RSAM-Seg: A SAM-based Approach with Prior Knowledge Integration for Remote Sensing Image Semantic Segmentation

Jie Zhang, Xubing Yang, Rui Jiang, Wei Shao and Li Zhang

***Abstract*—The development of high-resolution remote sensing satellites has provided great convenience for research work related to remote sensing. Segmentation and extraction of specific targets are essential tasks when facing the vast and complex remote sensing images. Recently, the introduction of Segment Anything Model (SAM) provides a universal pre-training model for image segmentation tasks. While the direct application of SAM to remote sensing image segmentation tasks does not yield satisfactory results, we propose RSAM-Seg, which stands for Remote Sensing SAM with Semantic Segmentation, as a tailored modification of SAM for the remote sensing field and eliminates the need for manual intervention to provide prompts. Adapter-Scale, a set of supplementary scaling modules, are proposed in the multi-head attention blocks of the encoder part of SAM. Furthermore, Adapter-Feature are inserted between the Vision Transformer (ViT) blocks. These modules aim to incorporate high-frequency image information and image embedding features to generate image-informed prompts. Experiments are conducted on four distinct remote sensing scenarios, encompassing cloud detection, field monitoring, building detection and road mapping tasks . The experimental results not only showcase the improvement over the original SAM and U-Net across cloud, buildings, fields and roads scenarios, but also highlight the capacity of RSAM-Seg to discern absent areas within the ground truth of certain datasets, affirming its potential as an auxiliary annotation method. In addition, the performance in few-shot scenarios is commendable, underscores its potential in dealing with limited datasets. Our code is available at: https://chief-byte.github.io/RSAM-Seg-Site.**

***Index Terms*—Segmentation, Deep learning, Segment Anything Model, Remote sensing image.**

I. Introduction

With the development of remote sensing satellite technology, high-resolution remote sensing images have been widely used in various fields, such as cloud detection, urban infrastructure assessment, agricultural land planning, and road condition analysis [1]–[5]. Cloud detection plays a pivotal role as the initial step in the data processing pipeline for earth observation and remote sensing technologies [6]. Urban infrastructure assessment leverages remote sensing capabilities to evaluate a diverse range of structures, including buildings, roads, and bridges, supporting maintenance and planning efforts [7]. Furthermore, in the realm of agricultural land planning, remote sensing assumes a crucial function by monitoring crop health, scrutinizing land use patterns, and optimizing irrigation management to enhance farming practices [8]. However, the satellite images often suffer from object occlusion, blurring, incomplete coverage and other issues, which pose challenges for identifying objects [9].

Thus, a multitude of methods have been proposed to address above issues, which can be generally divided into three categories: threshold-based algorithms, classical machine learning algorithms and deep learning algorithms [10]–[14]. Threshold-based algorithms utilize the spectral characteristics of remote sensing data to complete semantic segmentation tasks based on prior knowledge and judgment conditions provided by experts [15]–[18], Li *et al.* proposed an object-based approach to create a land-cover classification map. However, threshold-based methods rely on a large number of pre-designed rules and require remote sensing professionals to design and evaluate the rules, which leads to problems such as high costs, long processing times, and poor results [19]. In terms of classical machine learning algorithms, Support Vector Machines (SVM) and Random Forest (RF) have gained much attention [20]–[22]. Melgani *et al.* assessed the potential of SVM classifiers in high-dimensional feature spaces of hyperspectral remote sensing images [21]. The experimental results suggested that SVM classifiers are a viable option for classifying hyperspectral remote sensing data. However, SVM is subject to certain limitations. The choice of kernel function poses a challenge, where a smaller kernel width parameter may lead to overfitting and a larger value may cause excessive smoothing [23]. RF has been successfully used to map urban buildings and land cover categories [24], [25]. However, RF model trained on one region is not applicable or transferable to new regions [26] and depends on the number of variables used for splitting the tree nodes [27].

With the rapid development of deep learning, this technology has shown great potential in addressing segmentation challenges in remote sensing [28]. However, despite the accomplishments of deep learning methods in remote sensing segmentation, there are still several challenges that demand attention. One prominent challenge lies in the significant within-class variance and limited between-class variance observed

Jie Zhang, Xubing Yang, Li Zhang are with the College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China (e-mail:j.zhang.fn@gmail.com, xbyang, lizhang@njfu.edu.cn)

Rui Jiang is with the College of Telecommunications and InformationEngineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: j ray@njupt.edu.cn).

Wei Shao is with Shenzhen Research Institute, Nanjing University of Aeronautics and Astronautics, Guangdong 518038, China. (shaowei20022005@nuaa.edu.cn)

This work is supported by National Natural Science Foundation of China (NSFC) under grant (No. 61802193) and The Shenzhen Science and Technology Program under Grant JCYJ20220530172403008, (Corresponding author: Wei Shao, Li Zhang.)

in pixel values of objects of interest [29]. In addition, the quality and availability of labeled data play a crucial role in this regard, particularly when dealing with small datasets or rare classes [30]. Insufficient labeled data also hampers the model's ability to generalize well and accurately identify and segment objects of interest. Consequently, there is a notable lack of universality and ease of transfer across different remote sensing scenarios. Overcoming this challenge requires innovative strategies such as domain adaptation, transfer learning, or incorporating prior information that can help the model better leverage the available data and extract meaningful features that are specific to the remote sensing domain to enhance the model's ability to generalize and perform effectively in diverse remote sensing applications [31]–[35].

Recently, the general-purpose vision segmentation models has brought new and more effective solutions to the field of image segmentation [36], [37]. These models are pre-trained on large amounts of data and can be generalized to new tasks and data distributions through the use of prompt engineering, demonstrating outstanding capabilities in few-shot and zero-shot learning [35]–[37]. SAM is a new general-purpose vision segmentation model based on Natural Language Processing (NLP) developed by Meta [38], [39]. It focuses on promptable segmentation tasks and uses prompt engineering to adjust to various downstream segmentation tasks. SAM can automatically identify objects present in an image and immediately provide segmentation masks for any prompt by simply marking points to include or exclude objects, or by drawing bounding boxes to create segmentation [38], which is considered to be a game-changer in the field of image segmentation. Additionally, it achieves fully automated segmentation of potential objects within the images and aims to achieve effective segmentation of any object in any image, without the need for additional task-specific or dataset-specific adaptation (such as training). The segmentation accuracy of SAM on a wide range of diverse benchmark datasets show that SAM has a high generalization ability. Carefully tuned prompts could even surpass popular supervised-training models designed specifically for object segmentation tasks.

Although SAM has shown promising results on open datasets, its effectiveness is often limited when applied to specific downstream tasks, particularly remote sensing segmentation tasks. This is due to the complex characteristics of the interested objects in remote sensing images, such as blurriness, occlusion, and irregular shapes, which pose challenges for segmentation algorithms. In addition, the prompt requires manual input. As a result, there is a need to develop an domain specific SAM that can better handle these challenges and improve the overall performance of remote sensing segmentation tasks without manual prompt input.

To address these issues, we propose RSAM-Seg. Feature information is extracted from specific domains and inserted into the ViT blocks in the encoder to improve the performance in remote sensing field. By incorporating prior knowledge specific to remote sensing image data, such as emedding features and spectral features, the model adjusts better to the segmentation tasks of remote sensing images. To validate the effectiveness of the proposed methodology, experiments were conducted on cloud, buildings, fields and roads scenarios.

The main contributions of the work were summarized as follows:

1). Based on our extensive research and analysis, we have pioneered the application of SAM to object segmentation tasks in remote sensing images, and RSAM-Seg demonstrates better adaptability to remote sensing images.

2). RSAM-Seg eliminates the need for manual intervention to provide prompts, thereby streamlining the workflow of SAM.

3). RSAM-Seg can incorporate custom domain-specific prior information, making it adaptable to diverse tasks in the remote sensing field.

4). RSAM-Seg outperforms the original SAM and U-Net across multiple scenarios such as cloud, buildings, fields and roads in the experiments. Moreover, it discerns missing areas in dataset ground truths and demonstrates few-shot capability.

The rest of this article is organized as follows. Section II provides an overview of the related work in the field of remote sensing segmentation tasks. Section III presents the proposed method for the tasks. In Section IV, the datasets, experimental settings and performance metrics are described in detail. The experimental results and analysis are presented in Section V. Section VI offers a comprehensive discussion of the findings. Finally, Section VII concludes the article by summarizing the main contributions and highlighting future research directions.

II. Related Work

In recent years, deep learning methods have been widely applied to segmentation tasks in remote sensing [40], [41]. Since deep learning networks are typically trained using large datasets to learn specific features and patterns within the input data, which are then used to classify new data, they can be categorized into three types based on the availability of labeled training data: supervised learning, weakly-supervised learning, and unsupervised learning [42].

Over the past ten years of its development, supervised learning has witnessed the emergence of Convolutional Neural Network (CNN). CNN extracts local features from images through convolutional operations, and then reduces the dimensionality of the feature map through pooling operations [42]. DeepLab, a CNN-based model, utilizes techniques such as dilated convolution and multi-scale pyramid pooling to improve segmentation accuracy. It performs well in the field of remote sensing image processing and has been widely used in high-resolution remote sensing image segmentation tasks [43]. DeepLab V3 and DeepLab V3+ are the successors of DeepLab [43], [44]. DeepLab V3 applies global average pooling on the last feature map of the model. DeepLab V3+ brings about a decoder module to the DeepLab V3 to refine the boundary details. Liu *et al.* proposed FieldSeg-DA based on DeepLab V3+ to automatically extract individual arable fields (IAF) from Chinese Gaofen-2 images [44]. U-Net, a neural network based on CNN but utilizing the Encoder-Decoder architecture, has shown excellent performance in the field of image segmentation tasks [45]. Improved networks based on the U-Net structure have gained considerable attention for

their potential in remote sensing image segmentation. Sun *et al.* proposed L-UNet, which replaces the partial convolution layers of U-Net with Conv-LSTM and Atrous in order to improve both the quantity and quality of the network compared to the original U-Net [46]. Hou *et al.* prposed C-UNet on basis of the standard U-Net, where four more modules are added for road extraction tasks and show improved performance compared to standard U-Net [47].

In the field of weakly-supervised learning, semantic segmentation with weak supervision offers a potential solution to address the challenges associated with labeling complexity in land cover classification. Weakly-supervised Semantic Segmentation (WSS) methods often rely on the utilization of Class Activation Maps (CAMs), which is a CNN trained for image-level classification to perform rough localization of object areas based on global average pooling or gradient backpropagation, have been widely used for natural images [23], [48]–[50]. Fu *et al.* proposed WSF-Net, calculates CAMs using fused features of objects in remote sensing image especially in the water and cloud scenarios [51]. Wang *et al.* proposed U-CAM, which adapts CAMs for U-Net to perform cropland segmentation [52]. Nyborg *et al.* proposed the utilization of fix-point Generative Adversarial Network (GAN) for weakly-supervised cloud detection, referred as FCD [53]. Chen *et al.* utilized a WSS framework based on point labels with transfer method to accurately classify land cover with minimal human intervention [54]. Wang *et al.* proposed a novel RS-WSOD framework, which addresses the challenges of background noises and missing detections in remote sensing images [55]. Xu *et al.* proposed the Consistency-Regularized Region-Growing Network (CRGNet) for semantic segmentation of urban scenes, leveraging point-level labels [56].

In terms of unsupervised learning, unsupervised learning addresses the reliance of annotated data and domain shifts in high-resolution remote sensing imagery. Zhu *et al.* proposed Memory Adapt Net (MAN), which established an adversarial learning scheme in output space to bridge the domain distribution discrepancy between the source and the target domains to perform cross-domain segmentation tasks of the high-resolution remote sensing imagery [57]. Chen *et al.* proposed a category-certainty attention mechanism to effectively handle unadapted regions for semantic segmentation of high-resolution satellite imagery [58]. Li *et al.* employed an objective function that integrated multiple weakly supervised constraints to minimize the distributional shift of data between the source and target domains to address challenges related to sensor and landscape variations in diverse geographic locations [59]. Zhang *et al.* proposed a stagewise domain adaptation model called RoadDA that aimed to align the features of the source and target domains through interdomain adaptation using GAN to achieve promising road segmentation on unlabeled target images. [60]. Chen *et al.* proposed a unsupervised domain adaptation method and a contrastive-learning based and Memory-Contracted (MCD) module for building extraction in high-resolution remote sensing imagery [61]. Cai *et al.* proposed the segmentation model from two opposite directions where source domain images are transformed into images featuring the style of the target domain then adapt the classifier to the target domain to improve the performance of the cross-domain semantic segmentation in urban city areas [62].

Recently, few-shot learning, as a nascent method under weakly supervised learning, has gained attention in the field of remote sensing to address the issue of limited datasets [63]. Zhang *et al.* first introduced the concept of few-shot learning [64]. Liu Y *et al.* proposed NTRENet to distinguish ambiguous regions, which is benefit for satellite images [65]. Prior-knowledge based method utilizes pretraining on various other datasets to continuously accumulate learning ability and experience. Domain-specific knowledge is incorporated into the network backbone through various methods on the tasks of few-shot semantic segmentation in aerial images [63], [66], [67]. Cheng *et al.* proposed SPNet to tackle interclass similarity issues in remote sensing scenes during few-shot segmentation by considering the validity of prototypes [68]. Li *et al.* proposed SCL-MLNet to boost few-shot classification in remote sensing scenarios through the fusion of multi-scale spatial features and integration of self-supervised contrastive learning methods [69]. Liu *et al.* enforce the tunable parameters focusing on the explicit individual image and achieved high performances on domin-specific tasks [67]. By leveraging the acquired general knowledge, the model can achieve fast learning with only a small amount of labeled data.

III. Method

A. RSAM-Seg architecture

RSAM-Seg uses SAM as the backbone while retaining most of the structure of the decoder part. RSAM-Seg extracts features from remote sensing images without the necessity of human-provided prompts. To obtain more task-related information, the original encoder and decoder part of the model are modified. This adaptation enables better performance on remote sensing related tasks. To be specific, the ViT blocks of the encoder are modified by incorporating Adapter-Scale inside, and embedding Adapter-Feature between ViT layers to extract image information. We assume P^i refers to the prompts that generated from the extracted features of the image.

$$P^i = \mathrm{MLP}_{up}\left(\mathrm{GELU}\left(\mathrm{MLP}^i_{tune}\left(\mathrm{F_{pe}} + \mathrm{F_{hfc}}\right)\right)\right) \tag{1}$$

Where i denotes each individual adapter between ViT layers. $\mathrm{F_{pe}}$ and $\mathrm{F_{hfc}}$ stand for embedding features and High-Frequency Components (HFC) features. The mask decoder remains unchanged with no given prompt inputs and is fine-tuned using a pre-trained model. The architecture is shown in Figure 1.

B. Adapter details

1) Adapter-Scale: In the encoder, Adapter-Scale consists of three parts: Downscale, ReLu, and Upscale. The Downscale part uses a single Multi-Layer Perceptron (MLP) layer to reduce the dimensionality of the embedding. After applying the ReLu activation function, the embedding is restored to its original dimensionality using another MLP layer in the Upscale part. Two Adapter-Scale modules are inserted into each ViT block. The first is before the multi-head attention blocks and residual connections. The second is within the

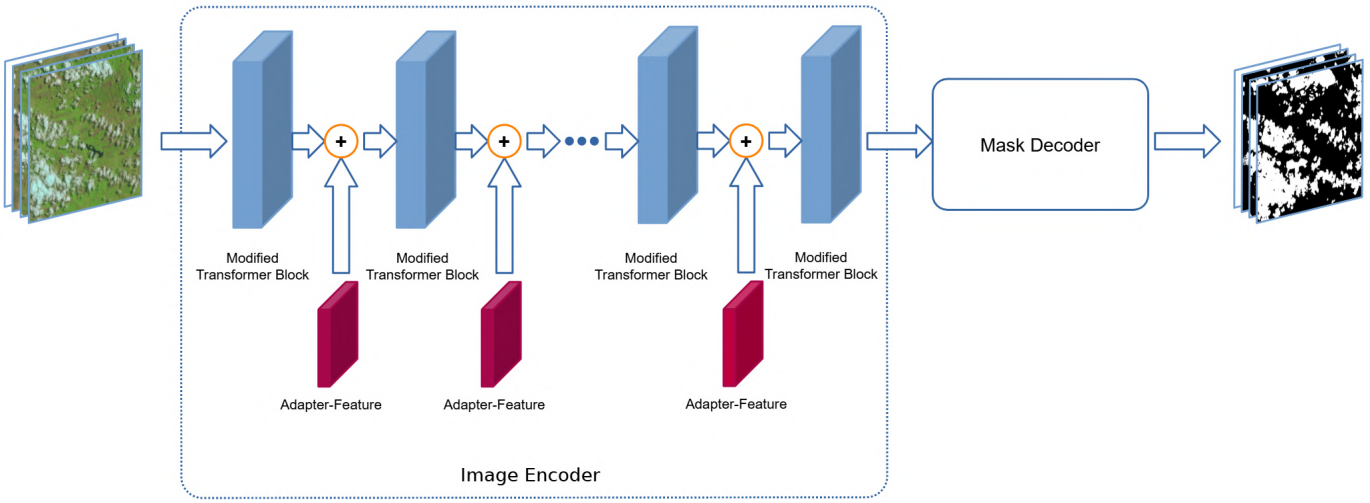

Fig. 1: The structure of RSAM-Seg. Adapter-Feature are inserted between modified ViT blocks while maintaining the mask decoder identical to the original SAM.

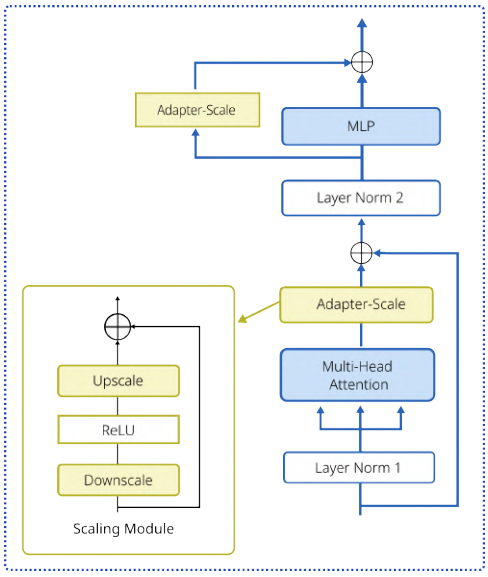

Fig. 2: The structure of the modified transformer block and Adapter-Scale in the encoder of RSAM-Seg.

residual structure of the MLP. Additionally, a scale factor of 0.5 is applied to each adapter. The structure of ViT blocks is shown in Figure 2.

2) Adapter-Feature: Between the ViT layers, Adapter-Feature consists of two MLPs. The first is the MLP_{tune}, which extracts features from remote sensing images to serve as prompts. The second MLP, MLP_{up}, is used to adjust the feature dimension to input into the ViT layer. The Adapter-Feature structure is shown in Figure 3.

In our work, both embedding features and high-frequency components features are tuned. In the part of embedding features, a linear layer with a scale factor is used to change the original embedding dimension. The structure of Adapter-Feature is shown in Figure 3.

In the part of HFC features, the HFC of the images are

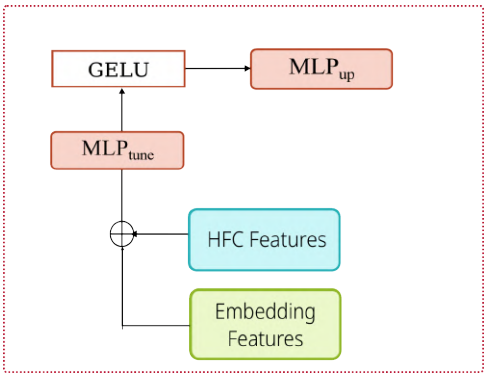

Fig. 3: The structure of the Adapter-Feature between the ViT blocks in the encoder of RSAM-Seg.

extracted and then inputted as prompts into the encoder. For an image I with dimensions of $H \times W$, high-frequency and low-frequency information can be extracted through Fast Fourier Transform (FFT) and inverse transforms. The high-frequency information of the image is of particular interest to us. `fft` and `ifft` are used to represent the Fast Fourier Transform and its inverse transform, respectively. The frequency components extracted from image I can be expressed by $f = \texttt{ftt}(I)$. Image I can also be restored through `ifft` by $I = \texttt{iftt}(f)$. To avoid the loss of information at the edges, a mask is used to selectively filter the high-frequency components,which can be done by shifting the low-frequency coefficients to the center of the image$(\frac{H}{2}, \frac{W}{2})$. The mask is generated with a mask ratio τ.

$$\mathbf{M}_h^{i,j}(\tau) = \begin{cases} 0, & \frac{4\left|\left(i-\frac{H}{2}\right)\left(j-\frac{W}{2}\right)\right|}{HW} \leq \tau \\ 1, & \textbf{otherwise} \end{cases} \tag{2}$$

where the symbol τ represents the proportion of the masked region. The HFC feature can obtain by:

$$I_{hfc} = \texttt{ifft}\left(f\mathbf{M}_h(\tau)\right) \tag{3}$$

IV. Experiment Settings

TABLE I. The detailed information of various datasets.

Scenarios	Region	Dataset	Total Area km^2	Resolution (m)
Building	Austin Chicago Kitsap CountyWest TyrolVienna	Inria	81	0.3
Cloud	-	38-Cloud	2,188,800	30
Field	France	Sentinel-2	785	6 0
Road	Thailand Indonesia India	DeepGlobe	2,220	0.50

A. Datasets

Since buildings, clouds, roads and fields are typical scenes in the realm of remote sensing, we select these four scenes to assess the performances of RSAM-Seg. Inria Aerial Image Labeling dataset, 38-Cloud dataset, DeepGlobe Roads dataset and Field Delineation dataset are utilized to respectively evaluate the building, cloud, field and road scenes [70]–[73]. The detailed information of four datasets are listed in Table I. And Figure 4 shows orignal images and ground truth masks of each dataset.

Inria : The Inria dataset has the coverage of 810 km² (405 km² for training and 405 km² for testing) and aerial orthorectified color imagery with a spatial resolution of 0.3 m. The ground truth data is separated for two semantic classes: building and not building. The original resolution of each image is 5000 × 5000 and then cropped to 1024 × 1024. Finally the building dataset contains 2380 labled patches and 500 unlabled patches as training and testing sets.

38 – Cloud : The 38-Cloud dataset is a public satellite cloud image dataset collected by the Landsat-8 satellite which includes 9 spectral bands. In this study, three commonly used bands are selected - Band 2 (blue), Band 3 (green), and Band 4 (red) - to compose a three-channel RGB image. The average cloud coverage of the Landsat-8 dataset is 51.6%. The original resolution of each image is 5000 × 5000 and then cropped to 1024 × 1024 patches. Finally the dataset contains 660 labled and 166 unlabeled patches as training and testing sets.

Sentinel – 2 : The Sentinel-2 field dataset has the resolution ranging from 10 to 60 meters in the visible, near infrared (VNIR), and short-wave infrared (SWIR) spectral zones, including 13 spectral channels. In this study, agricultural field scenes from France are selected. The dataset contains 1566 labled patches and 400 unlabeled patches as training and testing sets, each image is cropped to 224 × 224.

DeepGlobe : The dataset contains 6226 RGB images with resolution of 1024 × 1024, which covers images captured over Thailand, Indonesia and India. The satellite imagery mainly covers regions contained roads. In this study, owing to constraints in hardware capacity, a subset of 2500 patches is selected as the training set and 500 patches as the testing set.

B. Implementation details

In the experiment, ViT-L/14 version of SAM is utilized as the network backbone and trained all datasets using the AdamW optimizer. Additionally, cosine decay is applied to the learning rate and Binary Cross-Entropy (BCE) is used as the loss function. The network is trained for 60 epochs on all datasets. RSAM-Seg is implemented in PyTorch, an NVIDIA A40 (80GB) GPU are used for all experiments.

C. Performance metrics

Seven metrics are employed, including Jaccard index, precision, recall, specificity, F1 score, overall accuracy, and mIoU (mean Intersection over Union), to evaluate the performance of RSAM-Seg on different datasets, the baseline method chosen is U-Net, and the SAM operates in point mode in conjunction with the evaluation process. It is crucial to note that point mode encompasses two distinct variations: center(+) and center(-), representing the center point being labeled as the positive and negative class, respectively. In evaluation, center point coordinates of each image are inputted as prompts into SAM and examined its performance under both positive and negative center point labeling scenarios.

V. Results

In this section, we evaluate performance of RSAM-Seg compared to U-Net and original SAM on four datasets.

A. Results in various scenarios

1) Results in the cloud scenario: The quantitative results of various methods in the cloud scenario are summarized in Table II and the visualization results are listed in Figure 5.

From the "Cloud" scenario of Table II, RSAM-Seg significantly exceeds the basic SAM in all metrics. Compared to U-Net, RSAM-Seg exhibit superior performance in four comprehensive metrics, Jaccard, F1 score, overall accuracy and mIoU. RSAM-Seg demonstrates an average superiority over SAM under both modes by 36.7%.

From the images in the first row of Figure 5, it is evident that RSAM-Seg performs well specially when distinguishing thin cloud segments in the bottom-left. In contrast, SAM struggles to accurately identify cloud formations, resulting in large cloud areas being grouped into a single category. U-Net is able to segment thick clouds more accurately, but struggles with the segmentation of thin clouds. This indicates that RSAM-Seg is better suited for handling thin cloud scenarios.

2) Results in the field scenario: The quantitative results in the field scenario are summarized in Table II and the visualization results are listed in Figure 6.

The "Field" row of Table II reveals a enhancement across all metrics compared to the original SAM version. Specially, overall accuracy increased by 28.5%, and F1 score improved by 56%. Moreover, RSAM-Seg surpasses the baseline by 18% and 10% respectively.

It can be observed in the third row image of Figure 6 that RSAM-Seg performs well in distinguishing both regular

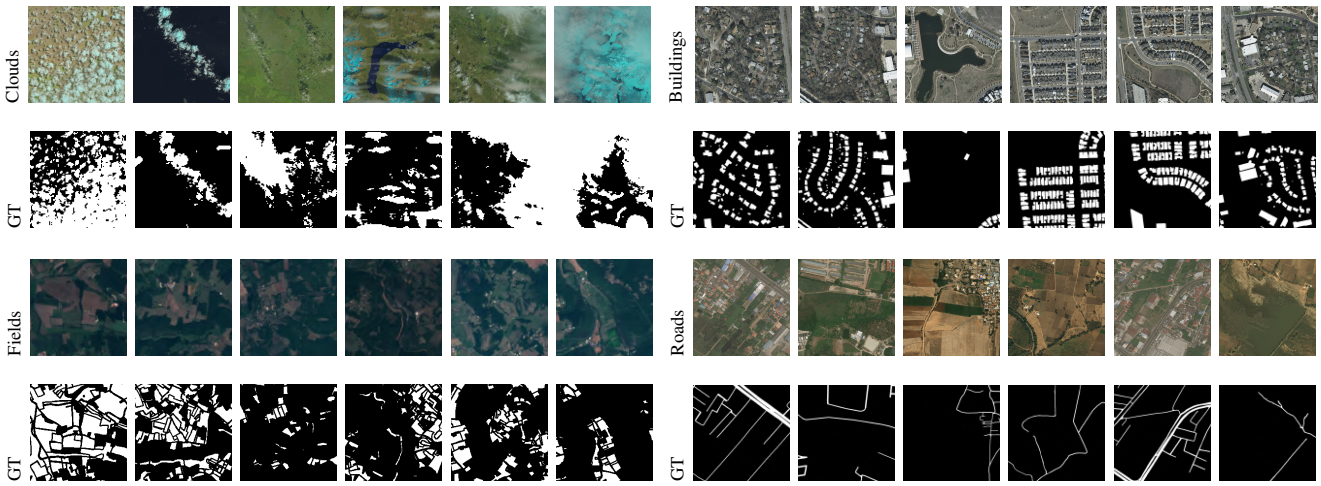

Fig. 4: The images in different datasets depict various scenes, including clouds, buildings, fields and roads. The image above shows a remote sensing image, with the corresponding mask displayed below. GT represents ground truth.

TABLE II. sons across multiple datasets in terms of Jaccard index, precision, recall, specificity, F1 score, overall accuracy, and mean intersection over union (mIoU).

Scenario	Method	Resolution	Jaccard	Precision	Recall	Specificity	F1 score	Overall accuracy	mIoU
Cloud	U-Net	1024*1024	0.6461	**0.8994**	0.6740	**0.8994**	0.7467	0.9021	0.7262
	SAM(center +)		0.1940	0.4107	0.3929	0.4107	0.2502	0.4838	0.2666
	SAM(center -)		0.0637	0.3652	0.0801	0.3652	0.0956	0.6221	0.3246
	RSAM-Seg		**0.731**	0.8301	**0.8396**	0.8301	**0.8152**	**0.9197**	**0.7646**
Field	U-Net	224*224	0.5011	0.6963	0.6484	0.6963	0.6391	0.7392	0.5545
	SAM(center +)		0.1798	0.5323	0.3399	0.5323	0.2627	0.513	0.2866
	SAM(center -)		0.065	0.5442	0.0751	0.5442	0.1176	0.5502	0.2989
	RSAM-Seg		**0.6346**	**0.76**	**0.7818**	**0.76**	**0.7592**	**0.8201**	**0.6634**
Building	U-Net	1024*1024	0.5047	0.7151	0.6318	0.7151	0.6496	0.9125	0.7017
	SAM(center +)		0.0046	0.1522	0.0062	0.1522	0.0083	0.807	0.4057
	SAM(center -)		0.0067	0.2146	0.0072	0.2146	0.0132	0.8433	0.4249
	RSAM-Seg		**0.7353**	**0.839**	**0.836**	**0.839**	**0.8337**	**0.9583**	**0.8424**
Road	U-Net	1024*1024	0.5286	0.6673	0.7276	0.6673	0.6774	0.974	0.7506
	SAM(center +)		0.0068	0.0244	0.0354	0.0244	0.0112	0.8706	0.4383
	SAM(center -)		0.0031	0.0216	0.005	0.0216	0.0061	0.9257	0.4644
	RSAM-Seg		**0.6195**	**0.7332**	**0.8104**	**0.7332**	**0.7548**	**0.9785**	**0.7982**

and irregular field scenes. SAM performs poorly in segmenting agricultural fields in densely populated areas and only identifies the agricultural field surrounding the given point, without taking the overall layout of the fields into account. U-Net struggles to accurately identify roads and other features separating agricultural fields. This suggests that RSAM-Seg is well-suited for handling complex and heterogeneous landscapes.

3) Results in the building scenario: The quantitative results in the building scenario are summarized in Table II and the visualization results are listed in Figure 7.

Examining the "Building" row in Table II, it's observable that the results surpass SAM across multiple evaluation metrics, while slightly exceeding the baseline by 5% and achieves a substantial average accuracy improvement of 42.71% under both operational modes of SAM in overall accuracy.

Upon scrutinizing the images in the central row of Figure 7, RSAM-Seg accurately distinguishes both the overall structures and scattered buildings. Furthermore, from the top-left images, RSAM-Seg effectively avoids interference from similar elevated structures within the scene. In contrast, SAM is limited by its dependence on prompts, resulting in only segmenting the area around the point prompt. Meanwhile, U-Net struggles with the segmentation when facing highway structures, leading to misclassification in certain scenarios. RSAM-Seg performs well in complex urban environments, highlighting its potential as a valuable tool for urban planning and management.

4) Results in the road scenario: The quantitative results in the road scenario are summarized in Table II and the visualization results are listed in Figure 8.

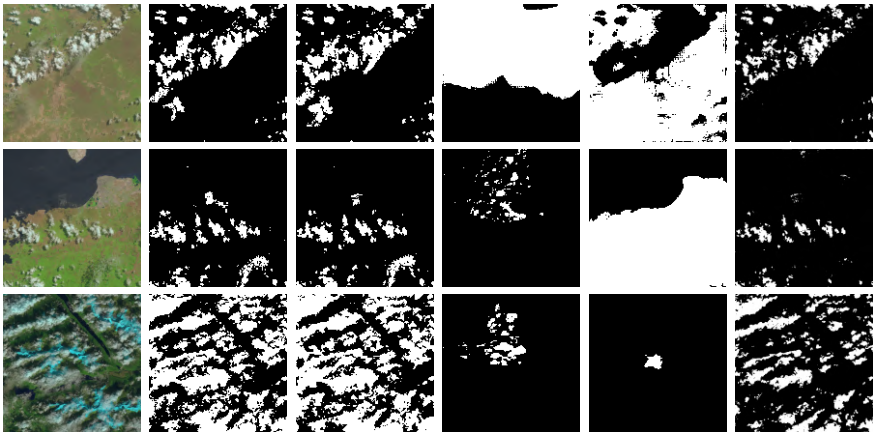

Fig. 5: Comparison of cloud segmentation results on 38-Cloud dataset with RSAM-Seg, SAM and U-Net.

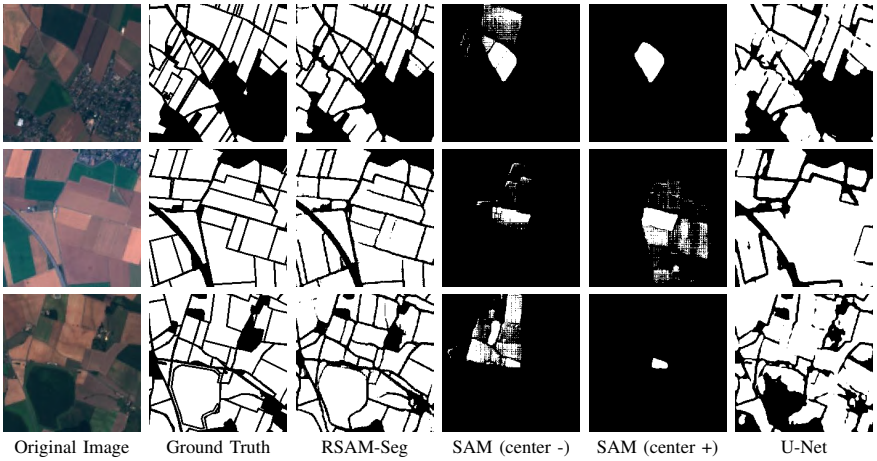

Fig. 6: Comparison of field segmentation results on Sentinel-2 dataset with RSAM-Seg, SAM and U-Net.

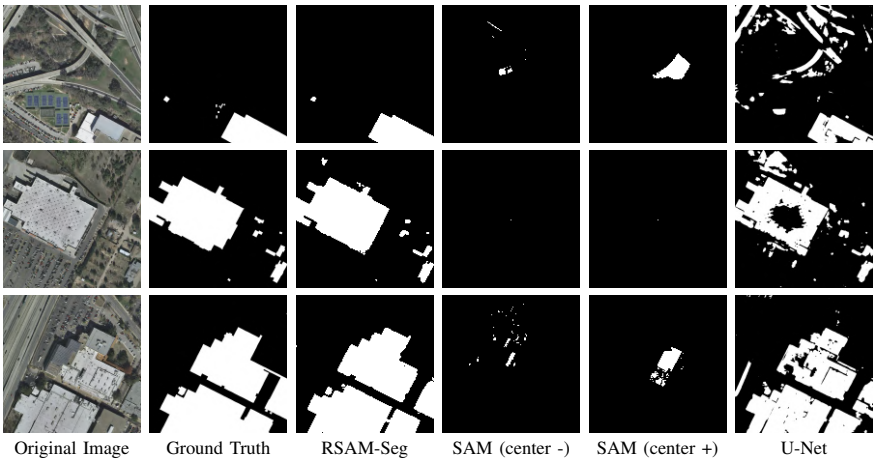

Fig. 7: Comparison of building segmentation results on Inria dataset with RSAM-Seg, SAM and U-Net.

The ”Road” row in Table II reveals RSAM-Seg outperforms the baseline with an 11% improvement in F1 score and SAM exhibits suboptimal performance around 45% in mIoU, which possibly attributed to the narrowness of the road and indistinct demarcation from the surrounding environment.

Considering the images in the third row of Figure 8, RSAM-Seg demonstrates the ability to distinguish densely roads, presenting well-defined and more complete road segments. SAM cannot effectively distinguish roads and is easily disrupted by surrounding environments, such as farmland, as seen in the experiments. Additionally, the U-Net is susceptible to misclassification of similar road-like areas, such as gaps

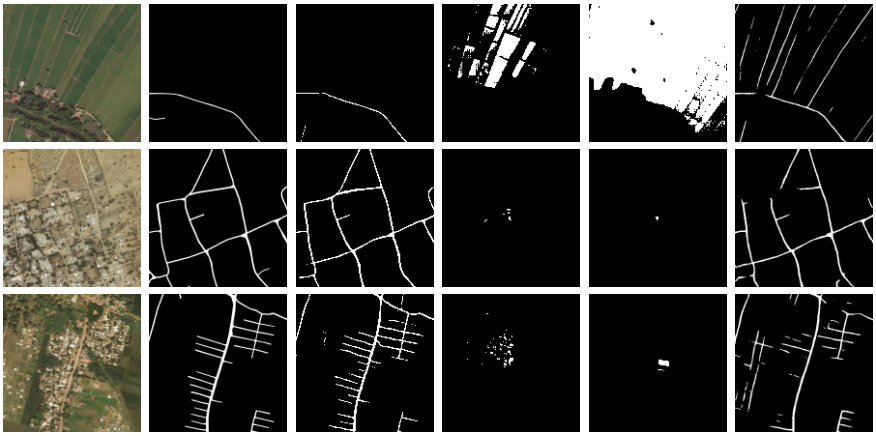

Fig. 8: Comparison of road Segmentation results on DG-Road dataset with RSAM-Seg, SAM and U-Net.

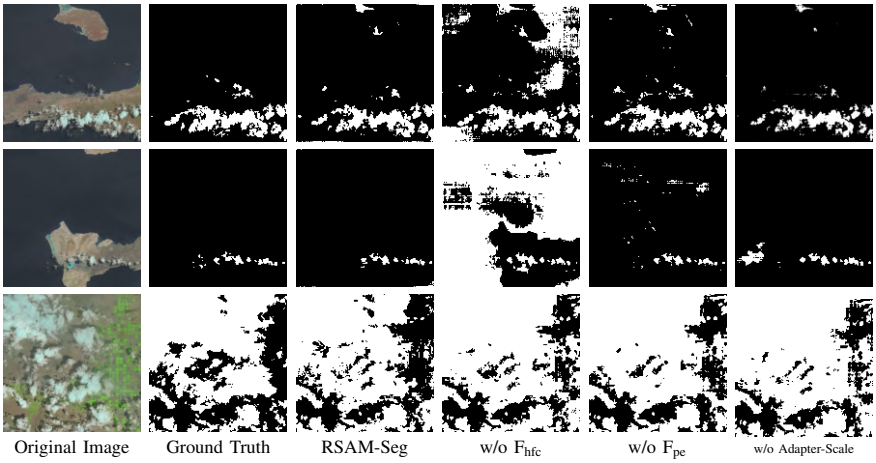

Fig. 9: Visualization results on the ablation study for RSAM-Seg on 38-Cloud dataset.

between farmland or buildings, when segmenting roads.

After conducting experiments on datasets from various remote sensing domains, it can be observed that SAM has the limitations that rely heavily on human annotations or prompts. However RSAM-Seg can perform significantly better than the original SAM approach and automate the segmentation process without the need for manually annotated data or prompts in specific remote sensing scenes. This modification enables SAM to better adapt to segmentation tasks in remote sensing imagery, making it a valuable tool for a wide range of remote sensing applications.

B. Ablation study

In order to systematically assess the contribution of different components in our proposed approach, an ablation study is conducted on 38-Cloud dataset and results are presented in Table III.

Table III presents the results on the 38-Cloud dataset, where the impact of removing F_{pe} and F_{hfc} within the Adapter-Feature, as well as removing Adapter-Scale from RSAM-Seg is evaluated. The experiments show that the F_{hfc} significantly improves the performance of RSAM-Seg on several evaluation metrics, indicating that it introduces high-frequency information into the model that is crucial for accurate image segmentation in remote sensing applications. Additionally, both the F_{pe} and Adapter-Scale modules contribute to the overall performance of RSAM-Seg in processing remote sensing imagery.

TABLE III. Ablation study for RSAM-Seg, showing the impact of Adapter-Scale and F_{pe}, F_{hfc} within Adapter-Feature on segmentation performance.

Method		Jaccard	Precision	Recall	Specificity	F1 score	Overall accu	mIoU
RSAM-Seg		**0.731**	**0.8301**	0.8396	**0.8301**	**0.8152**	**0.9197**	**0.7646**
Adapter-Feature	w/o F_{pe}	0.723	0.8146	0.8469	0.8146	0.8049	0.9131	0.7597
	w/o F_{hfc}	0.7118	0.7888	**0.8563**	0.7888	0.7885	0.892	0.7364
RSAM-Seg w/o Adapter-Scale		0.7287	0.8173	0.8562	0.8173	0.8114	0.9056	0.7608

The visualization results are listed in Figure 9, which can be observed that the F_{hfc} effectively reduces the interference from the surrounding environment of the clouds on the classification results. Furthermore, the combination of Adapter-Scale and F_{pe} further enhances the segmentation performance.

By analyzing both the quantitative and visualization results, the critical role of each component in the proposed method can be observed. These findings not only validate the effectiveness of RSAM-Seg but also provide valuable insights for future

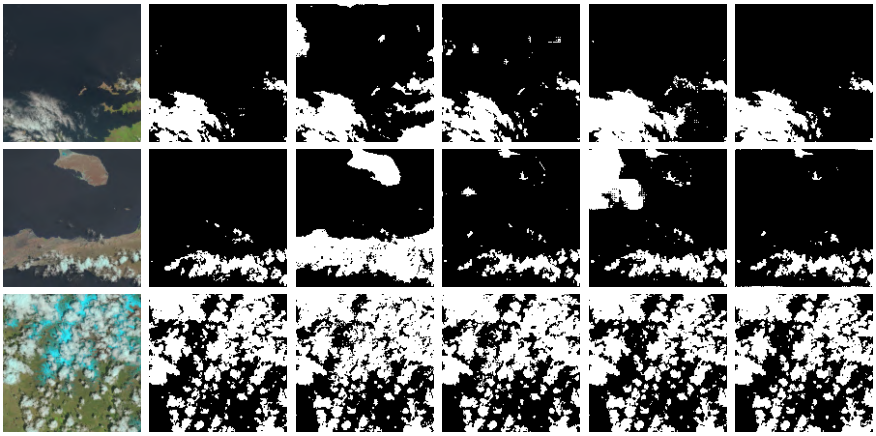

Fig. 10: Examples of few-shot segmentation results on 38-Cloud dataset.

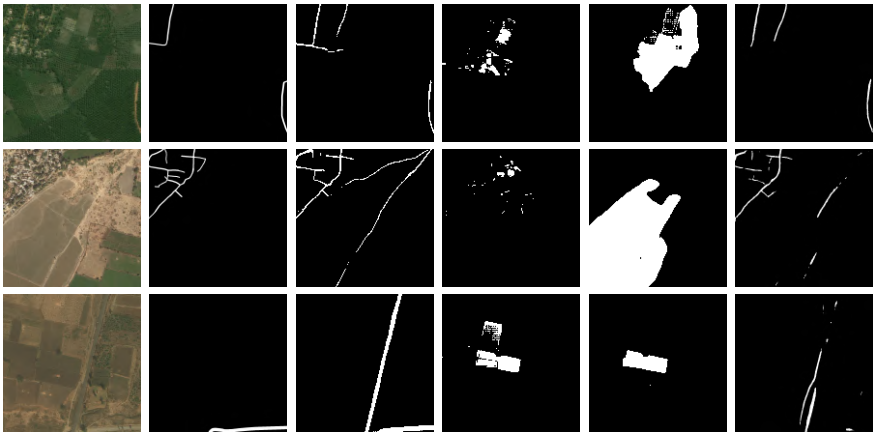

Fig. 11: Examples of completion results on the DG-Road dataset.

research and development in this field.

VI. Discussion

A. Few-shot scenario

In the experiment, RSAM-Seg exhibits a commendable degree of accuracy even in the challenging few-shot scenarios. Moreover, as the sample size expands, the performance of the model demonstrates a notable enhancement in terms of precision and predictive capability.

TABLE IV. The impact of dataset size on few-shot results

Dataset	Jaccard	Precision	Recall	Specificity	F1 score	Overall accu	mIoU
1% 38-Cloud	0.5552	0.7389	0.6777	0.7389	0.6561	0.7919	0.6412
10% 38-Cloud	0.6733	0.7984	0.8032	0.7984	0.7587	0.8738	0.7172
30% 38-Cloud	0.6940	0.7723	0.8597	0.7723	0.7770	0.8797	0.7234
70% 38-Cloud	**0.731**	**0.8301**	**0.8396**	**0.8301**	**0.8152**	**0.9197**	**0.7646**

The performance in the context of few-shot learning of RSAM-Seg is assessed using the 38-Cloud dataset. Table IV reflects the results under the condition where the original test set remains unchanged, 1% , 10%, 30% and 70% of images from the training set are randomly selected as new training subsets. Compared to the results of U-Net in Table II, RSAM-Seg demonstrates comparable efficacy to U-Net when only utilizing 10% of the dataset.

The visualization results are listed in the Figure 10, which reveals the potential of the methodology in the domain of remote sensing, particularly for few-shot image segmentation tasks.

B. Beyond Ground Truth

The experimental findings reveal a observation that our method surpasses the ground truth annotations of the dataset in certain scenarios, yielding segmentation results that exhibit superior accuracy and fidelity.

The Figure 11 shows the segmentation results in the road segmentation scenario. The snippets in the second row clearly demonstrate the capability of RSAM-Seg to segment roads that ground truth failed to identify, showcasing the completion capability of RSAM-Seg in delineating road regions from remote sensing imagery.

The segmentation results also indicate that the integration of domain-specific prior knowledge from remote sensing scenes into the SAM holds substantial promise for enhancing the construction of remote sensing datasets. RSAM-Seg exhibits

generalizability, positioning it as a formidable tool for auxiliary annotation purposes, thereby mitigating the burdensome costs associated with manual annotation and concurrently amplifying the overall efficiency of the process.

VII. Conclusion

We propose RSAM-Seg by incorporating specific prior information from the remote sensing domain and combined the high-frequency information of the images with their intrinsic features as prompts without manual prompt. Adapter-Feature and Adapter-Scale are integrated to enhance performance in semantic segmentation tasks involving remote sensing imagery. To evaluate the proposed methodology, comprehensive experiments are conducted in cloud, buildings, fields and roads scenarios. A meticulous comparative analysis is also conducted, benchmarking RSAM-Seg against the original architecture as well as the widely adopted U-Net model in the general semantic segmentation domain. The findings suggest that leveraging the incorporation of prior information, RSAM-Seg demonstrates promising capabilities in few-shot learning scenarios. Furthermore, RSAM-Seg holds potential as an auxiliary annotation tool, offering a novel approach to facilitate dataset creation while mitigating associated costs.

In the future, the primary focus will be on multi-object segmentation in few-shot scenarios, emphasizing the improvement of segmentation accuracy. Concurrently, there will be exploration into the optimization of efficiency and model compactness.

References

[1] B. Z. et al., "Progress and challenges in intelligent remote sensing satellite systems," *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing, vol. 15, pp. 1814–1822*.

[2] S. Mahajan and B. Fataniya, "Cloud detection methodologies: variants and development—a review," *Complex Intell. Syst., vol. 6, no. 2, pp. 251–261*.

[3] J. E. Patino and J. C. Duque, "A review of regional science applications of satellite remote sensing in urban settings," *Computers, Environment and Urban Systems, vol. 37, pp. 1–17*.

[4] S. Khanal, J. P. F. K. Kc, S. Shearer, and E. Ozkan, "Remote sensing in agriculture—accomplishments, limitations, and opportunities," *Remote Sensing, vol. 12, no. 22, p. 3783*.

[5] Z. C. et al, "Road extraction in remote sensing data: A survey," *International Journal of Applied Earth Observation and Geoinformation, vol. 112, p. 102833*.

[6] C. Gonzales and W. Sakla, "Semantic segmentation of clouds in satellite imagery using deep pre-trained u-nets," *IEEE Applied Imagery Pattern Recognition Workshop (AIPR), 1–7*.

[7] T. Blaschke, G. J. Hay, Q. Weng, and B. Resch, "Collective sensing: Integrating geospatial technologies to understand urban systems—an overview," *Remote Sensing, vol. 3, no. 8, pp. 1743–1776*.

[8] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *IEEE Geosci. Remote Sensing Lett., vol. 14, no. 5, pp. 778–782*.

[9] Rasti, Behnood, Chang, Yi, Dalsasso, Emanuele, Denis, Loic, Ghamisi, and Pedram, "Image restoration for remote sensing: Overview and toolbox," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 2, pp. 201–230, 2022.

[10] Zhang, Liangpei, Zhang, Lefei, and D. Bo, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, 2016.

[11] W. Wang, N. Yang, Y. Zhang, F. Wang, T. Cao, and P. Eklund, "A review of road extraction from remote sensing images," *Journal of Traffic and Transportation Engineering (English Edition), vol. 3, no. 3, pp. 271–282*, Jun. 2016.

[12] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing, vol. 65, no. 1, pp. 2–16*, Jan. 2010.

[13] M. D. Hossain and D. Chen, "Segmentation for object-based image analysis (obia): A review of algorithms and challenges from remote sensing perspective," *ISPRS Journal of Photogrammetry and Remote Sensing, vol. 150, pp. 115–134*, Apr. 2019.

[14] I. Kotaridis and M. Lazaridou, "Remote sensing image segmentation advances: A meta-analysis," *ISPRS Journal of Photogrammetry and Remote Sensing, vol. 173, pp. 309–322*, Mar. 2021.

[15] Moser, G., Serpico, S.B., and B. J.A., "Land-cover mapping by markov modeling of spatial-contextual information in very-high-resolution remote sensing images," *Proc. IEEE 101, 631–651*, 2013.

[16] Dey, V., Zhang, Y., Zhong, and M., "A review on image segmentation techniques with remote sensing perspective," *Wagner, W., Székely, B. (Ed.), ISPRS TC VII Symposium – 100 Years ISPRS. Vienna, pp. 31–42.*, 2010.

[17] Schiewe and J., "Segmentation of high-resolution remotely sensed data-concepts, ap- plications and problems," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 34, 380–385.*, 2002.

[18] Li, X., Myint, S.W., Zhang, Y., Galletti, C., Zhang, X., Turner, and B.L, "Object-based land-cover classification for metropolitan phoenix, arizona, using aerial photo- graphy," *Int. J. Appl. Earth Obs. Geoinf. 33, 321–330.*, 2014.

[19] Sharma, Richa, Ghosh, Aniruddha, Joshi, and P.K., "Decision tree approach for classification of remotely sensed satellite data using open source support." *J. Earth Syst. Sci. 122 (5), 1237–1247.*, 2013.

[20] Mountrakis and Giorgos., "Support vector machines in remote sensing: A review." *ISPRS Journal of Photogrammetry and Remote Sensing*, 2011.

[21] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Transactions on geoscience and remote sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.

[22] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions." *ISPRS Journal of Photogrammetry and Remote Sensing.*, 2016.

[23] M. Pal and P. M. Mather, "Support vector machines for classification in remote sensing," *International journal of remote sensing*, vol. 26, no. 5, pp. 1007–1011, 2005.

[24] M. Belgiu, L. Drăguţ, and J. Strobl, "Quantitative evaluation of variations in rule-based classifications of land cover in urban neighbourhoods using worldview-2 imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 87, pp. 205–215, 2014.

[25] R. R. Colditz, "An evaluation of different training sample allocation schemes for discrete and continuous land cover classification using decision tree-based algorithms," *Remote Sensing*, vol. 7, no. 8, pp. 9655–9681, 2015.

[26] A. Juel, G. B. Groom, J.-C. Svenning, and R. Ejrnaes, "Spatial application of random forest models for fine-scale coastal vegetation classification using object based analysis of aerial orthophoto and dem data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 42, pp. 106–114, 2015.

[27] A. Mellor, S. Boukir, A. Haywood, and S. Jones, "Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 105, pp. 155–168, 2015.

[28] Huang, Bo, Zhao, Bei, Song, and Yimeng, "Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery." *Remote Sens. Environ. 214 (September), 73–86.*, 2018.

[29] Diakogiannis, F. I., Waldner, François, Caccetta, Peter, Chen, and Wu., "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data." *ISPRS J. Photogramm. Remote Sens. 162 (April), 94–114.*, 2020.

[30] G.-J. Qi and J. Luo, "Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods," *IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 4, pp. 2168–2187*, Apr. 2022.

[31] M. Toldo, A. Maracani, U. Michieli, and P. Zanuttigh, "Unsupervised domain adaptation in semantic segmentation: A review," *Technologies, vol. 8, no. 2, p. 35*, Jun. 2020.

[32] Li, Wenmei, W. Ziteng, Wang, Yu, Wu, Jiaqi, Wang, Juan, Jia, Yan, Gui, and Guan, "Classification of high-spatial-resolution remote sensing scenes method using transfer learning and deep convolutional neural network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1986–1995, 2020.

[33] P. de Lima, R., and M. K., "Convolutional neural network for remote-sensing scene classification: Transfer learning analysis," *Remote Sens. 2020, 12, 86.*, 2020.

[34] L. Mengxi, S. Qian, C. Zhuoqun, and L. Jianlong, "Pa-former: Learning prior-aware transformer for remote sensing building change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[35] Q. Zeng and J. Geng, "Task-specific contrastive learning for few-shot remote sensing image scene classification," *ISPRS Journal of Photogrammetry and Remote Sensing,vol191, Pages 143-154*, Sept 2022.

[36] R. Wenqi, T. Yang, S. Qiyu, Z. Chaoqiang, and H. Qing-Long, "Visual semantic segmentation based on few/zero-shot learning: An overview," *IEEE/CAA Journal of Automatica Sinica*, pp. 1–21, 2023.

[37] Z. Chang, Y. Lu, X. Ran, X. Gao, and X. Wangg, "Few-shot semantic segmentation: a review on recent approaches," *Neural Comput and Applic, vol. 35, pp. 18251–18275*, Sept 2023.

[38] A. K. et al., "Segment anything," *arXiv, Apr. 05, 2023*, May 2023.

[39] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. jamin Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, , and D. Amodei., "Language models are few-shot learners," *NeurIPS 2020*, 2020.

[40] X. Yuan, J. Shi, and L. Gu., "A review of deep learning methods for semantic segmentation of remote sensing imagery." *Expert Systems with Applications, vol. 169, p. 114417.*, May 2021.

[41] Ball, J. E., Anderson, D. T., Chan, and C. S., "Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community," *Journal of Applied Remote Sensing, 11, 42609.*, 2017.

[42] Zhang, L., Du, and B., "Deep learning for remote sensing data: A technical tutorial on the state of the art." *IEEE Geoscience and Remote Sensing Magazine, 4,22–40.*, 2016.

[43] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, and A. L., "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." *IEEE Transactions on Pattern Analysis and Machine Intelligence,40, 834–848*, 2018.

[44] S. Liu, L. Liu, F. Xu, J. Chen, Y. Yuan, and X. Chen, "A deep learning method for individual arable field (iaf) extraction with cross-domain adversarial capability," *Computers and Electronics in Agriculture*, vol. 203, p. 107473, 2022.

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

[46] Z. Dong, S. An, J. Zhang, J. Yu, J. Li, and D. Xu, "L-unet: A landslide extraction model using multi-scale feature fusion and attention mechanism," *Remote Sensing*, vol. 14, no. 11, p. 2552, 2022.

[47] Y. Hou, Z. Liu, T. Zhang, and Y. Li, "C-unet: Complement unet for remote sensing road extraction," *Sensors*, vol. 21, no. 6, p. 2153, 2021.

[48] J. Ahn, S. Cho, and S. Kwak, "Weakly supervised learning of instance segmentation with inter-pixel relations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2209–2218.

[49] Z. Huang, X. Wang, J. Wang, W. Liu, and J. Wang, "Weakly-supervised semantic segmentation network with deep seeded region growing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7014–7023.

[50] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[51] K. Fu, W. Lu, W. Diao, M. Yan, H. Sun, Y. Zhang, and X. Sun, "Wsf-net: Weakly supervised feature-fusion network for binary segmentation in remote sensing image," *Remote Sensing*, vol. 10, no. 12, p. 1970, 2018.

[52] S. Wang, W. Chen, S. M. Xie, G. Azzari, and D. B. Lobell, "Weakly supervised deep learning for segmentation of remote sensing imagery," *Remote Sensing*, vol. 12, no. 2, p. 207, 2020.

[53] J. Nyborg and I. Assent., "Weakly-supervised cloud detection with fixed-point gans." *IEEE International Conference on Big Data (Big Data)*, 2021.

[54] G. Z. H. C. X. L. S. H. J. M. Z. L. H. L. Chen, Yujia and H. Wang., "A novel weakly supervised semantic segmentation framework to improve the resolution of land cover product." *ISPRS Journal of Photogrammetry and Remote Sensing.*, 2023.

[55] X. Z. Z. P. X. J. X. T. Wang, Guanchun and L. Jiao., "Mol: Towards accurate weakly supervised remote sensing object detection via multi-view noisy learning." *ISPRS Journal of Photogrammetry and Remote Sensing 196.*, 2023.

[56] Y. Xu and P. Ghamisi, "Consistency-regularized region-growing network for semantic segmentation of urban scenes with point-level annotations," *IEEE Transactions on Image Processing*, vol. 31, pp. 5038–5051, 2022.

[57] Y. G. G. S. L. Y. M. D. Zhu, Jingru and J. Chen., "Unsupervised domain adaptation semantic segmentation of high-resolution remote sensing imagery with invariant domain-level prototype memory." *IEEE Transactions on Geoscience and Remote Sensing 61.*, 2023.

[58] Y. G. G. S. Y. Z. J. Chen, J. Zhu and M. Deng., "Unsupervised domain adaptation for semantic segmentation of high-resolution remote sensing imagery driven by category-certainty attention." *IEEE Trans. Geosci. Remote Sens., vol. 60*, 2022.

[59] Y. Z. W. C. Z. W. Y. Li, T. Shi and H. Li., "Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation." *ISPRS J. Photogramm. Remote Sens., vol. 175*, 2021.

[60] J. Z. L. Zhang, M. Lan and D. Tao., "Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images." *IEEE Trans. Geosci. Remote Sens., vol. 60.*, 2021.

[61] P. H. J. Z. Y. G. G. S. M. D. Chen, Jie and H. Li., "Memory-contrastive unsupervised domain adaptation for building extraction of high-resolution remote sensing imagery." *IEEE Transactions on Geoscience and Remote Sensing 61 (2023): 1–15.*, 2021.

[62] Y. C. et al., "Bifdanet: Unsupervised bidirectional domain adaptation for semantic segmentation of remote sensing images." *Remote Sens.*, 2022.

[63] X. Yao, Q. Cao, X. Feng, G. Cheng, and J. Han, "Scale-aware detailed matching for few-shot aerial image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[64] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang, "Sg-one: Similarity guidance network for one-shot semantic segmentation," *IEEE transactions on cybernetics*, vol. 50, no. 9, pp. 3855–3865, 2020.

[65] Y. Liu, N. Liu, Q. Cao, X. Yao, J. Han, and L. Shao, "Learning non-target knowledge for few-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 573–11 582.

[66] J. Wu, R. Fu, H. Fang, Y. Liu, Z. Wang, Y. Xu, Y. Jin, and T. Arbel, "Medical sam adapter: Adapting segment anything model for medical image segmentation," *arXiv preprint arXiv:2304.12620*, 2023.

[67] W. Liu, X. Shen, C.-M. Pun, and X. Cun, "Explicit visual prompting for low-level structure segmentations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 434–19 445.

[68] G. Cheng, L. Cai, C. Lang, X. Yao, J. Chen, L. Guo, and J. Han, "Spnet: Siamese-prototype network for few-shot remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2021.

[69] X. Li, D. Shi, X. Diao, and H. Xu, "Scl-mlnet: Boosting few-shot remote sensing scene classification via self-supervised contrastive learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[70] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017.

[71] S. Mohajerani and P. Saeedi, "Cloud-net: An end-to-end cloud detection algorithm for landsat 8 imagery," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 2019, pp. 1029–1032.

[72] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[73] H. L. Aung, B. Uzkent, M. Burke, D. Lobell, and S. Ermon, "Farm parcel delineation using spatio-temporal convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 76–77.

Jie Zhang is currently pursuing the M.S. degree in computer science from Nanjing Forestry University, Nanjing, China. His research interests include computer vision and remote sensing image processing.

XUBING YANG received his B.S. degree in mathematics from Anhui University in 1997. He completed his M.S. and Ph.D. degrees in computer applications at Nanjing University of Aeronautics & Astronautics (NUAA) in 2004 and 2008, respectively. Since 2008, he joined Nanjing Forestry University and now worked as an associate professor at computer science and engineering department at NFU. His research interests include pattern recognition, machine learning, and neural computing.

RUI JIANG received the B.S. and Ph.D. degrees in electronic engineering from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2007 and 2013, respectively. In 2013 he joined the Department of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China, where He is now an associate professor. His research interests include machine learning and Internet of things (IoT).

WEI SHAO received the B.Sc. and M.Sc. degrees in information and computing science from Nanjing University of Technology, China in 2009 and 2012, respectively, and the Ph.D. degree in software engineering from Nanjing University of Aeronautics and Astronautics, China in 2018. He is currently an associate professor with College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China. His research interests include machine learning and bioinformatics.

LI ZHANG received the B.S. degree in computer science from Changsha University of Science & Technology, the M.S. and Ph.D degree in computer science from Nanjing University of Aeronautics and Astronautics (NUAA) in 2007, 2010 and 2015 respectively. He is currently an associate professor in the College of Information Science and Technology, Nanjing Forestry University. His research interests include machine learning, remote sensing and medical imaging analysis.