou, X., Lin, M., Sun, J.: ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856 (2018)
38. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16816–16825 (2022)
39. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. Int. J. Comput. Vision **130**(9), 2337–2348 (2022)